Olaya İlişkin Potansiyellerde tek-Epok temelli sınıflandırma

Classification of single epochs in Event Related Potentials

Kerime Dilşad ÇİÇEK[1], Oğuz BAYAT[2], Osman Nuri UÇAN[2]
Meslek Yüksekokulu Bilgisayar Programcılığı[1],
Fen Bilimleri Enstitüsü[2]
Altınbaş Üniversitesi
İstanbul, Türkiye
kerime.cicek@altinbas.edu.tr

Adil Deniz DURU[3]
Spor Bilimleri Fakültesi, Antrenörlük Eğitimi Bölümü[3]
Marmara Üniversitesi
İstanbul, Türkiye
deniz.duru@marmara.edu.tr

***Özetçe*—. Bu çalışma kapsamında olaya ilişkin potansiyellerde epok temelli sınıflandırma gerçekleştirilmiştir. Karar Ağaçları, Lojistik Regresyon, Rastgele Orman, Destek Vektör Makineleri ve XGBoost sınıflandırma yöntemlerini kullanarak EEG sinyallerinin sınıflandırma performansları değerlendirmiştir. Bu kapsamda, deneklere iki farklı uyaran rastgele sunularak EEG kayıtları elde edilmiştir. Elde edilen özellik seti, karar ağaçları, lojistik regresyon, rastgele orman, destek vektör makineleri ve xgboost sınıflandırıcılarına giriş olarak verilmiştir. Elde edilen test verisinin kısıtlı olmasından dolayı dataya Sentetik Azınlık Aşırı Örnekleme Tekniği(SMOTE) uygulanmış, yeni oluşturulan veri kümesi ile sınıflandırma yapılmıştır. Gerçekleştirilen çalışma sonucunda, en iyi performans eğitim verisinde %91 doğruluk oranı ile rastgele orman ve xgboost sınıflandırma metotlarında, test kümesinde ise xgboost_Tuned %62 oranında doğruluk ve %71 oranında F1 değeri elde edilmiştir. xgboost sınıflandırma metodu kullanılarak diğer sınıflandırıcılardan daha üstün sonuçlar bulunmuştur.**

Anahtar Kelimeler — EEG, SMOTE, XGBoost

***Abstract*—In the concept of this thesis, single trial event related potential measurements were classified. Classification performances of Decision Trees, logistic regression, random forest, Support Vector Machines and XGBoost methods are evaluated. In this context, EEG was collected during the presentation of two different stimulus. The resulting feature set is given as an input to decision trees, logistic regression, random forest, support vector machines, and xgboost classifiers. Due to the limited test data obtained, synthetic Minority oversampling technique(SMOTE) was applied to the data and classification was performed with the updated dataset. As a result of the study, 91% accuracy was obtained for the training dataset in random forest and XGBoost classification methods. For the test set xgboost_tuned has a 62% accuracy and 71% F1 value. To conclude, superior results were found from other classifiers using the xgboost classification method.**

Keywords —EEG, SMOTE, XGBoost

978-1-7281-2420-9/19/$31.00 ©2019 IEEE

I. GİRİŞ

Beyin elektriksel aktivitesi temelde, nöronlardaki dentrintlerin sinaptik uyarımları sırasında elektrik akımlarının akışından kaynaklanan veri şeklinde gözlenmektedir [1]. Elektroensofalogram, (EEG) kayıt sistemi elektrotlar, amplifikatörler, A/D dönüştürücü ve bir kayıt cihazından oluşur. Elektrotlar kafa derisinden sinyali alır, amplifikatörler analog sinyali EEG sinyallerinin genliğini büyütmek için işler. Böylece A/D dönüştürücü sinyali daha doğru bir şekilde dijitalleştirebilir ve bilgisayarda veriler görüntülenir [3]. Günümüzde EEG sinyallerinin sınıflandırılmasına yönelik öznitelik çıkarımı çalışmaları ivme kazanmıştır [2]. Sınıflandırma teknikleri olarak, denetimli ve denetimsiz makina öğrenmesi yöntemlerine başvurulmaktadır [4].
Denetimli (Supervised) Öğrenmede, sistemin öğrenme işlemini gerçekleştirebilmesi için danışman gerekmektedir. Öğrenilmesi istenilen model ile ilgili Girdi/Çıktı değerleri sisteme verilmelidir. Giriş ve çıkış değerleri eşleştirilerek sistemin öğrenmesi sağlanmaktadır [5]. Denetimli öğrenme problemleri, sınıflandırma problem olarak ele alınmakta ve eğitilmiş sistem tarafından oluşturulan model kullanılarak test seti üzerinden tahmin yapmak üzere kullanılmaktadır [6].
Denetimsiz(Unsupervised) Öğrenmede ise sisteme çıkış değeri verilmeden, sadece giriş değerlerini kullanarak, makinenin sisteme girilen örnekler arasında ilişki kurup model oluşturmaya çalışmasıdır [6].

Bu çalışmada, EEG verilerinin sınıflandırılması amacıyla denetimli öğrenme modellerinde kullanılan karar ağaçları, rastgele orman, lojistik regresyon, karar destek sistemleri ve xgboost sınıflandırıcıları kullanılarak performansları karşılaştırılmıştır.

II. MATERYAL VE METOD

A. Veri Seti

Bu çalışma kapsamında, sekiz katılımcı deneylere katılmıştır. Mutlak olarak, 100 mikrovolttan yüksek olan sinyallerin 200 milisaniye sonrası olan bölümler otomatik olarak, göz hareketleri kaynaklı artefaktlar ise el ile işaretlenmiş ve ileri analizlerden çıkarılmıştır. Elde edilen EEG veri seti, Brain Products actiCAP ile 1000 Hz örneklem frekansıyla toplanmıştır. EEG kaydı FP1, FP2, F3, Fz, F4, P3, P4, Pz, C3, C4, Cz, O1, O2 Oz, T7 ve T8 elektrotlarından olucak şekilde, 16 kanaldan ölçümler gerçekleştirilmiştir. Olaya ilişkin potansiyel ölçümleri (OİP), iki farklı tip uyaran kullanılarak gerçekleştirilmiştir. Deneyde kullanılan uyaranlar, MATLAB ortamında hazırlanan sunum programı ile katılımcılara sunulmuştur. Deneklerden elde edilen verilerin analizleri için Brain Products Analyzer paket programı kullanılıp, verilerin sınıflandırılmasında ise R programı kullanılmıştır. Bu çalışma, Marmara Üniversitesi Sağlık Bilimleri Enstitüsü Etik Kurulu tarafından 23.02.20 15-14 onay tarihi ve onay sayısı ile onanmıştır.

B. Problemin Tanımlanması

Deneklere go/nogo testi yapılmıştır. Bilgisayar ekranında deneklere X ve O harfleri uyaran olarak gelmiştir. 0,925 msn yanıp sönen harfler ve 0,75ms farenin sol tuşuna basma süresidir. Deneklerden harfi gördüklerinde, farenin sol tuşuna basması istenmiştir. Toplam 8 kişiden 501 no/go, 173 go deneme alınarak 1 sn'lik süre zarfında zaman serisi ölçümü yapılmıştır. Temin edilen veriye dayanarak 1 veya 2 olarak sınıflandırılan bir model kurulması istenmiştir.

C. Karar Ağaçları(Decision Tree)

Karar ağacı algoritmaları, özellik değerlerine bakarak verileri kategorize eden ağaçlardan oluşmaktadır. Belirli kümelerdeki örneklerin ait olduğu sınıflar alınıp, karar ağacı yaprağına eşit bir şekilde atanır. Yüksek karar düğümü kök düğümü olarak adlandırılır. Verileri en etkili şekilde sıralayabilen özellik kök düğüm olarak seçilir. Bu yol ile, eğitimin her alt bölümü için veriler çoğaltılır ve tüm veriler belirli sınıf partilerine bölünene kadar sınıflandırılır [7].

D. Destek Vektör Makinesi(DVM)

Destek vektör makineleri (DVM) regresyon ve sınıflandırma problemlerinde sıklıkla kullanılan makine öğrenme tekniklerinden biridir [8, 9]. DVM'leri iki aşamada çalışır, etiketli verilerin bir DVM modeli oluşturmak için kullanıldığı bir eğitim aşaması ve eğitimli modelin yeni bir veri kümesinin ait olduğu sınıfı belirlemek için kullandığı bir sınıflandırma aşamasıdır [9, 10].

E. Rastgele Orman(Random Forest)

Rastgele orman algoritmasında karar ağacı algoritmasının tersine pek çok karar ağacı bir arada bulunur. Karar ağaçlarında tek bir ağaç oluşturulurken burada pek çok tekli ağaç oluşmaktadır. Bu denli çok ağaç bulundurması ve rastgele değişkenler seçerek oluşturduğu tekli ağaçlar yüzünden rastgele orman adını almıştır. Oluşturulan her ağaç sonuç olarak bir çıktı verir. Bu çıktılar arasında yapılan oylama sonucunda ağaçlar arasında en çok oyu alan çıktı, tahmin edilen çıktı olarak belirlenmiş olur [11].

F. Lojistic Regresyon

Lojistik fonksiyona dayanan popüler bir sınıflandırma modelidir [12]. Lojistik fonksiyon, herhangi bir giriş x değerini alıp, 0 ve 1 arasında bir değere eşleyebilen s şeklinde bir eğridir. Bağımlı veya bağımsız değişkenler arasında doğrusal bir ilişki olduğunu varsayarak sınıflandırma yapılır. Bu nedenle, bir lojistik regresyon, gerçek problemi aşırı basitleştirecek ve yüksek bir önyargı ile sonuçlanan çok esnek olmayan bir sınıflandırma algoritması olarak kabul edilir [12].

G. Aşırı Gradyan Arttırma (XGBoost)

Aşırı Gradyan Arttırma (Extreme Gradient Boosting), Friedman tarafından önerilen boosting metodunun bir çeşididir [13]. Gradyan ağaç arttırma, güçlü bir sınıflandırıcı oluşturmak için zayıf sınıflandırıcılar kümesini birleştiren bir ağaç topluluğu artırma yöntemidir [14]. Hem Gradyan artırma hem de XGBoost aynı parametreyi takip eder. XGBoost sınıflandırma metodu ile ağaçların karmaşıklığını kontrol edilerek daha iyi performanslar elde edilir [14].

III. DENEYSEL SONUÇLAR

Çok kanallı 1 saniye boyutundaki pencereler bir epok olarak belirlenmiş ve etiketlenmiştir. Gürültü işaretleme ve belirleme işlemi sonrasında 674 adet epoka ulaşılmıştır 674 adet EEG trasesisinden ile elde edilen zaman-frekans bileşenlerinden sınıflandırıcının performansını ölçmek için, hata matrisi kullanılmıştır. Hata matrisi; model başarılarını değerlendirirken kullanılan temel kavramlardan olan doğruluk, duyarlılık, seçicilik, kesinlik ve F1 ölçütü gibi değerler hata matrisi yardımı ile elde edilir. Modelin başarısı, doğru ve yanlış sınıflara atanan örnek sayıları ile ilgilidir. Test sonucunda elde edilen sonuçların başarı oranı hata matrisi ile ifade edilir [15].

TABLO I. Hata Matrisi

		Öngörülen Sınıf	
		Hayır	Evet
Doğru Sınıf	Hayır	TN (Doğru Negatif)	FP (Yanlış Pozitif)
	Evet	FN (Yanlış Negatif)	TP (Doğru Pozitif)

Tablo I.'de görüldüğü gibi, FP (Yanlış Pozitif) ve FN (Yanlış Negatif); İkili bir sınıflandırma yaparken modelin yapabileceği iki tür hata vardır. Yanlış Pozitifler (FP) yani sınıflandırıcının hayır cevabı vermesi gerekirken evet cevabı vermesidir ve FN (Yanlış Negatifler) cevabın evet olması gerektiği durumda hayır cevabı almamızdır [16, 18].

TP (Doğru Pozitif) ve TN (Doğru Negatif); İkisi de gerçek pozitif ve negatif cevaplardır. TP(Doğru Pozitif) modelin tahmini ve gerçekte doğru olan durumdur, TN (Doğru Negatif) ise, modelin tahmini gerçekte negative olan doğru durumdur [16].

Doğruluk; sistemin doğru tahminde bulunduğu, Doğru Pozitif ve Doğru Negatif sayılarının veri kümesindeki değerlere bölünmesi ile elde edilir [8].

$$\text{Doğruluk (Accuracy)} = \frac{(TP+TN)}{(TP+TN+FN+FP)} \quad (1)$$

Duyarlılık; sistemin doğru tahmin ettiği pozitif örnek sayısının, doğru tahmin edilen pozitif ve yanlış pozitif sayısına oranıdır. Duyarlılık olarak da adlandırılan geri çağırma değeri, tüm gerçek örnekler arasında gerçek sınıflandırılmış örneklerin göreceli miktarı olarak tanımlanır.

$$\text{Duyarlılık(Precision)} = \frac{TP}{(TP+FP)} \quad (2)$$

Kesinlik; Sistemin doğru pozitif örnek sayısının, sistemde doğru ve yanlış tahmin edilen pozitif değerlere bölünmesidir. Elde edilen sonuç ise, doğru durumların ne kadarının doğru tahmin edildiğini bize vermektedir.

$$\text{Kesinlik(Recall/Sensivity)} = \frac{TP}{(TP+FN)} \quad (3)$$

Seçicilik (Specificity); Sınıflandırmanın gerçek negatif değerleri ne kadar doğru tahmin ettiği ile ilgilidir.

$$\text{Seçicilik(Specificity)} = \frac{TN}{TN+FP} \quad (4)$$

F1 Değeri; Kesinlik ve duyarlılık sonuçlarını beraber değerlendirip, bir değer hesaplamaktadır. Daha yüksek F1 değeri elde etmemiz daha başarılı modeller yakalamamızı sağlamaktadır.

$$F_1 = 2.\frac{precision.recall}{precision+recall} \quad (5)$$

Bu çalışmada veri seti içindeki en büyük ve en küçük değerler ele alınarak verilere Min-Max Normalizasyonu uygulanmıştır. Elde edilen veri setinin %80-%20 dağılımı uygulanarak bağımlı değişkenler için TABLO II de ki sonuçlar elde edilmiştir.

TABLO II. %80-%20 Bağımlı Değişken Veri Dağılımı

0=(1)	1=(2)
401	138
%74	%26

Tabloda da görüldüğü üzere, test verisinin kısıtlı olmasından dolayı veriye SMOTE(Sentetik Azınlık Aşırı Örnekleme Tekniği) suni veri temin edilmiştir. SMOTE gerçek veriler üzerinde belirli işlemleri gerçekleştirmek için ekstra eğitim verileri oluşturmak anlamına gelmektedir[17]. Sentetik veri üretiminden sonra bağımlı değişken veri dağılımı TABLO III görülmektedir.

TABLO III. Smote Bağımlı Değişken-Veri Dağılımı

0(=1)	1(=2)
276	276
%50	%50

Veri dağılımı ve normalizasyon işleminden sonra R kullanılarak, hata matrisi yardımıyla Doğruluk, Duyarlılık, Kesinlik, Seçicilik ve F1 metrikleri elde edilmiştir.

TABLO IV. Eğitim Verisi Sonuçları

Eğitim Data Set	*Doğruluk*	*Duyarlılık*	*Kesinlik*	*Seçicilik*	*F1*
Lojistik Regresyon	*0.86*	*0.88*	*0.85*	*0.88*	*0.86*
Karar Ağaçları	*0.82*	*0.81*	*0.81*	*0.84*	*0.82*
Rastgele Orman	*0.91*	*0.95*	*0.96*	*0.86*	*0.91*
Karar Destek Makinesi	*0.76*	*0.74*	*0.71*	*0.82*	*0.75*
Karar Destek Makinesi Tuned	*0.89*	*0.87*	*0.86*	*0.91*	*0.88*
XGBoost	*0.91*	*0.9*	*0.9*	*0.92*	*0.91*
XGBoost Tuned	*0.57*	*0.5*	*0.6*	*0.54*	*0.64*

R yardımı ile elde ettiğimiz eğitim verisi sonuçları TABLO IV de görüldüğü üzere, eğitim verimizin doğruluk oranı karar ağaçları %82, lojistik regresyon %86, rastgele orman %91, karar destek makineleri tuned %89 ve xgboost %91 olarak bulunmuştur. Rastgele orman ve xgboost sınıflandırma tekniklerinin eğitim verisi değerlendirilirken, %91 doğruluk oranını yakaladığı görülmektedir. Sınıflandırma sonuçları değerlendirirken modelin ne kadar başarılı olduğunu söyleyebilmek için yanlızca doğruluk değerlerine bakmak yeterli değerdir. Modelin doğru ve yanlış tahminleri ne kadar doğru tahminleyebildiğine bakmak için, kesinlik ve seçicilik değerleri ele alınır. Kesinlik(Sensitivity), makinenin doğru olan durumları yakalayabilme başarısı, xgboost da %90'ı bulurken, rastgele orman sınıflandırma metodun da %96'yı sağlamıştır. Makinenin yanlış durumları ne kadar doğru tahmin ettiğine bakacak olursak xgboost ile %92 bir başarı oranı sağlanmıştır. Genel olarak bakıldığında eğitim verisi için en iyi sonuçları rastgele orman algoritması ikinci olarak da xgboost sınıflandırma metotunda sağlanmıştır

TABLO V. Test Verisi Sonuçları

Test Data Set	Doğruluk	Duyarlılık	Kesinlik	Seçicilik	F1
Lojistik Regresyon	0.34	0.1860	0.3	0.45	0.4
Karar Ağaçları	0.41	0.22	0.38	0.51	0.49
Rastgele Orman	0.4	0.19	0.39	0.42	0.49
Karar Destek Makinesi	0.4	0.28	0.44	0.53	0.54
Karar Destek Makinesi Tuned	0.53	0.17	0.66	0.17	0.66
XGBoost	0.4	0.18	0.41	0.37	0.5
XGBoost Tuned	0.62	0.38	0.67	0.48	0.71

Test verisi sonuçlarına bakıldığında ise, kurulan tüm modellerin içerisinde en iyi sonucu XGBoost_tuned metodunun verdiğini görmekteyiz. Test kümesinin doğruluk oranı %62 olarak bulunurken, doğru durumları tahmin etme oranı da %67 olarak elde edilmiştir. En iyi ikinci yöntem ise %53 doğruluk oranı ile Destek Vektör Makinesi_tuned olmuştur.

IV. TARTIŞMA

Çalışmada, elde edilen EEG kayıtları, literatürde geçerliliği olan popüler sınıflandırma metotları kullanılarak, doğruluk, duyarlılık,kesinlik, seçicilik ve F1 performans metrik değerleri elde edilmiştir. TABLO IV ve TABLO V de görüleceği gibi modellerin çoğu eğitim datasıyla iyi sayılabilecek hata değeri ve tahmin gücü yakalamalarına rağmen test verisine aşırı uyum yaşamakta ve test verisi tahmin etmekte güçlük çekmektedir. Bu durumun en önemli sebebi ise elde edilen eğitim verisinin kısıtlı olmasıdır. Performans metriklerinin iyileştirilmesi adına, farklı sınıflandırma metotları kullanılarak EEG işaretlerinin sınıflandırılmasına yönelik çalışmalar yapılması amaçlanmaktadır.

KAYNAKLAR

[1] *Baillet S., Mosher, C. J., and Leahy, M. R. "Electromagnetic brain mapping", IEEE Signal Processing Magazine, USA, volume:18, November 2001,pp:14–30*

[2] *Duru, Adil Deniz . "Determination of Increased Mental Workload Condition From EEG by the Use of Classification Techniques". International Journal of Advances in Engineering and Pure Sciences 31 / 1 (Mart 2019): 47-52. https://doi.org/10.7240/jeps.459420*

[3] *Graimann, B. Allison, Z. B., Pfurtscheller G.," Brain-Computer Interfaces: Revolutionizing Human-Computer Interaction" Springer Science & Business Media, 2010*

[4] *Chen, G.and Hou, R., "A New Machine Double-Layer Learning Method and Its Application in Non-Linear Time Series Forecasting," in International Conference on Mechatronics and Automation, 2007. ICMA 2007, 2007, pp. 795 –799.*

[5] *Nilsson,N. J."Introduction to machine learning An early draft of a proposed textbook."(1996)*

[6] *Kevin, M. P., "Machine Learning A Probabilistic Perspective",London, 2012, pp:1-13*

[7] *Murthy, K. S., ''Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey'', Data Mining and Knowledge Discovery, vol.2, pp:345-389, December 1998*

[8] *İbrahim, T. H., Mazher, J.W.,Ucan, O.N., Bayat, O., "A grasshopper optimizer approach for feature selection and optimizing SVM parameters utilizing real biomedical data sets" Neural Computing and Applications, Accepted: 3 March 2018*

[9] *He, H. and Ma, Y.,, ''Imbalanced Learning: Foundations, Algorithms, and Applications'', IEEE Press Wiley, 2013, cit. on p. 12.*

[10] *Steinwart, I. and Christmann, A., "Support Vector Machines", Springer Publishing Company, Incorporated, 1st ed., 2008.*

[11] *Benyamin, ,D."A gentle introduction to random forests, ensembles, and performance metrics in a commercial system", pp: 13,2012*

[12] *Hosmer W. D. and Lemeshow, S. "Applied Logistic Regression", Wiley New York, 2 edition, 2000*

[13] *Friedman, H. J.,"Greedy Function Approximation: A Gradient Boosting Machine," The Annals of Statistics, Institute of Mathematical Statistics , 2001, pp. 1189-1232.*

[14] *Chen,T and Guestrin,C. in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16 (ACM Press, New York, New York, USA, 2016) pp. 785–794.*

[15] *Farah, S., Duru, D.A., Bayat,.O.,"Classification of Breast Cancer Using Data Mining ", American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS), ISSN (Online) 2313-4402*

[16] *Lopéz, V. and Albert,F., "An insight into classification with imbalanced data: Empirical results and current trendson using data intrinsic characteristics", Information Sciences, pp:113-141,2013*

[17] *Nitesh,V.C., B. W. Kevın, W. B., Lawrence, O.H. and Philip, K.,"SMOTE: Synthetic Minority Over-sampling Technique Nitesh," Journal of Artificial Intelligence Research, no. 16, pp. 732-735, 2002.*

[18] *Hüseyin Akbulut ; Selen Güney ; Hasan Birol Çotuk ; Adil Deniz Duru, Classification of EEG Signals Using Alpha and Beta Frequency Power During Voluntary Hand Movement, 10.1109/EBBT.2019.8741944, 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*