Name: Özgün Can Yürütken
Number: 04170106

EHB328 Ödev 6

K-means

Parametreler:

Seed(1): Seed sayısı ile küme merkezlerinin nereden başlayacağını belirleyebiliriz. Normalde her seferde değişen başlangıç merkezlerini her kümelemede kullanacağımız sabit bir seed sayısı ile aynı yerlerden başlamasını sağlayabiliriz. Büyük verilerde optimize edilebilecek bir parametre iken bizim için hangi sayı olduğu önemsiz.

displayStdDevs(True): Her bir kümedeki her bir öznitelik için, küme merkezine göre standart sapmasını (+-) olarak gösterir. Daha fazla bilgi için True seçtim.

numExecutionSlots(1): İşlem süresince işlemcide çalıştırılacak çekirdek sayısı. Küçük veri olduğundan önemsiz.

dontReplaceMissingValues(True): Verideki eksik hücreleri ortalama değer ile değiştirir. Verimizde eksik değer olmadığı için True (don't) seçtim.

debug(True): Ekstra bilgi için.

numClusters(3): Kümelemede oluşturulacak grup sayısı. Verimizde 3 ayrı sınıf olduğundan 3 kümeye ayırdım.

doNotCheckCapabilities(False): İşlem süresini kısaltmak için kullanılabilecek bir parametre. Veri küçük olduğundan önemsiz.

maxIterations(10): Kümelemedeki maksimum tekrar sayısı. Veri küçük olduğundan 10 seçtim. Sonuçta da görülebileceği gibi 3 tekrarda kümeleme sonuçlanıyor.

preserveInstancesOrder(True): ???

initializationMethod(Random): Kümeleme yapılırken kullanılacak yöntem. Kmeans++, Canopy veya farthest first.

distanceFunction(EuclideanDistance): Hata fonksiyonu. Veri küçük olduğundan ve lineer bağımlı olmadığından Öklid uzaklığı kullandım.

fastDistanceCalc(False): Optimizasyon için bir parametre, uzaklık ölçümü için cut-off değerleri kullanarak işlem süresini azaltır fakat daha fazla hataya yol açabilir. Optimizasyona ihtiyacımız olmadığından False seçtim.

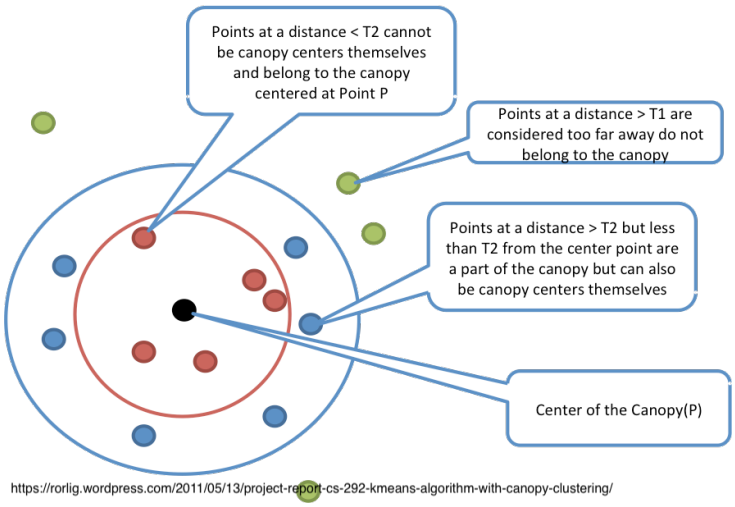
reduceNumberOfDistanceCalcsViaCanopies(True): Canopy yöntemi ile işlem sırasında daha az uzaklık hesabı yapılır.**canopyMinimumCanopyDensity(2.0):** Canopy yöntemi için bir parametre. Varsayılan olarak bıraktım.

canopyT1(-1.25): İlk çemberin yarıçapı. Varsayılan olarak bıraktım.

canopyT2(-1.0): İkinci çemberin yarıçapı. Varsayılan olarak bıraktım.

canopyPeriodicPruningRate(10000): Canopy yöntemi için bir parametre. Default olarak bıraktım.

canopyMaxNumCanopiesToHoldInMemory(100): Bellekte tutulacak maksimum canopy sayısı. Default olarak bıraktım.




```
Scheme:       weka.clusterers.SimpleKMeans -init 0 -C -max-candidates 100 -periodic-pruning 10000
-min-density 2.0 -t1 -1.25 -t2 -1.0 -V -M -N 3 -A "weka.core.EuclideanDistance -R first-last" -I
10 -O -num-slots 1 -S 1 -output-debug-info
Relation:     irisdata
Instances:    150
Attributes:   5
              sepallenght
              sepalwidht
              petallenght
              petalwidht
              class
Test mode:    evaluate on training data
=== Clustering model (full training set) ===
kMeans
======
Number of iterations: 3
Within cluster sum of squared errors: 7.817456892309573

Initial starting points (random):

Cluster 0: 7.7,3,6.1,2.3,Iris-virginica
Cluster 1: 6.3,2.5,4.9,1.5,Iris-versicolor
Cluster 2: 6.4,2.7,5.3,1.9,Iris-virginica

Reduced number of distance calculations by using canopies.
Canopy T2 radius: 0.874
Canopy T1 radius: 1.092

Final cluster centroids:                                        Cluster#
Attribute                 Full Data                 0                 1                 2
                            (150.0)           (50.0)            (50.0)            (50.0)
=============================================================================================
sepallenght                  5.8433            5.006             5.936             6.588
                            +/-0.8281         +/-0.3525         +/-0.5162         +/-0.6359

sepalwidht                    3.054             3.418              2.77             2.974
                            +/-0.4336          +/-0.381         +/-0.3138         +/-0.3225

petallenght                  3.7587             1.464              4.26             5.552
                            +/-1.7644         +/-0.1735         +/-0.4699         +/-0.5519

petalwidht                   1.1987             0.244             1.326             2.026
                            +/-0.7632         +/-0.1072         +/-0.1978         +/-0.2747

class                    Iris-setosa       Iris-setosa    Iris-versicolor     Iris-virginica
  Iris-setosa             50.0 ( 33%)       50.0 (100%)        0.0 (  0%)         0.0 (  0%)
  Iris-versicolor         50.0 ( 33%)        0.0 (  0%)       50.0 (100%)         0.0 (  0%)
  Iris-virginica          50.0 ( 33%)        0.0 (  0%)        0.0 (  0%)        50.0 (100%)

Time taken to build model (full training data) : 0.02 seconds

=== Model and evaluation on training set ===

Clustered Instances:
0       50 ( 33%)
1       50 ( 33%)
2       50 ( 33%)
```

Oluşturulan 3 kümenin, verideki 3 sınıf ile %100 oranında eşleştiği görülmüştür.

Random Forest

Parametreler: **(İlk örnekte açıklaması yapılan bazı parametreleri tekrar açıklamadım.)**

seed(1): K-means yöntemine benzer bir şekilde işleme aynı random değerlerden başlanmasını sağlar.

representCopiesUsingWeights: Örneklerin açıkça değil de ağırlıklar kullanılarak gösterilmesi. Gerek duyulmadığı için False seçtim.

storeOutOfBagPredictions: Out of bag durumlarını bellekte saklama. Varsayılanda bıraktım.

bagSizePercent(100): Her çantanın büyüklüğü (Eğitim verisinin bir yüzdesi olarak)

numDecimalPlaces(2): İşlemde kullanılacak basamak sayısı. Varsayılanda bıraktım.

batchSize: Çoklu tahmin yapma durumunda kullanılan bir parametre. Varsayılanda bıraktım.

printClassifiers(False): Sınıflandırma işleminde kullanılan her bir ağacı yazdırmak. Çıktıyı çok uzattığı için False seçtim.

numIterations(10): Sınıflandırmadaki ağaç sayısı. Küçük veri olduğundan 10 seçtim.

outputOutOfBagComplexityStatistics(False): Out of bag durumu oluştuğunda gösterilecek istatistikler.

breakTiesRandomly(True): Birkaç öznitelik eşit derecede iyi gözüküyor ise rastgele seçim yapar.

maxDepth(0): Ağaçların maksimum derinliği. Sonsuz için 0 değeri verdim.

computeAttributeImportance(True): Özniteliklerin önem sıralamasını yazdırmak.

calcOutOfBag(False): Out of bag hatası oluştuğunda yapılacak hesaplamalar. Error olmadığı için False seçtim.

numFeatures(4): Rastgele seçilecek öznitelik sayısı. 4 özniteliğimiz olduğundan 4 seçtim.

%100 başarılı bir sınıflandırma olduğundan performans metriklerinin hepsi 1.0 değerinde.

Precision	Recall	F-Measure
1.000	1.000	1.000
1.000	1.000	1.000
1.000	1.000	1.000
1.000	1.000	1.000

```
Scheme:       weka.classifiers.trees.RandomForest -P 100 -attribute-importance -I 10 -num-slots 1 -K 4 -M 1.0 -V
0.001 -S 1 -B -output-debug-info
Relation:     irisdata
Instances:    150
Attributes:   5
              sepallenght
              sepalwidht
              petallenght
              petalwidht
              class
Test mode:    evaluate on training data
=== Classifier model (full training set) ==
RandomForest

Bagging with 10 iterations and base learner

weka.classifiers.trees.RandomTree -K 4 -M 1.0 -V 0.001 -S 1 -B -output-debug-info -do-not-check-capabilities

Attribute importance based on average impurity decrease (and number of nodes using that attribute)

      0.79 (     4)  sepalwidht
      0.61 (     8)  sepallenght
      0.61 (    25)  petallenght
      0.55 (    20)  petalwidht


Time taken to build model: 0.01 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correctly Classified Instances         150               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0.0147
Root mean squared error                  0.0677
Relative absolute error                  3.3    %
Root relative squared error             14.3527 %
Total Number of Instances              150

=== 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
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Iris-versicolor
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Iris-virginica
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

  a  b  c   <-- classified as
 50  0  0 |  a = Iris-setosa
  0 50  0 |  b = Iris-versicolor
  0  0 50 |  c = Iris-virginica
```

Ada Boost

Parametreler: (İlk ve ikinci örnekte açıklaması yapılan bazı parametreleri tekrar açıklamadım.)

weightThreshold(100): Pruning işlemi için eşik ağırlık değeri. Varsayılanda bıraktım.

numIterations(100): Daha iyi sonuç vermesi için 100 seçtim.

resume(True): Sınıflandırıcının, belirlenen tekrar sayısına ulaşıldıktan sonra eğitime devam etmesini sağlar. Modelin boyutunu arttırabilir. True seçtim fakat sonuçta bir değişiklik olmadı.

useResampling(True): Yeniden ağırlıklandırma yerine yeniden örnekleme kullanılması. Daha iyi sonuç verdiği için True seçtim.

classifier(DecisionStump): Toplu öğrenmede kullanılacak zayıf sınıflandırıcı. Modele uygun olduğu için Decision Stump seçtim. (Decision Stump: Dal ayrımında sadece tek bir özniteliğe bakılan karar ağacı yöntemi.)

```
=== Run information ===

Scheme:       weka.classifiers.meta.AdaBoostM1 -Q -P 100 -resume -S 1 -I 100 -W weka.classifiers.trees.DecisionStump -output-debug-info
Relation:     irisdata
Instances:    150
Attributes:   5
              sepallenght
              sepalwidht
              petallenght
              petalwidht
              class
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===


Number of performed Iterations: 34


Time taken to build model: 0.04 seconds

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

Correctly Classified Instances         143               95.3333 %
Incorrectly Classified Instances         7                4.6667 %
Kappa statistic                          0.93
Mean absolute error                      0.0506
Root mean squared error                  0.1661
Relative absolute error                 11.3784 %
Root relative squared error             35.2302 %
Total Number of Instances              150

=== 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
                 0.920    0.030    0.939      0.920    0.929      0.895    0.965     0.946     Iris-versicolor
                 0.940    0.040    0.922      0.940    0.931      0.896    0.975     0.917     Iris-virginica
Weighted Avg.    0.953    0.023    0.953      0.953    0.953      0.930    0.980     0.954

=== Confusion Matrix ===

  a  b  c   <-- classified as
 50  0  0 |  a = Iris-setosa
  0 46  4 |  b = Iris-versicolor
  0  3 47 |  c = Iris-virginica
```

Precision	Recall	F-Measure	Class
1.000	1.000	1.000	Iris-setosa
0.939	0.920	0.929	Iris-versicolor
0.922	0.940	0.931	Iris-virginica
0.953	0.953	0.953	

Yorum

K-means kümeleme ve random forest sınıflandırma yöntemi, küçük ve karmaşık olmayan bir veri olduğu için %100 doğruluk ile sonuç verdi.

Ada-boost yönteminin %100 doğruluk vermemesinin sebebi kullanılan zayıf sınıflandırıcının zayıflığı olabilir.