|  |
| --- |
| **YILDIZ TEKNİK ÜNİVERSİTESİ – BİLGİSAYAR MÜHENDİSLİĞİ** |
| **Kolektif Öğrenme 3. Ödevi** |
| **Artımsal karar ağaçlarından oluşan bir topluluğun normal karar ağaçlarından oluşan toplulukla karşılaştırılması** |

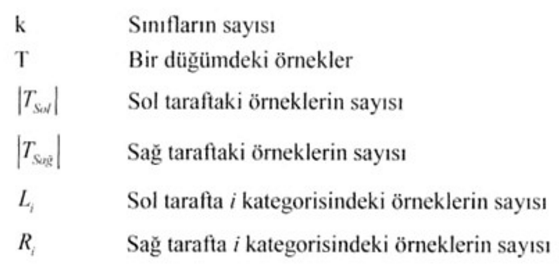
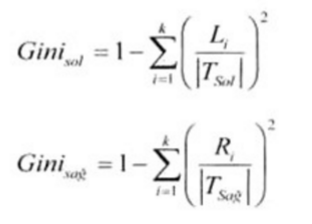
|  |
| --- |
| Melike Nur Mermer - 15501010  12.12.2016 |

**Giriş:**

Bu çalışma örneklerin geliş sırasının farklı olduğu durumlarda farklı modeller üretilen artımsal karar ağaçlarından oluşan bir topluluk ile örnek sırasına bakılmaksızın eğitilen klasik karar ağaçlarından oluşan bir topluluğun karşılaştırmasını konu almaktadır. Her iki algoritmada da temel öğrenicilerin eğitim kümeleri Bagging ile oluşturulmuş aynı eğitim kümeleridir. Artımsal ağaçlardan oluşan topluluk bu örnekleri teker teker ele alarak büyüyerek tüm eğitim örneklerini doğru sınıflandıracak şekilde her yaprakta sadece bir sınıfa ait örneklerin bulunduğu full treelerden oluşmaktadır. Normal karar ağaçlarından oluşan topluluk ise eğitim kümesinin tamamının birden verildiği ve en ayırt edici özellikten başlayarak düğümlerin oluşturulduğu Gini algoritmasıyla eğitilmiş full treelerden oluşmaktadır. Karşılaştırma sınıflandırma veri setlerinde gerçekleştirilmiş ve her iki topluluk için kararlar demokrasi usulü ile birleştirilmiştir. 10x2 fold CV ile her iki yöntemde elde edilen hata oranları ttest ile karşılaştırılmış ve sonuçları verilmiştir.

**Seçilen karar ağacı algoritması (Gini):**

Klasik ve artımsal karar ağaçlarının eğitiminde Gini algoritması kullanılmıştır. Gini algoritması veri setini her bir özellik için özellik değerlerine göre sol ve sağ olarak iki gruba ayırır ve her özellik için Ginisol ve Ginisağ değerleri aşağıdaki bağıntıya göre hesaplanır.



Elde edilen sol ve sağ Giniler ile her özelliğin toplam Ginisi aşağıdaki bağıntıdan bulunur.



Her özellik için hesaplanan Gini değerlerinden en küçük olanı seçilir ve düğümdeki veriler bu özelliğe göre belirlenen gruplara ayrılır. Ayrılan gruplarda özellikler için tekrar sağ ve sol Gini hesaplanarak toplam Ginisi en düşük özellik belirlenir. Tüm yapraklarda sadece aynı sınıfa ait örnekler bulunana kadar ağaç büyümeye devam eder.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **özellik1** | **özellik2** | **özellik3** | **özellik4** | **sınıf** |
| **1** | 1 | 1 | 0 | 1 | 1 |
| **2** | 1 | 1 | 0 | 0 | 2 |
| **3** | 1 | 0 | 0 | 1 | 2 |
| **4** | 0 | 0 | 1 | 1 | 1 |
| **5** | 1 | 0 | 0 | 0 | 1 |
| **6** | 0 | 1 | 0 | 1 | 2 |
| **7** | 0 | 1 | 0 | 0 | 2 |
| **8** | 1 | 0 | 1 | 1 | 1 |
| **9** | 0 | 1 | 1 | 1 | 1 |
| **10** | 1 | 0 | 1 | 0 | 1 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **özellik1** | | **özellik2** | | **özellik3** | | **özellik4** | |
| **sınıf** | **0** | **1** | **0** | **1** | **0** | **1** | **0** | **1** |
| **1** | 2 | 4 | 4 | 2 | 2 | 4 | 2 | 4 |
| **2** | 2 | 2 | 1 | 3 | 4 | 0 | 2 | 2 |

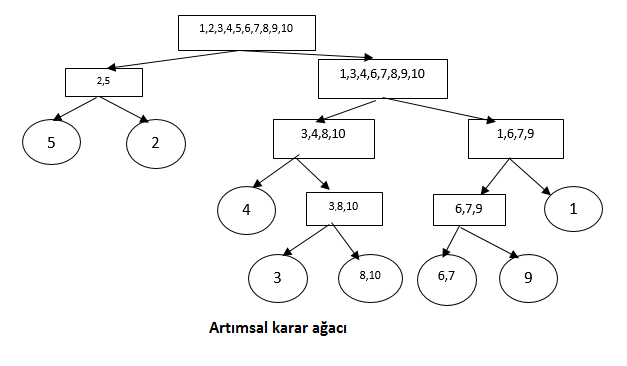
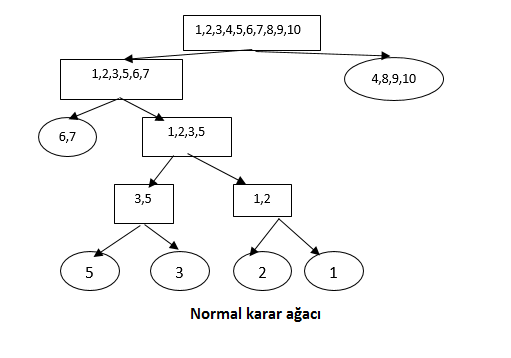
Gini algoritmasının artımsal ve normal yöntemlerde uygulanması sonucu oluşan karar ağaçlarını incelemek üzere küçük bir veri seti oluşturulmuştur. Bu veri seti ile oluşturulan karar ağaçları bir sonraki bölümde verilmektedir.

**Artımsal karar ağaçları:**

Örneklerin tamamının tek seferde verilmediği, yeni örnekler geldikçe modelin bu örnekler için de doğru sonuç verecek şekilde güncellendiği algoritmalar online algoritmalardır. Artımsal karar ağaçları da yeni gelen eğitim verilerine göre kendini güncelleyen bir online algoritmadır. Bu çalışmada kullanılan artımsal ağaç modeli Gini algoritmasını kullanarak en uygun özellik-eşik ikilisinin belirlendiği klasik bir karar ağacı algoritmasından yola çıkılarak oluşturulmuş bir modeldir. Bu artımsal ağaç modelinin ilk düğümünün oluşturulabilmesi için farklı sınıflardan iki örneğin ağaca verilmesi gerekmektedir. Bu sebepten Bagging ile oluşturulmuş eğitim setinin ilk iki örneği aynı sınıftansa 2. örnek sıradaki ilk farklı sınıfa ait örnekle yer değiştirir. Bu sayede eğitim kümesinin ilk iki örneğini alan artımsal ağaç bu iki örneği ayıracak en iyi özellik-eşik ikilisini belirleyerek ilk düğümünü oluşturur ve 3. örnekten başlayarak örnekleri bu ağaçta sınıflandırır. Ağacın gelen örnek için belirlemiş olduğu sınıf eğer gerçekten ait olduğu sınıfsa bir değişiklik yapılmaz ancak değilse ağaç yeni örneği doğru sınıflandıracak şekilde güncellenir. Bu güncelleme sırasında şu adımlar gerçekleştirilir.

1. Örneğin ağaçta düştüğü yaprak numarası belirlenir.
2. Bu yaprakta olan diğer örneklerle birlikte tüm örneklerin en iyi ayıracak özellik-eşik ikilisi Gini ile hesaplanır.
3. Yaprak yerine bir düğüm oluşturulur ve örnekler her yaprakta sadece aynı sınıfın örnekleri mevcut olacak düğümden çıkan yapraklara şekilde (yeni gelen örnek bir tarafa önceden bu yaprakta olanlar diğer tarafa) yerleştirilir.
4. Eski yaprak yerine yeni oluşturulan düğüm, numarası ile bulunarak ağaçtaki yerine yerleştirilir.

Örneklerin tamamı bu şekilde ağacı güncelleyerek sınıflandırılır. Sonuçta elimizde eğitim örnekleri için hatası 0 ve aslında sadece bu örnekler için oldukça özelleşmiş bir ağaç olur. Bunu göstermek için 2 sınıflı 10 örnekli küçük bir data ile normal ve artımsal karar ağaçları oluşturularak incelenmiştir.



10 örnekli veri için oluşturulan karar ağaçlarına bakıldığında normal karar ağacının artımsal ağaçtan daha az dal ve yapraktan oluştuğu söylenebilir. Artımsal karar ağacında sıradan gelen her örnek için eğer yanlış sınıfa düşmüşse ona özel yeni bir kural tanımlanır. Ağaç bu şekilde büyüyünce gerek duyulan karar mekanizması sayısı daha fazla olacaktır. Ağacın büyük olması karar sınırının çok hassas olması anlamına gelir. Eğitim örnekleri için çok hassaslaşmış bir öğrenicinin ise aşırı eğitilmiş (overfit) olması söz konusudur. Bu da hiç görmediği test örnekleri için elde edeceği başarının düşük olmasına sebep olur. Klasik yöntemde tüm veri için en ayırıcı özelliğin kök düğüme yerleştirilmesi bir şarta bakarak genelleştirme yapmak anlamına gelmektedir. Bu genelleştirme özelliği sayesinde klasik karar ağacı test örneklerinde daha başarılı olacaktır.

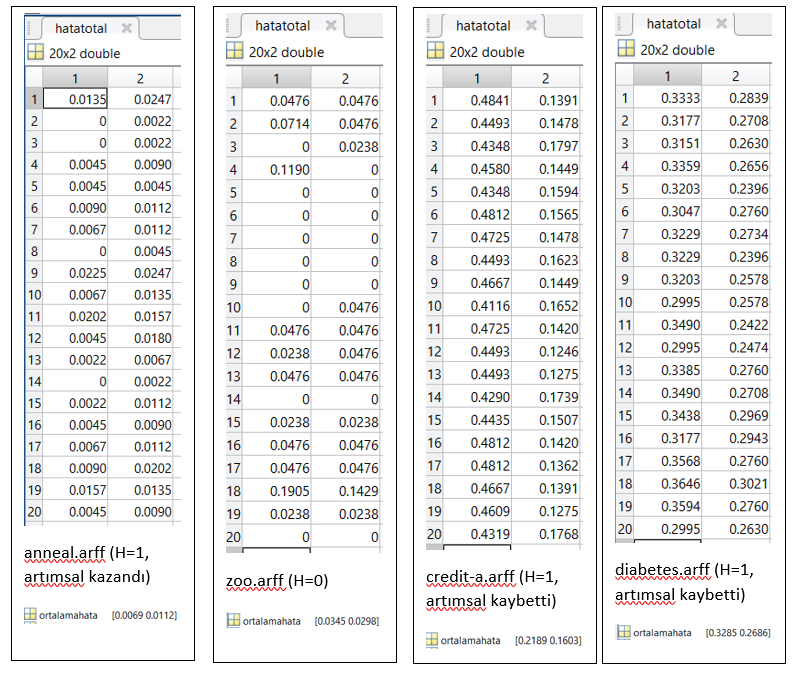
**Toplulukların oluşturulması ve kararların birleştirilmesi:**

Yukarıda bahsedilen yöntemler kullanılarak biri klasik yöntemle oluşturulmuş 10 tane ağaçtan diğeri artımsal yöntemle oluşturulmuş 10 tane ağaçtan meydana gelen 2 topluluk eğitilmiştir. Test aşamasında her iki topluluğun kendi içerisindeki kararları her bir temel öğrenicinin ortak verilecek kararda tek bir oy verme hakkının olduğu demokrasi usulü ile birleştirilmiştir. Tahminlerin birleştirilmesi için sınıf sayısı kadar elemanı olan bir dizi oluşturulmuştur. Her bir öğrenicinin verdiği oylarla bu dizinin ilgili sınıfa karşılık gelen indisteki elemanı artırılarak her sınıfa verilen toplam oylar belirlenir. Bu dizinin en büyük elemanının indisi ise kazanan sınıfı belirtmektedir.



**Test sonuçları:**

Ortak verilen kararın test örneğinin sınıfı ile aynı olup olmadığına bakılarak hata oranı belirlenmiştir. 10x2 fold CV ile 30 dataset üzerinde elde edilen hata oranları ile ttest yapılarak iki algoritma arasında istatistiksel açıdan fark olup olmadığı gözlenmiştir. Bazı datasetlerde elde edilen hata oranları ve ttest sonuçları aşağıda verilmektedir.



30 dataset üzerinde yapılan ttestler sonucu algoritmaların kazandığı/berabere kaldığı/kaybettiği datasetlerin sayıları aşağıdaki tabloda verilmektedir.

|  |  |  |
| --- | --- | --- |
|  | **Artımsal Ensemble** | **Normal Ensemble** |
| **Artımsal Ensemble** | **-** | **28/1/1** |
| **Normal Ensemble** | **1/1/28** | **-** |

**Sonuçlar:**

Yapılan testler sonucunda artımsal karar ağaçlarının genel kural oluşturabilme yeteneğinin olmayışı nedeniyle tekil başarılarının normal karar ağaçlarına göre düşük olduğu gözlenmiştir. Topluluk algoritmalarında tekil başarılar düşük olsa da ağaçların çok farklı kararlar üretmesi nedeniyle ortak kararın başarılı olma ihtimali vardır. Artımsal karar ağaçlarının başarılı olabilmesi için test örneklerinin eğitim örneklerine çok benzemesi gerekir. Sadece 1 datasette (anneal.arff) artımsal karar ağaçlarından oluşan topluluk daha başarılı olmuştur ve 1 datasette (zoo.arff) iki topluluk berabere kalmıştır. Bunların dışında kalan veri setleri için test örneklerinde tekil öğrenicilerin başarılarının çok düşük olması nedeniyle artımsal karar ağaçlarından oluşan topluluk başarısız olmuştur.

Artımsal ağaçların tekil başarılarını artırmak için budama yapılabilir. Ağaç büyürken yapılacak budama gürültülü örneklerin etkisini en aza indirgeyeceğinden başarıyı artırabilir.