SDÜ Orman Fakültesi Dergisi
SDU Faculty of Forestry Journal
2012, 13: 1-4
Araştırma makalesi/Research article

Sınıflandırma ve regresyon ağacı tekniği (SRAT) ile ekolojik verinin modellenmesi

Kürşad Özkan

Süleyman Demirel Üniversitesi, Orman Fakültesi, Isparta

İletişim yazarı/Corresponding author: kursadozkan@sdu.edu.tr , Geliş tarihi/Received: 07.07.2011, Kabul tarihi/Accepted: 08.02.2012

Özet: Orman ekosistemlerinde hedef türlerin yetişme ortamı özelliklerine göre modellenmesi ile ilgili parametrik olmayan yöntemlerin kullanımı gün geçtikçe artmaktadır. Parametrik olmayan yöntemlerden biride sınıflandırma ve regresyon ağacı tekniğidir (SRAT). Bu yöntem kullanılarak hem kategorik (sınıflandırma ağacı) hem de sürekli (regresyon ağacı) bağımlı değişkenler modellenebilmektedir. Bundan dolayı SRAT hayvan ve bitki türlerinin dağılım modellemeleri için ideal bir yöntemdir. Bu derlemede SRAT hakkında bilgi vermek amaçlanmıştır.
Anahtar kelimeler: Parametrik olmayan yöntemler, Tür dağılımı, Tür çeşitliliği, Coğrafi modelleme, Potansiyel dağılım

Modelling ecological data using classification and regression tree technique (CART)

Abstract: Nonparametric methods have been increasingly used in order to model the distribution of target species in the forest ecosystems by means of environmental factors in recent years. One of the nonparametric methods is classification and regression tree tecnique (CART). By using CART, both categorical and continuous dependent variables can be modeled. That is why CART is a suitable tecnique for the distribution, productivity and diversity models of animal and plant species. This review was written to give information about CART.
Keywords: Nonparametric methods, Species distribution model, Species diversity, Spatial modeling, Potantial distribution

1. Giriş

Orman ekosistemlerinde modelleme çalışmaları genelde hedef türlerin yetişme ortamına uygunluğuna, verimliliğine ve tür çeşitliliğine odaklanmıştır. Bu modelleme çalışmalarında çeşitli analitik yöntemlere başvurulmaktadır. Geleneksel veya doğrusal yöntemler olarak tür dağılımında lojistik regresyon analizi ve verimlilik/tür çeşitliliği dağılımında çoklu regresyon analizi ilgi grupları tarafından en fazla bilinenlerdir. Bu yöntemler uzun yıllar boyunca ekoloji alanında çalışan araştırmacılar tarafından kullanılmıştır. Bununla birlikte özellikle son yıllarda parametrik ve doğrusal olmayan, hiyerarşik ve/veya kural tabanlı yöntemlerin kullanımında hızlı bir artış vardır. Ekolojik veri değerlendirme ve modelleme çalışmalarında geleneksel olmayan yöntemlerin seçiminde verinin yapısı büyük rol oynamaktadır. Ekolojik veri genelde karışıktır, dengesizdir, eksiktir, aykırı ve/veya uzak gözlem içerebilmektedir. Ekolojik ilişkiler birçok yerde doğrusal değildir ve muhatap değişkenlerin birçoğu normal dağılım göstermeyebilir. Dahası doğrusal modeller kullanıldığında, bağımsız değişkenler üstünden bir bağımlı değişkenin varyasyonu yüksek derecede açıklansa dahi, özellikle ekolojik araştırmalarda sıklıkla karşılaşılan bağımsız değişenler arasındaki yüksek korelasyondan doğan çoklu bağlantı problemi modelleri geçersiz kılmaktadır.

Ekolojik araştırmalarda geleneksel olmayan yöntemlere olan talep artışının bir diğer önemli sebebi, bu yöntemlere yönelik birçok yazılımın yapılmış ve kullanıma sunulmuş olmasıdır. Bu bağlamda S-PLUS ve DTREG paket programları en fazla tercih edilenlerdir.

Geleneksel olmayan yöntemlerden en fazla tercih edilenlerinden biri sınıflandırma ve regresyon ağacı tekniği (SRAT) olarak isimlendirilmektedir. Bu derleme SRAT hakkında bilgi vermek amacıyla yazılmıştır.

2. Sınıflandırma ve regresyon ağacı tekniği

SRAT parametrik olmayan kural tabanlı bir tekniktir. SRAT'ın temel amacı bağımlı değişkene göre ana veri matrisini (bağımsız değişkenler matrisi) homojen alt gruplara ayırmaktır. Alt grupların oluşturulmasında veri dallanan bir ağaç şeklinde hiyerarşik bir düzende sunulur. Ağaç şekil içindeki ara düğümlerde en iyi ayırımı yapmış olan bağımlı değişkenler gösterilir. Bu düğümlerin dallarında ayırıcı bağımlı değişkenlerin kritik değeri verilir. Yapraklar bağımlı değişkenin değerlerini gösterir. Kök düğüm (ilk düğüm noktası) noktasından itibaren yapraklara kadar (en son düğüm) hatlar bulunmaktadır. Bu hatlar boyunca sınıflar arası ayrımın maksimize edildiği ve her sınıfın içindeki varyasyonun minimize edildiği ayırımların kuralları gösterilmektedir. Bu yaklaşım kullanılarak hem kategorik hem de sürekli bağımlı değişkenler modellenebilmektedir. Eğer bağımlı değişken kategorik ise yöntemin adı sınıflandırma ağacı (SA), sürekli ise regresyon ağacı (RT) ismini alır (Chu vd., 2009; McKenny ve Pedlar, 2003; Navarrate ve Espinosa, 2011; Breiman vd., 1984; Özkan ve Mert 2010; De'ath ve Fabricius 2000).

2.1. Sınıflandırma ağacı

Bitki ve hayvan ekolojisinde hedef türlerin dağılım modellenmesi en önemli konulardan biridir ve sınıflandırma ağacı genelde türlerin dağılım modellemesi için kullanılmaktadır Bu sebepten bağımlı değişken var-yok şeklinde iki kategoriyi içermektedir (Breiman vd., 1984; De'ath ve Fabricius, 2000).

Sınıflandırma ağacı yönteminde ikili bağımlı değişkenin saflığına karar verilirken Gini katışıklık ölçümünü kullanmaktadır.

Bir *t* düğümü için, Gini katışıklık indeksi (*g(t)*) aşağıdaki formülle belirlenmektedir.

$$g(t) = \sum_{j \neq i} p(j \mid t)\, p(i \mid t) \tag{1}$$

Burada *i* ve *j* hedef (bağımlı) değişkenin kategorileridir.

Türlerin dağılımına yönelik modellemelerde bağımlı değişkenlerin ikili (var-yok) kategorilerinden oluştuğundan dolayı indeks için eşitlik aşağıdaki gibidir.

$$g(t) = 2\, p(1 \mid t)\, p(2 \mid t) \tag{2}$$

Düğümdeki bütün yenilenmiş kodlar sadece bir kategoriye ait olduğunda -ki bu durumda düğüm saftır- indeks değeri sıfıra eşittir. Bir düğümün en iyi kestirimini yapacak bağımsız değişkeni bulmak için bağımsız değişken setindeki her değişken değerlendirilir ve en iyi değere sahip olan -ki o katışıklı değerindeki en yüksek azalmayı göstermektedir- değişken seçilir. Herhangi bir *t* düğümü için düğümün bir aday ayırıcısı olan *s* hem sağ taraf ayırımını (t_R) hem de sol taraf ayırımını (t_L) gerçekleştirir.

Değer aşağıdaki formül ile belirlenir.

$$\phi(s,t) = g(t) - p_L g(t_L) - p_R g(t_R) \tag{3}$$

Burada P_R sağ taraftaki bağımlı *t* düğümündeki durumların oranını verirken P_L sol taraftaki *t* düğümündeki durumların oranını vermektedir. Her bir düğümde ikili *s*'in bir aday *S* seti belirlenebilmektedir ve kök düğümü olan t_1 den başlayarak ayırıcı *s** bütün muhtemel *S'*ler arasında daha büyük bir katışıklık azalama değeri ile aranmaktadır.

$$\phi(s^*, t_1) = \max_{s \in S} \phi\,(s, t_1) \tag{4}$$

İdeal bir ayırıcı *s* veri setini $g(t_L)=g(t_R=0$ olacak şekilde iki alt gruba ayırmaktadır. Tekrarlamalı bölme algoritmasi, homojenlik sağlanana ve bazen tek bir durum bir sınıfta kalana kadar devam etmektedir. Bütün nihayi düğümlerin en ideal saflığa gelmesi ile maksimum ağaç elde edilmiş olur.

2.2. Regresyon ağacı

Regresyon ağaçlarında sınıflar yoktur. Bu sebepten regresyon ağacı tekniğinde sınıflandırma ayırım kuralları Gini indeksi kullanılarak uygulanamaz. Regresyon ağacındaki ayırımlar iki sonuçlanan düğüm için tahmin edilen toplam varyansın minimize olmasının gerekliliği anlamına gelen "artıkların karelerini azaltma algoritmasına" göre gerçekleştirilir (Breiman vd., 1984; De'ath and Fabricius, 2000).

Regresyon ağacı yönteminde her düğümde minimizasyon (azaltma) problemi aşağıda gibi çözülür.

$$\arg\min_{xj \le x_j^R,\, j=1,\ldots,M} \left[P_l Var(Y_l) + P_r Var(Y_r)\right] \tag{5}$$

Burada P_l ve P_r sırası ile sol ve sağ düğümlerin olasılıklarıdır. *M* eğitim setindeki değişkenlerin sayılarıdır. Değişken *j* "x_j" olarak gösterilmektedir. x_j^R ise değişken x_j nin en iyi ayırım değerini göstermektedir. $Var(Y_l)$, $Var(Y_r)$ karşılıklı sağ ve sol alt düğümler için sorumlu vektörlerdir. $x_j \le x_j^R, j = 1....,M$ optimal ayırım sorgulaması anlamına gelmektedir.

Artıkların karelerini azaltma algoritması Gini ayırım kurallarına benzemektedir. Eğer sınıf *k*'nın nesneleri değer "1" , diğer sınıfların nesneleri değer "0" ataması yapılır ise, o zaman bu değerlerin örnek varyansı $p(k|t)[1- p(k|t)]$'e eşit olur. Katışıklık ölçümü *i(t)* aşağıdaki yolla bulunur.

$$i(t) = 1 - \sum_{k=1}^{K} p^2(k \mid t) \tag{6}$$

Burada $p(k|t)$ düğüm *t* içinde sınıf *k*'nın koşullara bağlı özelliklerini, *K* sınıf sayısını, *k* sınıf indeksini ve *t* düğüm indeksini göstermektedir.

2.3. Optimal ağaç

Herhangi bir sınırlandırma yapmadan, sınıflandırma ve SRAT ile elde edilen ilk ağaç modele maximum ağaç ismi verilir. Veri setindeki aykırı veya uzak gözlemlerden dolayı maksimum ağaç genellikle aşırı eğitilmiştir. Bu sorunu ortada kaldırmak için budama yapılması ve optimal ağacın elde edilmesi gerekir. Uygun ağaç boyutunu bulmak için yaklaşımlardan biri veri setinin bir kısmının test için diğer kısmının eğitim için ayrılması şeklindedir. Eğitimle elde edilen ağaç model üstünden test setinde hem bütün ağaç hem de alt ağaçlar için hata değerleri hesaplanır ve en küçük hata değerine sahip alt ağaç optimal ağaç olarak belirlenir. Ancak bu yolla genellikle ideal ağacın elde edilmesi pek mümkün olmamaktadır. Dahası bu yaklaşımın kullanılması için geniş bir veri setine ihtiyaç vardır. Bunun yerine araştırmacıların daha fazla tercih ettikleri çapraz geçerlilik (cross-validation) testi uygulanabilir. Çapraz geçerlilik testinde (1) veri eşit orana (genellikle on eşit parçaya) ayrılır ve her defasında bir altgrup (verinin %10'u) test için kullanılmak üzere veriden çıkartılır ve diğer kalan veriler ile model inşa edilir. (2) Bu işlem verinin ayrıldığı parça sayısı kadar (10 defa) gerçekleştirilir ve böylece bütün veri kullanılmış olur. (3) Her defasında inşa edilen modeller ilgili test grupları ile kontrol edilir. Daha sonra bütün alt gruplar birleştirilir, 2. ve 3. adımlar ağacın her boyutu için gerçekleştirilir. Modellerin değerlendirmesi ile en düşük hata değerine sahip ağaç optimal ağaç olarak kabul edilir (Breiman vd., 1984; De'ath ve Fabricius, 2000; Moisen, 2008).

2.4. Modellerin değerlendirilmesi

Model değerlendirmelerinde hata değerlerinin hesaplanmasında genelde sınıflandırma ağacı için sınıflandırma hata oranı kullanılır. Bunun dışında kappa katsayısı, khi kare, odds oranı ve duyarlılık testleri sınıflandırma ağacı modellerin değerlendirilmesi için kullanılabilir (Manel vd., 2001; Liu vd., 2005; Özkan ve Mert 2010). Regresyon ağacı modellerinin hata değerlendirmelerinde ortalama karakök hatası veya artıkların ortalama sapması kullanılmaktadır. Bunun yanında regresyon katsayısı, etkinlik katsayısı, Akaike kriteri ve Bayesian kriteri regresyon ağacı modellerinin değerlendirmeleri için kullanılabilecek diğer yaklaşımlardır (Aertsen vd., 2010).

3. Kuralların yazılması ve coğrafi modelleme

Optimal ağaca karar verildikten sonra, analiz bitmiştir ve bu ağaç model artık kullanılmaya hazırdır. Ağaç modeller ihtiyaç duyuldukça belli alanlarda hedef değişkenin kestirimi yapmak için kullanılabilir. Diğer yandan ağaç modeller hedef değişkenlerin coğrafi modellemeleri içinde kullanılabilir. Ancak coğrafi modelleme için sınıflandırma ve regresyon ağaçlarının oluşumunu sağlayan bağımlı değişkenlerinin her birine ait dijital altlık haritalarının mevcut olması gerekir. Sınıflandırma ve regresyon ağaçlarının coğrafi modellemesi için modeldeki her nihai düğüm değeri kestirim değeri olarak kullanılır ve her hücreye bu kesittim değerleri ilgili bağımsız değişkenlerin kritik değerleri esas alınarak atanır. Böylece hedef değişkenlerin coğrafi modellemesi gerçekleştirilmiş olur (Özkan ve Mert, 2010).

Sınıflandırma ve regresyon ağacın coğrafi modellemesinde her bir nihai düğüm için (B_n) ilgili bağımsız değişken veya değişkenlerin sürekli (formül 8) veya kategorik (formül 9) olma durumuna göre aşağıdaki (eğer-ise kuralları) formüller kullanılır.

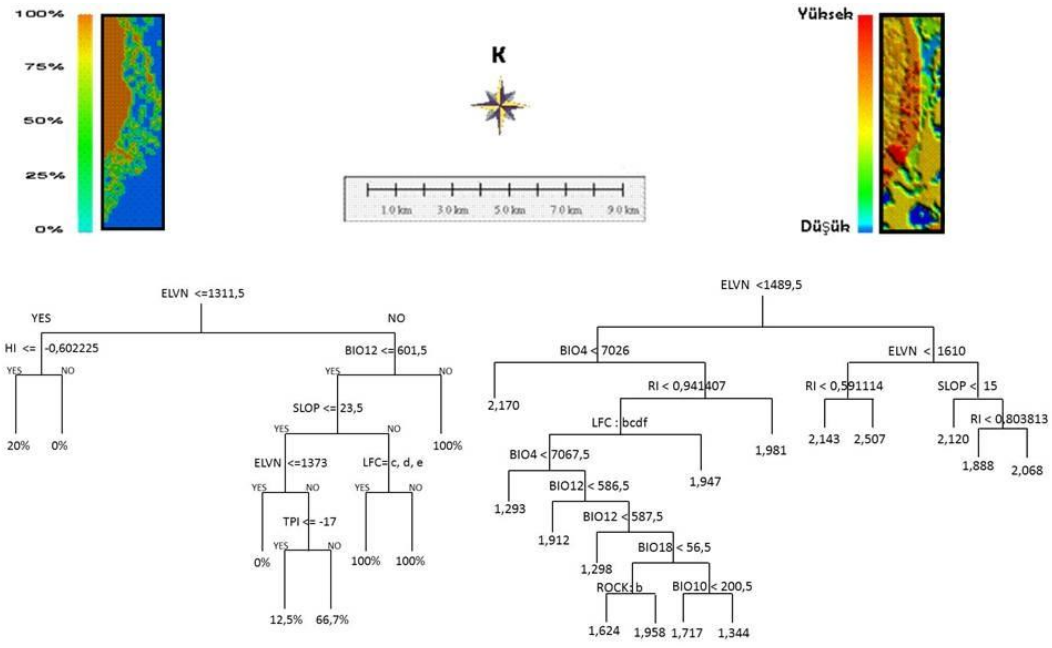
$$=EĞER(VE((X_{L1ij} \# Nd_1 ; X_{L2ij} \ldots X_{Lnij} \# Nd_n); \longrightarrow (B_n) \quad (8)$$

$$=EĞER(VE(VEYA(X_{LC1ij}=c_1;X_{LC1ij}=c_2;\ldots X_{LC1ij}=c_n); \longrightarrow (B_n) \quad (9)$$

Burada $X_{L1ij}\ldots..X_{Lnij}$ ağaç üzerinde belli bir hatta ilk seviyeden *(L1)* (tepe düğümü) son seviyeye (nihai düğüm) (*Ln*) kadar *i.* sütun ve *j.* satır için ayırıcı değişkenleri ve onların kritik değerlerini göstermektedir. *Nd* ayırıcı değişkenlerin ilgili hat boyunca her bir seviyedeki kritik değerlerini göstermektedir. X_{LC1ij} kategorik ayırıcı değişkenleri göstermektedir ve c_n belli bir kategorik değişkenin eğer-ise kurallarından elde edilen kategorileridir.

4. Uygulama örneği

Sınıflandırma ve regresyon ağacı tekniğinin uygulama çıktılarına örnek göstermek amacıyla Şekil 1'de Yukarıgökdere yöresinde dikdörtgen şeklinde kesinlen belli bir alanda henüz daha yazım aşamasında olan iki çalışmanın (Toros sedirinin (*Cedrus libani* A. Rich) ve tür çeşitliliğinin (*H*)) ağaç ve coğrafi dağılım modelleri verilmiştir. Şekil 1 de ki ağaç modellere bakılacak olursa sınıflandırma ağacı yönteminin uygulandığı Toros sedirinin dağılımında başta yükselti (ELVN) olmak üzere sıcaklık indisi (HI), yıllık toplam yağış (BIO12), eğim (SLOP), arazi şekli (LFC) ve yamaç konumu indisi (TPI) rol oynamaktadır. Regresyon ağacının uygulandığı tür çeşitliliği dağılımında da yükselti (ELVN) en önemli yetişme ortamı faktörüdür. Bunun dışında, mevsimsel sıcaklık sapması (BIO4), radyasyon indeksi (RI), eğim (SLOP), arazi şekli (LFC), yıllık toplam yağış (BIO12), en sıcak üç aylık dönemin yağışı (BIO18), en sıcak üç aylık dönemin ortalama sıcaklığı (BIO10) ve anakaya (ROCK) tür çeşitliliği ağacı modelinin oluşmasında katkıda bulunan diğer değişkenlerdir.



Şekil 1: Yukarıgökdere (Isparta) Yöresi'nde bir kesit alanda Toros sedirinin dağılımının sınıflandırma ağacı yöntemi (sağda) ve tür çeşitliliği (H) dağılımının regresyon ağacı yöntemi (solda) ile elde edilen ağaç ve coğrafi dağılım modelleri

5. Öneriler

Ekosistemlerin planlaması ile faydalanma ve sürdürülebilirlik arasındaki denge için en temel ve en fazla istenen ekolojik bilgi hedef türlerin veya tür gruplarının potansiyel değer modelleri ve/veya haritalarıdır. Ekolojik anlamda bir potansiyel değer modeli ve/veya haritasının üretilmesi, bilimsel olarak gerçek veriler üstünden ilişkilendirme ile gerçekleştirilebilmektedir. Bu sayede potansiyel değer tabanlı modeller elde edilebilir.

Ekolojik ilişkilerin karmaşıklığı doğrusal modellerin zaaflarını ortaya çıkarmaktadır. Özellikle ormanları dağlık yerlerde bulunan ülkemizde, geleneksel analitik yaklaşımlar ile ekolojik ilişkilerin yeterince açıklanamadığı bir gerçektir.

SRAT geleneksel yöntemlere alternatif olabilecek hiyerarşik, kural tabanlı, nonparametrik bir yöntemdir. SRAT ile hem kategorik hem de sürekli değişkenler modellenebilmektedir. SRAT ile doğrusal olmayan ilişkilerin algılaması ve modellenmesi mümkün olmaktadır. SRAT kullanılabilecek nonparametrik yöntemlerden sadece biridir. Bunun dışında genelleştirilmiş doğrusal modelin bir uzantısı olan genelleştirilmiş eklemeli model, yapay sinir ağları veya bulanık mantık uygulamalarına da başvurulabilir (Özkan, 2010).

Orman ekoloji alanında model tabanlı çalışmaların bir diğer önemi, üretilen modellerin iklim değişim senaryolarının dijital verilerine uyarlanabilmesidir. Model tabanlı haritaların potansiyel değer göstermesi hedef bölgelerde özellikle ağaçlandırma, restorasyon ve koruma çalışmalarında günümüz ve geleceğe yönelik doğru kararların verilebilmesi için büyük önem taşımaktadır.

Kaynaklar

Aertsen, W., Kint, V., Orshoven, J., Özkan, K., Muys, B., 2010. Comparison and ranking of different modelling techniques for prediction of site index in Mediterranean mountain forests, Ecological modelling 221, 1119-1130.

Breiman, L., Friedman, J.H., Olshen, R., Stone, A.C.G., 1984. Classification and regression trees. Wadsworth International Group, Belmont, California, USA.

Chu, C.M, Tsai, B.W., Chang, K.T., 2009. Integrating decision tree and spatial cluster analysis for landslide susceptibility zonation. World Academy of Science, Engineering and Technology 59:470-483.

De'ath, G., Fabricius K.E., 2000. Classification and regression trees: A powerful yet simple technique for ecological data analysis. Ecology, 81(11): 3178-3192 .

Liu, C., Berry, P.M., Dawson, T.P., Pearson, R.G., 2005. Selecting thresholds of occurance in the prediction of species distribution, Ecography 28, 385-393.

Manel, S., Williams, H.C., Ormerod, S.J., 2001. Evaluating presence-absence models in ecology: the need to account for prevalence, Journal of Applied Ecology, 38, 921-931.

Mckenney, D.W., Pedlar, J.H., 2003. Spatial models of site index based on climate and soil properties for two boreal tree species in Ontario, Canada. Forest Ecology and Management, 175:497-507.

Moisen, G.G., 2008. Classification and Regression Tree. In: Jorgensen SE (ed) In Encyclopedia of Ecology, pp.582-588.

Navarrate E., Espinosa M., 2011. Using the non-parametric classifies CART to model wood density. Journal of Data Science 9:261-270.

Özkan, K., 2010. Orman ekosistem çeşitliliği haritalama çalışmaları için ekolojik alan çeşitliliğinin belirlenmesi üzerine bir öneri. SDÜ Orman Fakültesi Dergisi 2:136-148

Özkan, K., Mert, A., 2010. Isparta Yukarı Gökdere Yöresinde Kasnak Meşesinin Senaryolarına göre 2050 ve 2080 yıllarında muhtemel potansiyel yayılış alanlarının coğrafi modellemesi, Çölleşme ile Mücadele Sempozyumu 17-18 Haziran, 2010, Anitta Otel, Çorum.