DUFED 5(2) (2016) 77-80

Dicle Üniversitesi Fen Bilimleri Enstitüsü Dergisi

dergi anasayfa: http://www.dufed.org

C4.5 Karar ağaçlarında genetik algoritma ile budama

C4.5 Decision tree pruning using genetic algorithm

Abdülkadir Gümüşçü[1], Ramazan Taşaltın[1], İbrahim Berkan Aydilek[2]

[1] *Harran Üniversitesi, Mühendislik Fakültesi, Elektrik-Elektronik Mühendisliği Bölümü, Şanlıurfa, Türkiye*
[2] *Harran Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Şanlıurfa, Türkiye*

MAKALE BİLGİSİ

Geliş Tarihi: 10 Aralık 2015
Revizyon Tarihi: 26 Ocak 2016
Kabul Tarihi: 27 Ocak 2016
Elektronik Yayın Tarihi: 23 Kasım 2016
Basım: 23 Aralık 2016

ÖZET

Karar ağaçları sınıflandırma ve değer tahmini amacıyla kullanılan makina öğrenme algoritmalarından biridir. Karar ağaçlarını oluşturmak amacıyla birçok yaklaşım önerilmiştir. Bu yaklaşımlardan biri olan C4.5 karar ağaçları metodu birçok alanda sıklıkla kullanılmaktadır. Ağaç yapısını kurmada kullanılacak veri setinin nitelik sayısının fazla olması, ağaç yapısında gereksiz dallar ve düğüm noktalarına sebep olmaktadır. Bunun sonucunda gereksiz oluşturulan dallar ve düğüm noktaları aşırı öğrenmeye, aşırı öğrenme ise sınıflandırma başarı oranını olumsuz yönde etkilemektedir. Bu çalışmada aşırı öğrenmenin etkilerini azaltmak için yeni bir budama algoritması önerilmiştir. WEKA ortamında çalıştırılan C4.5 algoritmasının Güven Faktörü (Confidence Factor) genetik algoritma ile optimize edilerek başarılı sonuçlar elde edilmiştir.

Anahtar sözcükler: Genetik algoritma, Karar ağacı, Budama

ABSTRACT

Decision tree is a machine learning algorithm that is used for classification and regression. Many approaches were proposed to build decision trees. C4.5 decision tree that is one of these approaches, is frequently used in many fields. Large number of attributes of the data set that is used for building decision tree causes unnecessary branches and nodes on decision tree. Unnecessary branches and nodes cause overfitting. Overfitting negatively affects classification success rate. In this paper, a novel pruning algorithm is proposed to reduce the effects of overfitting. Successful results were obtained by optimizing confidence factor (CF) of C4.5 algorithm executed in Weka using genetic algorithm.

Keywords: Genetic algorithm, Decision tree, Pruning

1. Giriş

Karar ağaçları sınıflandırma ve değer tahmini amacıyla kullanılan makina öğrenme algoritmalarından biridir. Karar ağacını oluşturmak amacıyla bir çok yaklaşım önerilmiştir (1,2). Bu çalışma C4.5 karar ağaçlarında sınıflandırma problemi üzerine yoğunlaşacaktır. Genel olarak en uygun ağaç yapısını oluşturmak için iki işlem adımını takip etmemiz gerekmektedir. Birincisi eğitim veri seti ile ağaç yapısını kurmak, ikincisi ise kurulan ağaç yapısında budama işlemini yürütmektir. Veri setlerinde nitelik sayısı arttığında kurulacak olan ağaç gereksiz düğüm noktalarını da oluşturmaktadır. Aşırı öğrenme olarak adlandırılan bu durum başarı oranını olumsuz yönde etkilemektedir. Budama işlemi genellikle aşırı öğrenmeden kaynaklı olumsuz etkileri yok etme amacıyla uygulanmaktadır. Weka ortamında C4.5 algoritması için güven katsayısı (CF) adında bir parametre istenmekte ve bu katsayı budama işleminin boyutunu belirlemektedir. CF katsayısı [0,1] arasında değişen bir değer almakta ve 0'a yaklaştıkça budama boyutu küçülmekte, 1'e yaklaştıkça budama boyutu artmaktadır. Bu çalışmada budama işlemini daha efektif kullanma adına Weka geliştirme ortamında bulunan güven parametresi genetik algoritma ile optimize edilerek adaptif bir budama işlemi önerilmiş ve kayda değer iyileştirmeler sağlanmıştır.

Yazışma adresi: Abdülkadir Gümüşçü
E-posta: agumuscu@harran.edu.tr

Makalenin ikinci bölümünde çalışmamız ile ilgili literatür özeti sunulacaktır. Üçüncü bölümde ise karar ağaçları, karar ağaçlarında budama ve genetik algoritmanın temelleri açıklanacaktır. Dördüncü bölümde önerilen metod anlatılacak ve kullanılan veri setleri detaylandırılacaktır. Beşinci bölümde ise sonuçlar sunulacak olup sonuçlar üzerinden yapılacak tartışma ve gelecekte konu ile ilgili yapılabilecek çalışma önerileri ile sonlandırılacaktır.

2. Literatür

Budama iki ana başlık altında açıklanabilir. Bunlar ön budama ve son budama olarak adlandırılırlar. Bu çalışmada son budama üzerine önerilmiş bir metot önerilmektedir. Son budama işleminde eğitim veri setine göre karar ağacı tam olarak kurulduktan sonra budama yapılır. Genel olarak son budama ön budamaya göre daha iyi sonuçlar vermektedir.

Literatürde birçok karar ağacı budama algoritması önerilmiştir (3-7). Breiman maliyet-karmaşıklık budama (CCP) [3] adında çok kullanılan bir metot önermiştir. Bu çalışma son budama metodu önermektedir ve bu metot ağaç kombinasyonlarından en küçük ve en düşük hata oranına sahip ağacı seçme temeline dayanmaktadır. Bu metot tüm ağaç kombinasyonlarını oluşturduktan sonra seçme işlemini yaptığından çok fazla niteliğe sahip veri setlerinde metodun işlenme süresini ciddi derecede arttırmaktadır. Önerdiğimiz metot sadece ideal güven katsayısını araştırdığından daha az süre gerektirmektedir. Benzer şekilde, En Düşük Hata Budama metodu (MEP), Niblett and Bratko tarafından önerilmiştir (4). MEP modelinde ise, alttan ve üstten arama modelini en düşük hata oranını verecek budanmış ağacı bulmak için kullanılmıştır. Azaltılmış Hata Budaması (REP) ve Kötümser Hata Budaması (PEP) metotları J.R. Quinlan tarafından önerilmiştir (5). Bunların yanında J.Chen et al. Genetik algoritma kullanarak budama metodu önermiştir (6). Bu metot ise en küçük ağaç yapısını en yüksek başarı oranı ile aramak için genetik algoritmayı kullanmıştır. Bu metot genetik algoritmayı dalların olup olmaması üzerinde belirlemek amacı ile kullanmaktadır. Dolaysıyla bu metodun çok fazla niteliğe sahip veri setleri için kullanılması işlem süresini arttıracaktır. Önerdiğimiz metot ise sadece güven katsayısını optimize ettiğinden daha az işlem gerektirmektedir. Esposito et al. ise birçok budama algoritmasını analiz etmiş ve detaylı şekilde karşılaştırmıştır (7).

3. Yöntem

3.1. C4.5 Karar Ağaçları ve Budama

Karar ağaçları pek çok alanda kullanılmaya başlanan bir sınıflandırma algoritmasıdır. C4.5 ise bir karar ağacı sınıflandırma algoritmasıdır. Bu çalışmada J. Ross Quinlan tarafından geliştirilmiş C4.5 sınıflandırma algoritması [1] üzerinde yeni bir budama algoritması önerilecektir. Genel olarak karar ağaçları, anlaşılması ve yorumlanması kolay olduğundan kullanımı yaygındır. C4.5 iki işlem adımı ile gerçekleştirilmektedir. Bunlardan ilki ağacı oluşturma işlemi ve diğeri ise budama işlemidir. Bu çalışmada budama işlemi için bir algoritma önerilmiştir.

Bir karar ağacı yapısı kök, düğüm, dal ve yaprak'tan oluşur. Ağaç yapısında en altta kalan kısım yaprak en üstte olan kısıma ise kök adı verilir. Veri setinde bulunan her bir nitelik ise düğüm noktalarını temsil etmektedir. Düğümler arası bağlantıyı sağlayan kısım ise dal olarak adlandırılır.

Hangi nitelik değerine göre dallanmanın gerçekleşeceğine karar verme, karar ağaçlarını oluşturmadaki en önemli işlem adımı olarak değerlendirilebilir (8). Karar verme kıstasları olarak Bilgi kazancı [1], Gini indeksi [3] ve Towing kuralı [3] yaygın olarak kullanılmaktadır. Bu çalışmada önerilen algoritma için bilgi kazancı karar verme kıstası olarak kullanılmıştır. Bu yönteme göre her nitelik için entropi bazlı bir değer ile ilgili niteliğin sonuca etkisi hesaplanmaktadır.

η adet sınıfa sahip olunduğu ve bu sınıf değerlerinin de T kadar tekrar edildiği varsayılırsa bir sınıfa ait olasılık değeri;

$$P_i = \frac{c_i}{|T|} \quad \textbf{(1)}$$

şeklinde hesaplanabilir. C_i bir sınıfa ait sınıf değerlerinin sayısını temsil etmektedir. Bu sınıflara ait entropi değeri $H(T)$ ise;

$$\mathrm{H(T)} = -\sum_{i=1}^{n} P_i \log_2 P_i \quad \textbf{(2)}$$

şeklinde hesaplanır. Veri setinde Y nitelik değerlerine göre T sınıf değerleri $T_1, T_2 ..., T_n$ şeklinde alt kümelere ayrıldığı göz önünde bulundurulduğunda Y nitelik değerleri kullanılarak T sınıf değerlerinin bölünmesi sonucunda elde edilecek bilgi kazancı $IG(Y,T)$);

$$IG(Y,T) = H(T) - \sum_{i=1}^{n} \frac{|T_i|}{|T|} H(T_i) (3) \quad \textbf{(3)}$$

şeklinde hesaplanır. kümesi içinniteliğinin değerinin belirlenmesinde ayrılma bilgisi;

$$SI(Y) = -\sum_{i=1}^{n} \frac{|T_i|}{|T|} \log_2 \left(\frac{|T_i|}{T} \right) (4) \quad \textbf{(4)}$$

şeklinde hesaplanır. Bilgi kazancının ayrılma bilgisi'ne oranı bize ilgili niteliğin ayrılmasının ne kadar bilgi kazancı sağlayacağını verir. Bu şekilde her nitelik için kazanç bilgisi hesaplanarak en yüksek kazanç bilgisine sahip niteliğe göre ağaç yapısı ayrılır.

Karar ağaçlarının yapısı oluşturulurken diğer bir önemli işlem adımı ise budama işlemidir. Budama işlemi iki şekilde yapılabilmektedir (9). Ağaç yapısı oluşturulurken belli bir oranda ağaç büyüyünce ağacın daha fazla büyümemesi için bölünmeyi durdurmak ön budama olarak adlandırılmaktadır. İkinci olarak, ağaç tamamen oluşturulduktan sonra oluşturulmuş olan bölünme noktalarını çıkararak budama yapma işlemi ise son budama olarak adlandırılmaktadır. Bu çalışmada önerilen metot son budama algoritmaları sınıfına girmektedir. Bu çalışmada Weka ortamında çalıştırılan C4.5 algoritmasının Güven Faktörü (Confidence Factor) optimize edilerek başarılı sonuçlar elde edilmiştir.

3.2. Genetik Algoritma

1975'te John Holland tarafından geliştirilen Genetik Algoritma, doğal seçim ilkelerine dayanan bir arama ve optimizasyon yöntemidir. Genetik algoritmalar, birçok alanda etkili bir arama tekniği olarak kabul edilmektedir (10). Genetik algoritmalar, başlangıç popülasyonu ve bu popülasyondan üreme, çaprazlama ve mutasyon gibi metotlar kullanarak yeni popülasyonlar oluşturmaya dayanır(11). Bu çalışmada genetik algoritma en iyi güven katsayısını bulmak amacı ile kullanılmıştır.

4. Önerilen Yöntem

Bu makalede önerilen Genetik Algoritma ile Son Budama (GASB) metodu, UCI veritabanında (12) bulunan 4 adet veri seti ile test edilmiştir. Kullanılan veri setlerinin özellikleri Tablo 1'de verilmiştir. Ayrıca Tablo 1'de veri setinin zorluk derecesi hakkında bilgi verecek olan Nitelik Sayısı (NS) / Örnek Sayısı (ÖS) şeklinde tanımlanan bir katsayı da verilmiştir. Bu katsayı nitelik sayısının, örnek sayısına oranını temsil etmektedir.

Tablo 1: Kullanılan veri setlerinin özellikleri.

Veri Seti	Örnek Sayısı	Nitelik Sayısı	NS/ÖS
IRIS	150	4	0.026
PIMA	768	18	0.023
GLASS	214	10	0.046
WINE	178	13	0.073

C4.5 karar ağaçlarında eğer tüm nitelikler ağaç yapısına dâhil edilirse; test aşamasında bazı örnekler için yanlış sonuçlar çıkabilmektedir. Bunun sebebi aşırı öğrenme olarak bilinmektedir. Bu çalışmada aşırı öğrenmeyi önleyen son budama işlemini daha etkin kullanmak için CF katsayısı optimize edilmiştir. Weka yazılım ortamında C4.5 algoritması için Güven Faktörü (CF) parametresi kullanılmaktadır. Bu parametre ağacın gelişiminden sonra ağacın budaması işleminin derecesini belirlerken değeri de 0 ile 1 arasında değişmektedir. CF katsayısını azaltmak; son budamanın küçülmesine, CF katsayısı arttırmak ise; son budamanın büyümesine sebep olur.

Önerilen metoda göre ilk önce veri seti niteliklerine göre C4.5 karar ağacı kurulmuştur. Son budama işlemi için Genetik Algoritma ile CF katsayısı belirlenmiş ve ağaç seçilen CF katsayısı ile budanmıştır. Budama işleminden sonra 10 katlı çapraz doğrulama işlemi ile sınıflandırma işlemi doğruluk oranları hesaplanmıştır.

5. Sonuç ve Tartışma

Bu çalışmada, C4.5 karar ağaçları için yeni bir son budama yöntemi sunulmuştur. Önceki Weka ortamında yapılan C4.5 karar ağacı algoritmalarında sabit bir CF katsayısı kabul edilmiştir. Bu çalışmada ise CF katsayısı genetik algoritma ile iyileştirilerek daha iyi başarı oranları elde edilmiştir. Önerilen algoritmanın sonuçları Tablo 2'de verilmiştir.

Tablo 2: Budamasız karar ağacı ile Önerilen metot ile budanmış karar ağacı sınıflandırma başarı oranları karşılaştırma tablosu.

Veri Seti	NS/ÖS	Budamasız Başarı Oranı	Önerilen Metot Başarı Oranı
IRIS	0.026	94.20%	91.98%
PIMA	0.023	69.10%	75.38%
GLASS	0.046	59.40%	69.70%
WINE	0.073	91.30%	92.15%

Önerilen metot PIMA, GLASS ve WINE veri setleri kayda değer bir iyileştirme sağlamıştır. Fakat IRIS veri seti için aynı durum söz konusu değildir. Bunun sebebi IRIS

veri setinin sadece 4 özellik içermesidir. 4 özellik bulunan IRIS veri seti için kurulacak ağaç yapısında bulunacak dallanma sayısı en fazla 4 olabilecektir. Son budama işlemi de dallanma sayılarını azaltacağından özellik sayısı az olan veri setlerinde bilgi kaybı yaşanmaktadır. Bundan ötürü budama işlemi yapmak başarı oranına olumlu etki yapmamaktadır. Sonuçlardan da görüldüğü üzere genel olarak budama yapılmadan başarı oranı yüksek olan veri setleri için budama işlemleri fazla bir katkı sağlamamaktadır.

Daha önceki çalışmalarda kullanılan diğer bir yöntem ise budamayı tamamen genetik algoritma ile yapma metodudur (6). Bu metot sonuçları ile Önerilen Metot sonuçları Tablo 3'de kıyaslanmıştır.

Tablo 3: Genetik Algoritma (6) ile budanmış karar ağacı ile Önerilen metot ile budanmış karar ağacı sınıflandırma başarı oranları karşılaştırma tablosu.

Veri Seti	NS/ÖS	(6) Başarı Oranı	Önerilen Metod Başarı Oranı
IRIS	0.026	94.20%	91.98%
PIMA	0.023	68.60%	75.38%
GLASS	0.046	60.40%	69.70%
WINE	0.073	91.30%	92.15%

Tablo 3'de görüldüğü üzere (6)'da önerilen metot sadece IRIS veri setinde daha iyi başarı oranını sağlamıştır. PIMA, GLASS ve WINE veri setlerinde (6)'da önerilen metoda kıyasla iyileştirme sağlanmıştır. Bu iyileştirmenin sebebi budama işleminin en ideal şekilde yapılması olarak kabul edilebilir. (6)'da önerilen metottaki değişken fazlalığı, hem genetik algoritmanın işlem yüküne hemde başarı oranına olumsuz etki yapmaktadır. Önerdiğimiz metotta sadece güven katsayısının optimize edilmesiyle budamanın hacminin belirlenmesi, hem işlem yükünü hem de sınıflandırma başarısını olumlu etkilemiştir.

Gelecekte bu çalışma ile ilgili özellik sayısı kısıtlı olan veri setleri için iyileştirmeler yapılabilir. Ayrıca önerilen metot farklı veri setlerine uygulanarak sınıflandırma başarı oranları kıyaslanabilir.

Kaynaklar

1. J. R. Quinlan, C4.5: Programs for Machine Learning: Morgan Kaufmann, 1993.
2. J. R. Quinlan, "Induction of decision trees," Machine Learning, vol. 1, pp. 81-106, 1986.
3. Breiman L, Friedman J, Olshen R, Stone C (1984) Classification and regression trees. Wadsworth International, Belmont.
4. Niblett T, Bratko I (1986) Learning decision rules in noisy domains. In: Proceedings of expert systems'86. Cambridge University Press, New York, pp 25–34.
5. J. R. Quinlan, "Simplifying decision trees," Int. J. Hum.-Comput. Stud, vol. 51, pp. 497-510, 1999.
6. Jie Chen, Xizhao Wang, Junhai Zhai, "Pruning Decision Tree Using Genetic Algorithms" International Conference on Artificial Intelligence and Computational Intelligence, 2019, pp 244–248.
7. Esposito F, Malerba D, Semeraro G (1997) A comparative analysis of methods for pruning decision trees. IEEE Trans Pattern Anal Mach Intell 19(5):476–491.
8. T. Kavzaoğlu, İ. Çölkesen, "Karar Ağaçları İle Uydu Görüntülerinin Sınıflandırılması: Kocaeli Örneği",Harita Teknolojileri Elektronik Dergisi , vol. 2, no:1, pp. 36-45, 2010.
9. Quinlan J.R., 1987, "Simplifying decision trees", International Journal of Man-Machine Studies, 27, 221-234.
10. I.B. Aydilek, A. Arslan, A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm, Information Sciences 233 (2013) 25–35.
11. T. Marwala, S. Chakraverty, Fault classification in structures with incomplete measured data using autoassociative neural networks and genetic algorithm, Curr. Sci. India 90 (2006) 542–548.
12. Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.