VERİ BİLİMİ DERGİSİ
www.dergipark.gov.tr/veri

Dengesiz Veri Setli Sağkalım Verilerinde Cox Regresyon ve Rastgele Orman Yöntemlerin Karşılaştırılması

Pelin AKIN [1]*, Yüksel TERZİ [1],

[1]*Ondokuz Mayıs Üniversitesi, İstatistik Bölümü, Samsun, Türkiye*

Özet

Cox regresyon modeli, temel olarak, hastaların sağkalım süresi ile bir veya daha fazla faktörlerin yaşam süreleri üzerindeki etkilerini incelemek amacıyla yaygın olarak kullanılan bir regresyon modelidir. Yüksek sayıdaki verilerin oluşu, verilerde doğrusal olmayan durum, yüksek derecede etkileşim ve yüksek boyutlu ilişkileri açıklamada kullanılabilecek Cox Regresyon yöntemine alternatif olarak makine öğrenme yöntemleri kullanılmaya başlanılmıştır. Bu çalışmada, veri seti Ondokuz Mayıs Üniversitesi göğüs hastalıkları servisinde yatmakta olan akut lösemi hastalarından elde edilmiştir. Analizden önce, çıktı değişkenin kategorisindeki dengesizliği düzeltmek için sentetik azınlık aşırı örnekleme (Smote) yöntemi uygulandı. Daha sonra, her hastanın riskini belirmek için rastgele orman ve Cox Regresyon yöntemleri kullanılmıştır. Bu iki yöntem uyum indeks, roc eğrisinin altında elde edilen alan (AUC) ve hata oranına göre karşılaştırılmıştır. Sonuç olarak, rastgele orman sağkalım analizinde Cox regresyonuna alternatif bir yöntem olarak kullanabilir.

Makale Bilgisi

Başvuru:
04/11/2019
Kabul:
04/05/2020

Anahtar Kelimeler: *Rastgele orman, Smote tekrar örnekleme, Cox regresyon,Dengesiz veri seti*

Comparison of Cox Regression and Random Forest Methods Survival Data with Imbalanced Data Set

Abstract

The Cox proportional-hazards model is essentially a regression model commonly used statistical in medical research for investigating the association between the survival time of patients and one or more factors . Alternative machine learning methods were introduced to the Cox Regression method, which can be used to explain the high number of data, nonlinear status, higher-order interactions and high dimensional covariates. In this study, patients who have been in Chest diseases service in the Hospital of Ondokuz Mayıs University. Before analysis, the smote sampling method was applied because the categories of the output variable were unbalanced. In this study, Random Forest and Cox Regression were used to determine the risk of each patient in leukemia. These two methods are compared to the C-index, area under the ROC curve (AUC) and error rate. According to the result, it was found that random forest is used as an alternative to Cox regression in survival analysis.

Keywords: *Random forest, Smote resampling method, Cox regression, Imbalanced dataset*

* İletişim e-posta: pelin.akin@omu.edu.tr

** Bu çalışmanın bir kısmı II. International Conference on Data Science and Applications 2019'da sözlü olarak sunulmuştur.

1 Giriş

Tıbbi araştırmalarda klasik istatistik yöntemlerini kullanmak bazen yetersiz kalmaktadır. Bunun temel nedeni sansürlü veri içermesidir. Sansürlü veri, bir bireyin sağkalım süresi hakkında her zaman tam bir bilgiye ulaşılamayabilir. Sansürlü veri olmasının genellikle üç nedeni vardır. İlki, birey gözlem esnasında ölebilir. İkincisi, birey gözlemden geri çekilebilir. İlgilenilen olay dışında bir başka nedenden dolayı ölebilir veya uygulanan yöntemlerden beklenmeyen bir sonuç alınabilir. Üçücüncü olarakda, birey gözlemin sonunda hala yasıyor olabilir. Sansürleme, sağdan sansürleme ve soldan sansürleme ve aralıklı sansürleme olarak üç gruba ayrılır [1]. Yaşam süresini belirlerken açıklayıcı değişkenlerin etkisini incelemek için genellikle Cox regresyon analizi kullanılmaktadır. Cox regresyon, 1972 yılında Cox tarafından geliştirilmiştir [1].

Veri sayısının fazla ve daha karmaşık olduğu sağkalım verilerinde kullanılabilecek yeni modeller araştırılmaya başlanmıştır. Bu çalışmada makine öğrenme yöntemlerinden olan rastgele orman yöntemi sansürlü veri setine kullanılmıştır. Bu yöntem için R programında ranger paketi kullanılmıştır.

Makine öğrenmesi; yapısal işlev olarak öğrenebilen ve veriler üzerinden tahmin yapılan algoritmaların genel adıdır. Makine öğrenme algoritmalarından rastgele orman algoritması, birden fazla karar ağacının birleşiminden oluşur [2].

Son yirmi yılda rastgele orman teknikleri hastalık prognozu ve tahmini için yaygın bir şekilde uygulanmıştır. Akman ve ekibini yaptığı çalışmada, rastgele orman yöntemini sağlık verisinde uygulamıştır. Ve çok sayıda değişkenin olduğu DNA veri seti gibi binlerce gen arasından önemli olanları tespit etmek için rastgele orman yöntemini kullanılabilir olduğunu iddia etmişlerdir[3] .Exarchos ve arkadaşları ağız kanserine yakalanan hastaların oral skumoz hücreli karsinomun (OSCC) tekrarlama tahmini üzerine bir çalışma önermiştir. Bu çalışmada Bayes network, yapay sinir ağları, destek vektör, karar ağaçları, rastgele orman olmak üzere beş farklı sınıflandırıcı kullanarak performansları karşılaştırılmıştır [4]. Ama sansürlü veri içeren verisetlerinde çok fazla çalışma bulunmamaktadır. Weathers (2017) ekibinin yaptığı çalışmada, rastgele orman, koşullu çıkarım orman (Conditional Inference Forest) ve Cox regresyon modellerini üç farklı veri seti kullanarak karşılaştırmıştır. Karşılaştırmada uyum indeksi ve hata tahminleri kullanılmıştır. Rastgele orman sonucu Cox regresyondan daha iyi bir performans göstermiştir [5].

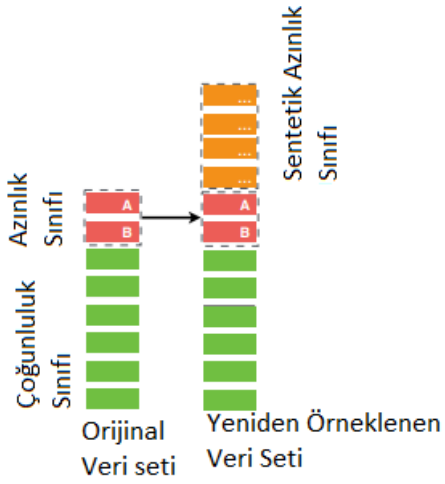
Bu çalışmada veri seti olarak Ondokuz Mayıs Üniversitesi Göğüs Hastalıkları servisinde yatan akut lösemi 165 hastadan elde edilen veriler kullanılmıştır [6]. Literatürdeki çalışmadan farklı olarak veri setinde dengesiz veri seti problemi vardır. Dengesiz veri seti hedef değişken kategorileri eşit olmadığı durumda ortaya çıkar. Bu problemi ortadan kaldırmak için sentetik azınlık aşırı örnekleme (Smote) yöntemi kullanılmıştır. Daha sonra elde edilen veri setine rastgele orman algoritması ve orijinal veri setine Cox regresyon uygulandı ve sonra sonuçlar karşılaştırıldı.

2 Yöntemler

2.1 Sentetik azınlık aşırı örnekleme (Smote) Yöntemi

Dengesiz veri seti problemi hedef değişkenin etiketinde eşit bir dağılım bulunmamasıdır. Bu durum, bir sınıf tahmininde zorluklar çıkarmaktadır. Bu problemi çözmek için farklı yöntemler kullanılmaktadır. Smote en fazla tercih edilen yöntemdir.

Smote'da her azınlık sınıfı örneği alınır ve bu örneğe ait k komşusunun herhangi birine ya da tümüne bakılarak sentetik örnekler oluşturulur (Şekil 1) [7].



Şekil 1. Sentetik azınlık aşırı örnekleme (Smote) Yöntemi

2.2 Cox Regresyon

Cox regresyon, sağkalım analizinde en çok tercih edilen yöntemlerden biridir. Cox regresyon yönteminde, yaşam süresi bağımlı değişken ile yaşam süresini etkileyen bağımsız değişkenlerin neden sonuç ilişkisi belirlenmeye çalışır. Anlık ölüm (hazard) oranı aşağıdaki gibi hesaplanır.

$$\mathrm{h(t,x)} = h_0(t)\exp\left(\sum \beta_i X_i\right) \quad (1)$$

Burada X_i bağımsız değişkenleri, β_i regresyon katsayıları vektörü ve $h_0(t)$ temel hazard fonksiyonu olarak tanımlanır [1].

Bağımsız değişkenler vektörü x ve sağkalım süresi t olsun. Bir bireyin bağımsız değişkenlere göre hazard fonksiyonu h(t;x) ile gösterilir. Cox tarafından parametre tahmin edilmesinde kullanılmak üzere önerilen kısmi olabilirlik fonksiyonu aşağıda gösterilmiştir [1].

$$L_p(x/\beta) = \prod_{i=1}^{k} \frac{\exp\left(\sum_{i=1}^{p} \beta_i X_i\right)}{\sum_{j\in R(t_i)} \exp\left(\sum_{j=1}^{p} \beta_j X_j\right)}$$

2.3 Rastgele orman

Rastgele orman algoritması topluluk öğrenme yöntemlerinden biridir. Topluluk öğrenme yöntemleri, farklı modelleri birleştirerek sonuçları iyileştirmeyi amaçlar. Rastgele orman algoritması birden fazla karar ağaçlarının topluluğundan elde edilmektedir. Rastgele orman algoritması avantajlarından biri, hem sürekli hem kesikli değişkenlerin birlikte kullanabilmektedir. Ayrıca büyük küçük boyutlu veri setlerinde kullanılabilmektedir. Rastgele orman ise doğrululuk oranı diğer algoritmaya göre yüksek çıkmaktadır. Dezavantaj olarak ise algoritma siyah kutu yani ağaç yapısı görülmemektedir [8].

Sağkalım analizinde rastgele orman algoritmasında kullanabilmek için R programında ranger paketi

kullanılmaktadır. Diğer rastgele orman algoritmasından farklı olarak bağımlı değişken olarak sağkalım fonksiyonu kullanılmaktadır.

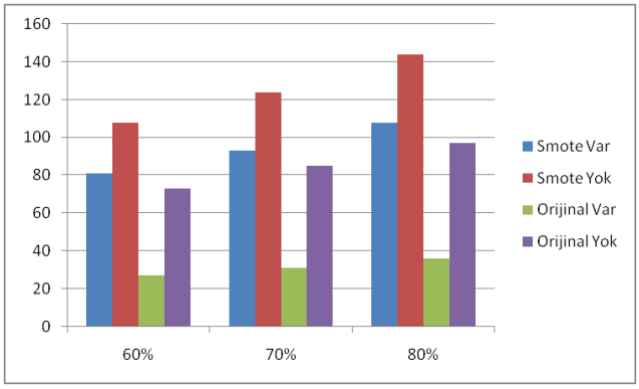
3 Bulgular

3.1 Uygulama Verisi

Ondokuz Mayıs Üniversitesi göğüs hastalıkları servisinde 1992-2002 yıllar arasında akut lösemi 165 hastadan elde edilen veriler kullanılmıştır [6]. Veri setinde on dört tane bağımsız değişken, bir tane hedef değişken ve zaman değişkenler vardır. Bağımsız değişkenler arasında sadece yaş ve gün nicel diğer değişkenler niteldir. Hedef değişken 121 hastanın sansürlü, 44 hastanın ölüm vardır. Hedef değişkenin kategorileri arasındaki dengesizliği gidermek için Smote yöntemi uygulanmıştır.

3.2 Sentetik azınlık aşırı örnekleme (Smote) yöntemi sonucunda veri seti

Makine öğrenmesinde veri seti eğitim ve test olmak üzere ikiye bölünür. Bu çalışmada, %60, %70, %80 olmak üzere üç farklı ayırma oranı kullanılmıştır. Eğitim verisini ayırdıktan sonra sentetik azınlık aşırı örnekleme yöntemi uygulanmıştır. Şekil 2'de orijinal ve Smote yöntemi kullanıldıktan sonraki dağılım verilmiştir. Örneğin, %60-%40 olarak ayırdığımızda hedef değişkeni 73 sansürlü, 26 ölümdür. Smote yöntemi kullandıktan sonra 108 sansürlü 81 tane ölüm olarak sentetik veri üretilir. Bu şekilde , dengesiz veri problemi ortadan kalkmaktadır.



Şekil 2. Hedef değişkenin Smote yönteminden sonraki dağılımı

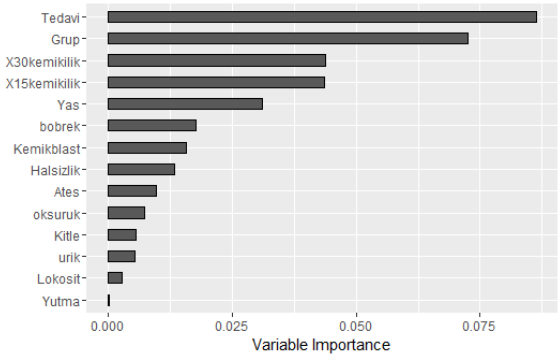
3.3 Rastgele orman analiz sonuçları

Tablo 1'de lösemi hastalarının ölüm riskini belirlemek için rastgele orman sonuçları verilmiştir. Uyum indeksi ve roc eğrisinin altında kalan (ROC) değeri en yüksek ve hata oranı küçük olanı tercih edilir. Yani, smote yöntemi ile elde edilen veri ve %80 eğitim verisi olarak ayırdığımız veri seti en iyi performansı vermiştir.

Tablo 1. Orijinal veri seti ve Smote ile elde edilen veri setinin rastgele orman sonuçları

	SMOTE			**Orijinal Veri Seti**		
	Uyum indeks	AUC	Hata oranı	Uyum indeks	AUC	Hata oranı
%80-%20 (Eğitim- Test)	0,831	0,756	0,168	0,748	0,656	0,252
%70-%30 (Eğitim- Test)	0,811	0,803	0,189	0,745	0,672	0,255
%60-%40 (Eğitim- Test)	0,799	0,797	0,201	0,642	0,590	0,358

Şekil 3'de rastgele orman analiz sonucundaki önemli değişkenlerin sıralaması görülmektedir. Önemli ilk üç değişken tedavi grupları, grup ve 30. Gün kemik iliği remisyonu olarak belirlenmiştir.



Şekil 3. Veri setinin önemli değişkenleri

3.4 Cox Regresyon analiz sonuçları

Tablo 2'de Cox regresyon analizinin sonuçları özetlenmiştir. Yutma güçlülüğü, tedavi planı, 30. gün kemik remisyon değişkenleri istatistiksel olarak anlamlıdır. Yutma güçlüğü hazard oranı 26,7565 olup, yutma güçlülüğü olanların ölüm riski olmayanlara göre yaklaşık 26 kat daha fazladır. Tedavi planının hazard oranı ise 0,3076 olduğundan, Tedavi 2 (Aml) tedavi planı olanların ölüm riski diğer tedavi planı olana göre yaklaşık %70 daha azdır. 30. gün kemik iliği remisyon değişkenin hazard oranı 0,214 olduğundan, remisyon olanların ölüm riski olmayanlara göre yaklaşık %79 daha azdır.

Tablo 2. Cox regresyon analiz sonuçları

	Tahmin	Hazard oran	Standart hata	p-değeri
Grup	-1,233	0,291	1,169	0,292
Yaş	0,023	1,023	0,045	0,610
Yutma	**3,287**	**26,756**	**1,420**	**0,021**
Halsizlik	-0,001	0,992	0,374	0,982
Öksürük	0,026	1,026	0,476	0,957
Ateş	-0,067	0,935	0,361	0,852
Kitle	0,841	2,318	0,509	0,099
Lökosit	0,464	1,591	0,471	0,325

Ürik	-0,437	0,646	1,145	0,703
Böbrek1	-0,481	0,618	0,456	0,291
Böbrek2	-0,777	0,460	0,859	0,366
Kemikblast1	-0,946	0,388	1,391	0,496
Kemikblast2	-0,451	0,637	1,380	0,744
Tedavi1	-0,314	0,730	1,210	0,795
Tedavi2	**-1,179**	**0,308**	**0,565**	**0,037**
X15 Kemikilik1	0,128	1,137	0,442	0,771
X15 kemikilik2	0,011	1,011	0,461	0,981
X30 kemikilik	**-1,539**	**0,214**	**0,442**	**0,000**

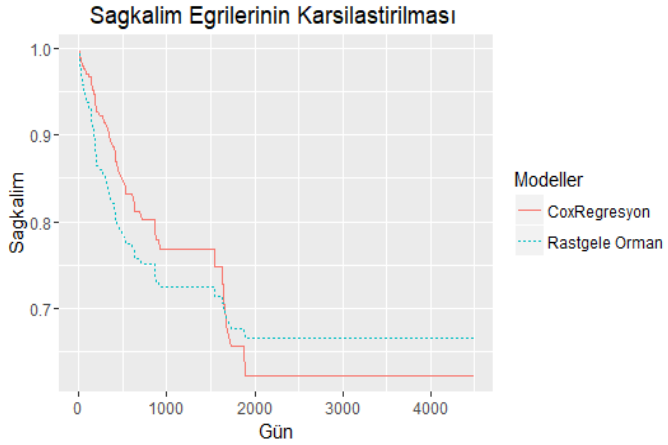
4 Sonuçlar

Bu çalışmada, lösemi hastalarının ölüm riskini belirmek için Cox regresyon ve rastgele orman yöntemleri kullanılmıştır. Makine öğrenme algoritmalarında en çok karşılan problemlerden biri dengesiz veri problemidir. Bu problemi ortadan kaldırmak için Smote yöntemi seçildi. Smote ile elde edilen veri setine rastgele orman yöntemi uygulandı. Bu modeli ile Cox regresyon karşılaştırmak için uyum indeks ve hata oranı kullanıldı. Tablo 3 'de görüldüğü gibi rastgele orman ve Cox regresyon sonuçları yakındır. Modelleri detaylı bir şekilde yani zamana bağlı olarak grafiklerini inceleyelim.

Tablo3. Modellerin uyum indeksi ve hata oranı değerleri

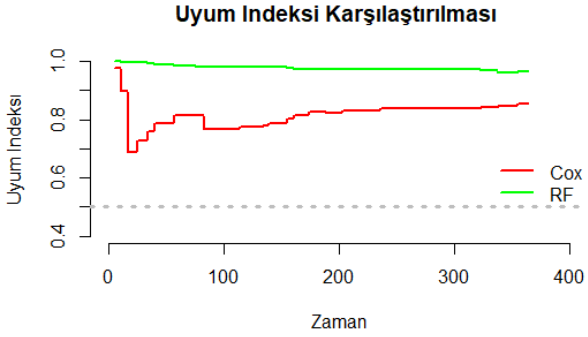
	Uyum indeks	**Hata oranı**
SMOTE %80-%20 (Eğitim- Test)	0,831	0,168
Cox regresyon	0,812	0,046

Sağkalım eğrilerin karşılaştırılması Şekil 4'de verilmiştir. Cox regresyon daha iyi gözükmesine rağmen zaman arttıkça rastgele orman analizi daha iyi tahmin sonucu vermiştir.



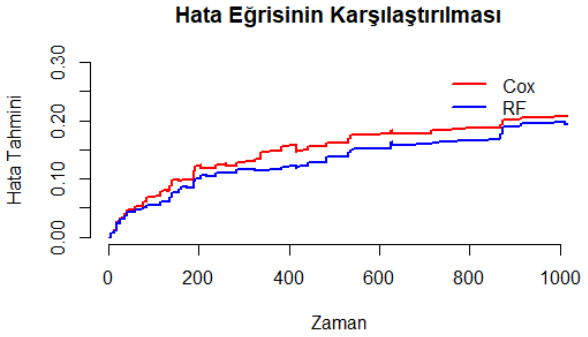
Şekil 4. Cox regresyon ve rastgele orman yöntemlerin Sağkalım eğrilerinin karşılaştırılması

Şekil 5'de ise zamana bağlı olarak modellerin uyum indekslerinin karşılaştırma grafiği verilmiştir. Rastgele orman Cox regresyona göre daha iyi bir performans göstermiştir.



Şekil 5. Cox regresyon ve rastgele orman yöntemlerin uyum indeksin karşılaştırılması

Şekil 6'de ise hata eğrileri karşılaştırılmıştır. Grafikte bu iki model için trendin aynı olduğu ancak Rastgele ormanın biraz daha yüksek olduğu görülmektedir.



Şekil 6. Cox regresyon ve rastgele orman yöntemlerin hata eğrilerinin karşılaştırılması

Sonuç olarak, veri sayısının fazla ve daha karmaşık olduğu sağkalım veri setlerinde Cox regresyonun yetersiz kaldığı durumlarda rastgele orman yöntemi alternatif olarak kullanabilir. Buna ek olarak sağkalım verilerin makine öğrenme yöntemlerin tercih edilmemesi sebepleri arasında olan dengesiz veri problemini çözmek için smote yöntemi ham veriye göre iyi bir performans göstermiştir.

Kaynaklar

1. Kleinbaum, D.G., *Survival Analysis, a Self-Learning Text.* Biometrical Journal: Journal of Mathematical Methods in Biosciences, 1998. **40**(1): p. 107-108.
2. Biau, G., *Analysis of a Random Forests Model.* Journal of Machine Learning Research, 2012. **13**(Apr): p. 1063-1095.
3. Akman, M., Y. Genç, and H. Ankarali, *Random forests yöntemi ve sağlık alanında bir uygulama.* Turkiye Klinikleri Journal of Biostatistics, 2011. **3**(1): p. 36-48.
4. Exarchos, K.P., Y. Goletsis, and D.I. Fotiadis, *Multiparametric decision support system for the prediction of oral cancer reoccurrence.* IEEE Trans Inf Technol Biomed, 2012. **16**(6): p. 1127-34.
5. Weathers, B., *Comparision of Survival Curves Between Cox Proportional Hazards, Random Forests, and Conditional Inference Forests in Survival Analysis.* 2017.
6. Dirican, A., *Kliniğimizde akciğer kanseri tanısı alan hastaların prospektif olarak değerlendirilmesi ve sağkalıma etki eden faktörlerin belirlenmesi* 2004, Ondokuz Mayıs University.
7. Chawla, N.V., et al., *SMOTE: Synthetic minority over-sampling technique.* Journal of Artificial Intelligence Research, 2002. **16**: p. 321-357.
8. Breiman, L., *Random forests.* Machine Learning, 2001. **45**(1): p. 5-32.