20501105 Özge Nur Ergün

Online Karar Ağaçları ile Ensemble Oluşturma

Proje kapsamında Python programlama dili ve Colab platformunda çalışılmıştır. Klasik karar ağacı yapısında düzenleme yapılarak eğitim verilerini incremental olarak ele alan online bir ağaç oluşturulmuştur. Tekil online ağaç ve online ağaçlardan oluşan bir ensemble 36 UCI sınıflandırma veri seti üzerinde test edilerek model performansları ölçülmüş, klasik ensemble'ların performansları ile karşılaştırılmıştır.

Online ağaca eğitim verileri teker teker beslenmektedir. Beslenen ilk örnek geçici olarak leaf oluşturur. Örnekler sıra ile beslenmeye devam edilerek farklı sınıftan bir örnek geldiğinde bu leaf node'a çevrilir. Yeni gelen örneği en iyi ayrıştıracak attribute belirlenir. Information gain hesabı sonucu belirlenen bu attribute'e göre node iki dala ayrılır. Gelen diğer örneklerin ağaçta hangi yaprağa düştüğü belirlenir, yaprağın sınıfı örneğin sınıfı ile aynıysa ağaca yeni bir node eklenmez. Eğer ağaç gelen örneği yanlış sınıflandırıyorsa örneğin düştüğü leaf node'a çevrilir ve information gain'e göre node 2 dala bölünür. Veri setindeki tüm örnekler sıra ile bu şekilde ağaca beslenerek online bir karar ağacı oluşturulmuş olur.

Yeni örnek geldikçe ağaç büyüdüğünden dolayı örneklerin veriliş sırası ağacın kural setini değiştirmektedir. Bu bilgiden yola çıkılarak aynı veri setinin farklı sıralarda verildiği 10 adet online ağaç oluşturulmuş ve bu ağaçların çıktıları demokrasi usulü birleştirilerek tahminleme yapılmıştır.

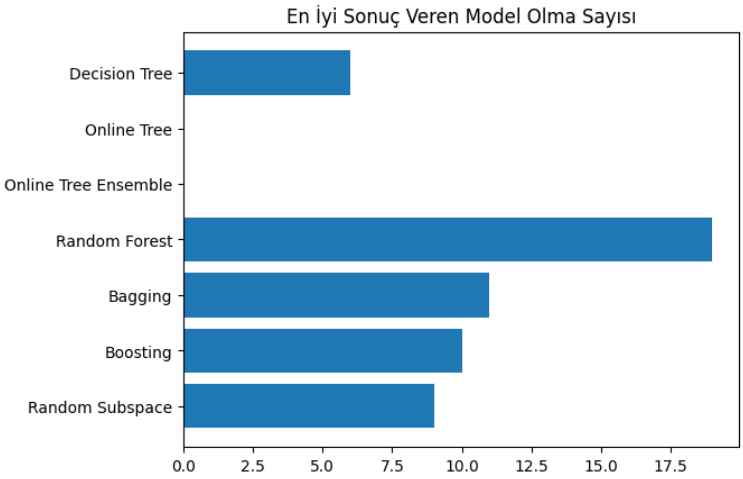
Oluşan base learner'lar incelendiğinde örneklerin veriliş sırasının ağaçta büyük değişikliklere yol açabildiği görülmüş olup buradan yola çıkılarak base learner'ların karar çeşitliliğinin yüksek olduğu söylenebilir. Kolektif öğrenmenin iki önemli noktasından biri olan çeşitlilik sağlanmış olsa da base learner'ların bireysel başarılarının düşük olduğu görülmüş, bu nedenle ensemble sonuçları beklenen seviyede başarılı olmamıştır.

Aşağıda modellerin örnek bazı verisetleri üzerindeki accuracy değerlerine yer verilmiştir.

Dataset	**Decision Tree**	**Online Tree**	**Online Tree Ensemble**	**Random Forest**	**Bagging**	**Gradient Boosting**	**Random Subspace**
zoo	**100.00**	94.40	98.69	**100.00**	**100.00**	**100.00**	**100.00**
primary-tumor	34.21	35.66	40.78	44.74	39.47	**48.68**	47.34
labor	66.66	72.27	77.51	**93.33**	80.00	66.66	73.33
ionosphere	86.36	74.17	80.54	**94.31**	**94.31**	92.00	92.04
heart-statlog	75.00	60.04	62.67	**80.88**	73.53	76.47	**80.88**
credit-a	75.72	52.38	54.83	**85.55**	84.93	84.97	83.24
anneal	**100.00**	98.78	99.29	**100.00**	99.55	99.10	**100.00**
balance-scale	75.79	68.74	72.83	78.98	78.98	**83.43**	82.80

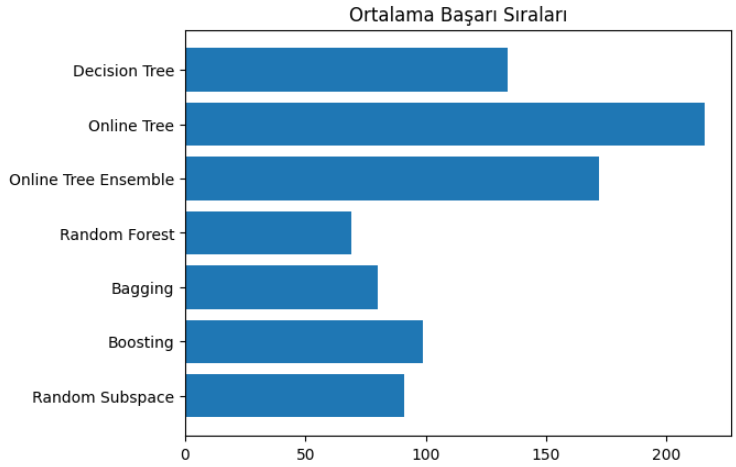
Aşağıda 36 veri seti üzerinde hangi modelin kaç kere en iyi model olduğuna dair bilgiler verilmektedir.

Model	En İyi Model Olduğu Veri Seti Sayısı
Decision Tree	6
Online Tree	0
Online Tree Based Ensemble	0
Random Forest	19
Bagging	11
Boosting	10
Random Subspace	9



Aşağıda modellerin ortalama sıralamaları verilmektedir.

Model	Ortalama Sıralama
Random Forest	69
Bagging	80
Random Subspace	91
Gradient Boosting	99
Decision Tree	134
Online Tree Based Ensemble	172
Online Tree	216



SONUÇLAR

- Test sürecinde kullanılan modeller içerisinde en başarılı olan random forest olmuştur. En başarısız ise online tekil ağaç olarak görülmektedir.

- Verinin veriliş sırası online ağacın yapısını değiştirdiğinden dolayı online ağaçlardan oluşturulan ensemble'daki base learner'ların karar çeşitliliği yüksektir. Bu durum, tekil online ağaçlara kıyasla ensemble'ın daha başarılı olmasına neden olmuştur.

- Online ağaçlar veri setinin geneline bakarak kural oluşturmadığı için tüm veri setini genelleyebilecek kurallar oluşturulamamaktadır. Eğitim veri setindeki her örneği doğru sınıflandıracak şekilde ağacı büyütmek overfitting'e yol açabilmektedir.

- Yanlış sınıflandırılan her örnek için yeni bir node eklendiğinden dolayı online ağaçlar klasik karar ağaçlarına kıyasla daha büyük olmaktadır.

- Özellikle ağacın ilk node'ları oluşturulurken denk gelen bir outlier ağacın başarısının düşük olmasına yol açacaktır. Buradan yola çıkılarak online ağaçların gürültüye karşı hassas oldukları söylenebilir.

- Outlier içermeyen veya train – test verileri arasında paralellik bulunan veri setlerinde online ağaçların başarılı sonuç vermesi mümkündür.

- Online ağaçların en büyük avantajı sonradan toplanan verilerin sıfırdan eğitim sürecine gerek olmadan doğrudan modele dahil edilebilmesidir.