EM za vise klasa

EM (Expectation-Maximization) algoritam je metod za klasterizaciju podataka koji iterativno pronalazi parametre modela koji najbolje odgovaraju podacima. Počinje sa inicijalizacijom parametara modela, uključujući centre klastera, matrice kovarijanse (za Gausijanac mješoviti model) i početne verovatnoće pripadanja klasterima. Zatim, iterira između koraka očekivanja, gde se procenjuje verovatnoća pripadanja svakoj tački klasteru, i koraka maksimizacije, gde se ažuriraju parametri modela na osnovu ovih procena. Ovaj proces se ponavlja sve dok se ne postigne konvergencija, optimizirajući parametre kako bi se maksimizovala verovatnoća podataka datih modelom.

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

Clustered Instances

0       198 (  4%)
1      1435 ( 26%)
2       904 ( 16%)
3      1251 ( 23%)
4       633 ( 12%)
5      1078 ( 20%)


Log likelihood: 215.2771


Class attribute: classification
Classes to Clusters:

    0    1    2    3    4    5  <-- assigned to cluster
    0  287  710  103    0    0 | DoS Hulk
    5  567    0    3    0  525 | FTP-Patator
  190   41  193   47  628    0 | DoS Slowhttptest
    3  540    1    3    0  553 | SSH-Patator
    0    0    0 1095    5    0 | PortScan

Cluster 0 <-- No class
Cluster 1 <-- FTP-Patator
Cluster 2 <-- DoS Hulk
Cluster 3 <-- PortScan
Cluster 4 <-- DoS Slowhttptest
Cluster 5 <-- SSH-Patator

Incorrectly clustered instances :       1946.0   35.3883 %
```

KMeans binarni

K-means je popularan algoritam za klasterizaciju podataka koji radi tako što pokušava da grupiše tačke podataka u predefinisani broj klastera na osnovu njihovih karakteristika. Ovaj algoritam radi na sledeći način: prvo se biraju početni centri klastera slučajnim izborom ili sistematski na osnovu podataka. Zatim se iterativno izvršavaju dva koraka: prvo, svaka tačka podataka se dodeljuje najbližem centru klastera, na osnovu udaljenosti, čime se formiraju klasteri; drugo, centri klastera se ažuriraju tako što se postavljaju na srednje vrednosti tačaka u njihovim klasterima. Ovaj proces se ponavlja dok se centri klastera ne stabilizuju ili dok se ne postigne unapred određen broj iteracija. Konačni rezultat je skup klastera gde su tačke unutar istog klastera što sličnije jedna drugoj, dok su tačke između različitih klastera što različitije. K-means je efikasan i jednostavan algoritam, ali može biti osetljiv na početno postavljanje centara klastera i može imati poteškoće sa podacima koji imaju nepravilne oblike ili različite veličine klastera.

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

Clustered Instances

0       36978 ( 17%)
1      185645 ( 83%)


Class attribute: classification
Classes to Clusters:

      0      1  <-- assigned to cluster
   3852 107460 | BENIGN
  33126  78185 | ATTACK

Cluster 0 <-- ATTACK
Cluster 1 <-- BENIGN

Incorrectly clustered instances :       82037.0  36.8502 %
```