Churn Analizinde Sınıflandırma ve Kümeleme Tekniklerinin Uygulanması

The Implementation of Classification and Clustering Techniques on Churn Analysis

Ahmet ELBİR[1], Hamza Osman İLHAN[1], Mehmet Furkan AYDIN[1], Yunus Emre DEMİRBULUT[1]
[1]Bilgisayar Mühendisliği Bölümü, Yıldız Teknik Üniversitesi, İstanbul, Türkiye
{aelbir, hoilhan}@yildiz.edu.tr
{y.emre.demirbulut, aydin.mehmetfurkan}@gmail.com

***Özetçe*— Telekomünikasyon firmalarının en büyük sorunlarından biri, firmalar arası potansiyel müşteri transferleridir. Bu sorunun önüne geçmek amacıyla ayrılma ihtimali olan müşterilerin önceden tespit edilmesi büyük önem taşır. Yapılan çalışmada, churn analizi olarak belirtilen potansiyel müşteri ayrılma eğilimlerinin analizleri üzerinde makine öğrenmesi tekniklerinden sınıflandırma ve kümeleme algoritmalarının başarımları ölçülmüş ve karşılaştırılması yapılmıştır. Sınıflandırma tekniklerinden K en yakın komşular, karar ağaçları, rastgele ormanlar, Destek Vektör Makineleri ve Naïve Bayes yöntemleri, kümeleme yöntemlerinden ise K-ortalama, Hiyerarşik kümeleme yöntemleri uygulanmıştır. Yöntemlerin başarımları hata oranı, kesinlik, duyarlılık ve F-ölçütü performans ölçütlerine göre değerlendirilmiştir.**

***Anahtar Kelimeler*—Churn Analizi; Makine Öğrenmesi Teknikleri; Sınıflama; Kümeleme**

***Abstract*— One of the most important problems of telecommunication companies is the potential transfer of customers between the firms. In order to avoid this problem, it is very important to identify customers who are likely to leave. In this study, the performance of the classification and the clustering algorithms in machine learning techniques has been evaluated and compared on the analysis of potential customer trends, which have been reported as churn analysis. K nearest neighbors, decision trees, random forests, support vector machines and naïve bayes methods were tested in scope of classification idea. Additionally, K-Means and hierarchical clustering methods were tested. The performances of the methods have been evaluated according to the accuracy, precision, sensitivity and F-measure performance metrics.**

Keywords—Churn Analysis; Machine Learning; Classification; Clustering

I. GİRİŞ

Yıllar içerisinde büyük gelişim göstermiş telekomünikasyon sektöründe, giderek artan firma sayısı ve kısıtlı müşteri ortamından dolayı telekomünikasyon firmaları arasındaki rekabet büyük oranda artmıştır. Yeni müşterilere ulaşmadaki zorluklar ve var olan müşterilerin firma değiştirmesindeki maliyetin yüksekliği firmaları yeni arayışlara itmiştir. Bu rekabeti takiben firmaların en büyük hedefleri; yeni framüşteriler kazanmakla beraber, var olan müşterilerinin rakip firmalara geçişinin önüne geçmek olmuştur. Firmalar açısından churn analizinin en önemli amaçları: firma değiştirmek üzere olan müşterilerin tahmin edilmesi ve ayrılması muhtemel müşterilerin elde tutulması için neler yapılabileceğinin tahmin edilmesidir [1].

Churn konusunda önceki yıllarda yapılan birçok çalışmada, değişkenler üzerindeki ilişkileri deneysel olarak araştıran bir churn modeli oluşturmaktan ziyade, belirli özelliklerin kullanıcının ayrılması üzerindeki etkisi incelenmiştir [1, 2]. Ancak iyi bir churn analizi bir müşterinin rakip firmaya geçme kararını önceden öngörebilme yeteneğine sahip olmalıdır. Bunun yanında churn analizinde olası ayrılma eğilimi olan müşteriler için firmalar, müşterileri kaybetmemeleri için uygulayacakları kampanyalar hakkında yol gösterebilmelidir [3]. Bunun gibi karmaşık analizlerin yapılma gerekliliği, firmalardaki veri setlerinin büyümesine, böylece klasik istatistik tekniklerinin kullanımındaki zorlukların artmasına neden olmuştur. Böylece, büyük veri ile uğraşılan her analizde olduğu gibi churn analizinde de makine öğrenmesi konularının kullanımını gerekli kılmıştır [4].

Literatürde, makine öğrenmesi ile yapılan churn analizlerinde farklı yaklaşım teknikleri bulunmaktadır. Lineer regresyon [5], karar ağaçları [6], yapay sinir ağları [7] gibi teknikler farklı veri setleri üzerinde çalışılmıştır. Churn analizi konusundaki engellerden biri kullanıcı veri setlerinin kolayca ulaşılabilir olmamasıdır. Telekomünikasyon firmaları bu konuda araştırma yapan kişi ve kuruluşlara müşterilerinin basit demografik bilgileri (isim, yaş, yaşadığı şehir vb.) dışında çok fazla bilgi vermemektedir. Bu yüzden literatürde gerçekleştirilen çalışmalar daha öznel kalmaktadır. Sunulan bu çalışmada ise daha önceden halka açık olarak paylaşılan veri seti kullanılacaktır.

Bu çalışma ASYU-2017 Konferansında sunulmuş ve sadece özet olarak
Bildiri Özetleri Kitabı (ISBN: 978-605-84722-8-0)'nda yayımlanmıştır.

Sunulan çalışma, makine öğrenmesi tekniklerinden sınıflama ve kümeleme yöntemleri müşteri kaybının önüne geçmek için ayrılma ihtimali yüksek olan müşterilerin tahmininin yapılmasındaki başarımlarının ölçülmesini içermektedir. Bu anlamda sınıflandırma tekniklerinden, K en yakın komşular, karar ağaçları, rastgele ormanlar, Destek Vektör Makineleri ve Naïve Bayes yöntemleri, kümeleme yöntemlerinden ise K-ortalama, Hiyerarşik kümeleme yöntemlerinin başarımları sunulmaktadır.

Bildiri düzeni şu şekilde olacaktır: uygulanan makine öğrenmesi teknikleri ve kullanılan veri setine ait bilgiler Bölüm 2'de verilecektir. Bölüm 3'de, elde edilen sonuçlara değinilecektir. Son olarak ise yöntemlerin Churn analizi üzerindeki başarımları yorumlanacaktır.

II. MATERYAL VE UYGULANAN YÖNTEMLER

A. Veri Seti ve Önişleme Süreci

Telekomünikasyon firmaları müşteri verilerinin gizliliği konusundaki yasalardan dolayı CRM (Customer Relationship Management) verilerini geliştiricilerle veya analistlerle paylaşmamaktadır [8]. Çalışma, daha önceden yurtdışındaki telekomünikasyon firmalarının halka açık paylaştığı veri seti üzerinde gerçeklenmiştir.

Veri seti 3333 müşterinin bilgilerini içermektedir. Müşterilerin %15'i firma değiştirmiş müşterileri oluşturmaktadır. Veri setinde hiçbir özellik eksik eleman içermemektedir [8,9]. Veriler kategorik ve numerik değerlerden oluşmaktadır. Müşterilerin telefon numarası, günün çeşitli saatlerindeki konuşma süreleri ve ücretleri ile adres bilgisi gibi demografik bilgileri bulunmaktadır. Çalışma kapsamında veri seti üzerinde makine öğrenme teknikleri test edilmeden önce, telefon numarası gibi sonuç üzerinde etkisi olmayan gereksiz sütunlar veri setinden çıkartılarak sınıflama/kümeleme için daha gerekli özelliklerin kullanılması amaçlanmıştır. Telefon numarası gibi özellikler her kullanıcı için farklı olduğundan sonuç üzerine etkisi olmayacaktır. Ayrıca kategorik tipte özellikler makine öğrenmesi kapsamında özellik olarak kullanılabilmesi için mantıksal (boolean) tipine dönüştürülmüştür. Sınıflandırma ve kümeleme yöntemlerinde başarım hesabı yapabilmek için veri seti ara yüz içerisinden k katlamalı çapraz doğrulama (k-fold cross validation) kullanabilmektedir. Bunun dışında kullanıcı isterse veri setinin belirli bir yüzdesini eğitim ve test veri kümeleri olarak belirleyebilmektedir.

B. Sınıflandırma Yöntemleri

1) Karar Ağaçları: Eğiticili öğrenmede kullanılan model tabanlı algoritmalardan bir tanesidir. Ağaç veri yapısı kullanılarak özellikler ve özelliklere ait değerler sınıflandırma yapmak için eğitilir. Ağaçta olması gereken en önemli özellik, en belirleyici olan özelliğin mümkün oldukça en üst seviyede olmasıdır. En belirleyici özelliğin tespitinde bilgi kazancı hesabı gerçekleştirilir. Daha detaylı bilgi için [10] incelenebilir.

2) K-En Yakın Komsular: Sınıflandırma sırasında çıkarılan özelliklerden, sınıflandırılmak istenen yeni bireyin daha önceki bireylerden k tanesine yakınlığına bakılır. Yakınlık hesaplama işleminde ise k-means ve hiyerarşik kümelemede kullanılan öklid uzaklığı, manhattan uzaklığı gibi mesafe hesaplama yöntemleri kullanılabilir. Uygulamada farklı *K* değerleri (3, 5, 7, 15) test edilmiştir.

3) Naïve Bayes: Olasılık tabanlı algoritmalardan birisidir. Temel olarak veri setinden elde edilecek önceki olasılık değerlerine bakarak gelecek olasılık değerlerini hesaplamaya ve bu sayede sınıflandırma yapmaya dayalıdır.

4) Destek Vektör Makineleri: Model tabanlı yöntemlerde bir diğeri de Destek Vektör Makineleridir. Temel olarak iki sınıfı ayırabilmek için bu iki sınıfın sınır bölgesi üzerinde optimizasyon yaparak ayırt edici doğrusal veya doğrusal olmayan fonksiyon üretme işlemidir. Çekirdek tabanlı öğrenme tekniğidir [11].

5) Rastgele Ormanlar: Bir diğer model tabanlı algoritmlardan olan rastgele ormanların temelini karar ağaçları oluşturur. Bu yöntemde temel öğrenici olarak karar ağacı seçilir ve öğrenme işlemi birden fazla karar ağacının oluşturulması ve oluşturulan bu ağaçların birleştirilmesiyle yeni bir öğrenicinin meydana getirilmesi işlemlerini takip eder[12].

C. Kümeleme Yöntemleri

1) K-Ortalamalar: Eğiticisiz öğrenme yöntemlerinden olan K-Ortalamalar algoritması uzaklık tabanlı kümeleme algoritmalarından bir tanesidir. Bir veri setinin k adet gruba ayrılması için uzaklık ve ağırlık merkezi değerlerini kullanır ve veri setinin etiketlenmesini sağlar.

2) Hiyerarşik: Veriler arasındaki uzaklık ve benzerlik değerlerini kullanarak çeşitli seviyelerde hiyerarşiler oluşturulur. Bu oluşturulan hiyerarşileriden faydalanarak veri setinin kaç gruba ayrılacağı dendogram vb grafiksel çizimler kullanılarak veriler üzerinde etiketleme yapılabilir [13].

D. Performans Ölçütleri

Genel olarak sınıflandırma veya kümeleme uygulamalarında performans metrikleri, Şekil 1'de sunulan karışıklık matrisinden türetilir. True Pozitif (TP), oluşturulan makine öğrenmesi modeli vasıtasıyla örneklerin doğru sınıflara atandığı durumları belirtmektedir. Sınıflandırılmasında yanlış gruplanmış örnekler False Pozitif (FP), diğer bir deyişle Tip I hatası altında toplanmıştır. False Negatif (FN), istenen sınıfta yanlış sınıflandırılmış örnekleri belirtir ve bu da Tip II

Bu çalışma ASYU-2017 Konferansında sunulmuş ve sadece özet olarak
Bildiri Özetleri Kitabı (ISBN: 978-605-84722-8-0)'nda yayımlanmıştır.

hatası olarak adlandırılır. Son notasyon, istenmeyen sınıflardaki istenmeyen örnekler için doğru sınıflandırmaların numerik başarımları True Negatif (TN) olarak gösterilmektedir.

		Öngörülen Sınıf	
		Sınıf=1	Sınıf=0
Doğru Sınıf	Sınıf=1	a	b
	Sınıf=0	c	d

a: TP (True Pozitif) c: FP (False Pozitif)
b: FN (False Negatif) d: TN (True Negatif)

Şekil 1. Karşıtlık Matrisi

Sunulan çalışmada, makine öğrenme yöntemlerinin churn analizinde başarımları, karşıtlık matrisi üzerinden hesaplanan doğruluk, kesinlik, duyarlılık ve F-score değerleri ile sunulmuştur.

Denklem (1)'de belirtilen doğruluk (Accuracy) ölçütü, makine öğrenmesi uygulamalarında esas ve temel performans metriğidir. Bütün test seti içindeki doğru sınıflandırılmış örneklerin yüzdesini ifade eder. %100'lük bir doğruluk, test setindeki verilen örneklerin hepsinin doğru olarak sınıflandırıldığını gösterir. Bununla birlikte, doğruluk metriğinde elde edilen yüksek sonuçlar, modelin başarısını tam olarak veremez. Doğruluk yalnızca gerçekten doğru sınıflandırılmış örnekleri belirtir. Ancak, doğruluk paradoksu [14] olarak belirtilen durumlarda model başarısının tam olarak kanıtlanabilmesi için karışıklık matrisinin tüm dağılımının incelenmesi gerekmektedir.

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

Modelin, Tip I ve Tip II hataları açısından değerlendirilmesi için Duyarlılık (Sensitivity) ve hassaslık (Precision) diğer performans metrikleri olarak kabul edilmektedir. Duyarlılık ve hassaslık sırasıyla denklem (2) ve (3) ile hesaplanabilir.

$$Duyarlılık = \frac{TP}{TP + FN} \quad (2)$$

$$Hassaslık = \frac{TP}{TP + FP} \quad (3)$$

F-score, denklem (4) kullanılarak hesaplanabilir. Elde edilen sonuç, 1'de en iyi değeri gösterirken, en kötü skor 0 noktasını temsil etmektedir. F-score metriğinde, FP ve FN'nin sonuçlara dâhil edilmesinden dolayı doğruluk oranlarına göre daha güvenilirdir

$$FMeasure = 2 \times \frac{Precision \times Sensitivity}{Precision + Sensitivity} \quad (4)$$

III. UYGULAMA VE SONUÇLAR

Sınıflandırma başarımlarını test etmek için kullanılan teknikler, K-En Yakın Komşular, Destek Vektör Makineleri, Karar Ağacı, Naive Bayes ve rastgele ormanlar algoritmalarıdır. Yukarıda verilen yöntemler veri setinin tamamı üzerinde carpraz doğrulama yöntemi kulanılarak test edilmiştir. Test sırasında kullanılan carpraz doğrulama degeri 10 seçilmiştir. Sonuçlar Tablo 1'de sunulmuştur.

Tablo 1. Sınıflandırma Algoritmalarının Başarım Değerleri

	Acc.	**Pre.**	**Rec.**	**F-Score**
Karar Ağacı	0.94	0.94	0.98	0.97
K-En Yakın K.	0.89	0.90	0.99	0.94
Naive Bayes	0.86	0.92	0.92	0.92
DVM	0.85	0.86	1	0.92
Rast. Ormanlar	**0.94**	**0.95**	**0.99**	**0.97**

Algoritmalar arası karşılaştırmalar dışında; algoritma parametrelerinin sonuçlar üzerindeki etkileri de çalışmada incelenmiştir. K En Yakın Komşular algoritması üzerinde komşuluk sayısının ve uzaklık ölçüsünün sonuçlar üzerindeki etkileri incelenmiştir. Komşuluk sayısı arttıkça algoritmanın dogruluk oranının arttırdığı gözlemlenmiştir. Uzaklık ölçüleri arasında yapılan testlerde "euclidean" ve "manhattan" ölçümleri birbirine yakın sonuçlar vererek en iyi uzaklık ölçüsü olarak degerlendirilmiştir. Karar ağaçları tekniği uygulamasında parametre olarak dallanma aşamasında kullanılan kriter ve dallanmada seçilecek özelliğin random seçilip seçilmemesi test edilmiştir. Dallanma kriteri olarak "gini" ve "entropy" kavramları test edilmiştir. Dallanmada seçilecek özellğin rastgele seçilmesi yerine "entropy" ve "gini" degerleri en yüksek çıkan özelliğin root olarak seçilmesi ile daha iyi sonuçlar alınmıştır. Rastgele Ormanlar algoritmasının testi sırasında oluşturulan orman için yaratılan ağaç sayısının sonuçlara etkisi de incelenmiştir. Ağaç sayısındaki artışın algoritmanın dogruluk değerini arttırdığı gözlemlenmiştir.

Churn verisinin etiketli olmadığı veri setlerinde verilen veri setini 2 kümeye ayırıp Churn tahmini yapılabilmesi için K-Means ve Hiyerarşik kümeleme algoritmaları test edilmiştir. Kullanılan veri seti etiketli bir veri seti oldugundan etiketler silinip kümeleme algoritmasından yeni etiketler elde edilir. Daha sonra bu yeni etiketler orjinal sınıf verisi ile karşılaştırılarak algoritmaların

Bu çalışma ASYU-2017 Konferansında sunulmuş ve sadece özet olarak
Bildiri Özetleri Kitabı (ISBN: 978-605-84722-8-0)'nda yayımlanmıştır.

başarım yüzdesi elde edilir. Deneysel testlerle elde edilen sonuçlar Tablo 2'de gösterilmiştir.

Tablo 2. Kümeleme Algoritmalarının Başarım Değerleri

	Accuracy	Precision	Recall	F-Score
K-Means	0.534	0.89	0.52	0.66
Hiyerarşik	**0.854**	**0.86**	**1**	**0.92**

IV. TARTIŞMA

Elde edilen sonuçlara göre, Rastgele Ormanlar algoritmasının accuracy yüzdesine göre %94.7 oranla en iyi sonucu verdigi gözlemlenmiştir. Sonrasında ise %94.3 ile Karar ağaçları gelmektedir. SVM algoritması çok özellikli veri setlerinde gösterdigi düşük performans nedeniyle %85.5 ile en başarısız algoritma olmuştur. Rastgele ormanlar algoritmasının en iyi sonuç vermesindeki ön önemli etken birden fazla sınıflandırıcı üreterek, bu sınıflandırıcıların tahminlerini kullanarak en iyileme sınıflaması üzerine çalışmasıdır. Naive Bayes algoritması sınıflandırma işleminde olasılıkçı yaklaşım kullanır. Bu yaklaşım nedeniyle sınıflandırmada kullanılacak her bir özelliğin bağımsız olması gerekliliği doğar. Bu yaklaşım naive bayes algoritmasının kullanım alanı kısıtlamasına rağmen kategorik verilerin ağırlıklı olduğu veri setlerinde başarımı oldukça yüksektir.

K-Ortalamalar algoritmasının gürültülere karşı hassasiyeti yüksek olduğundan başarım yüzdesi düşük çıkmaktadır. K-Ortalamalar algoritmasının başsarım yüzdesi bu hassasiyet dolayısıyla churn analizi işleminde yeterli düzeyde degildir. Doğruluk yüzdesi %53.4 olup sınıflandırma algoritmalarının performans başarımına yaklaşamamıştır. Hiyerarşik kümeleme algoritmasında ise tek baglantı tekniği ile elde edilen sonuçlar kıyaslama yapılan kümeleme algoritmasının oldukça önünde çıkarak %85.4 doğruluk ölçümü yapılmıştır. Ancak bütün teknikler bir araya getirilip kıyaslama yapıldıgında, inceleme yapılan veri seti üzerinde sınıflandırma tekniklerinin kümeleme tekniklerinden önde olduğu görülmektedir.

KAYNAKÇA

[1] Hadden, John, et al. "Churn prediction: Does technology matter." International Journal of Intelligent Technology 1.2 (2006): 104-110.

[2] Hung, Shin-Yuan, David C. Yen, and Hsiu-Yu Wang. "Applying data mining to telecom churn management." Expert Systems with Applications 31.3 (2006): 515-524.

[3] Ahn, Jae-Hyeon, Sang-Pil Han, and Yung-Seop Lee. "Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry." Telecommunications policy 30.10 (2006): 552-568.

[4] Fan, Wei, and Albert Bifet. "Mining big data: current status, and forecast to the future." ACM sIGKDD Explorations Newsletter 14.2 (2013): 1-5.

[5] Owczarczuk, Marcin. "Churn models for prepaid customers in the cellular telecommunication industry using large data marts." Expert Systems with Applications 37.6 (2010): 4710-4712.

[6] Wei, Chih-Ping, and I-Tang Chiu. "Turning telecommunications call details to churn prediction: a data mining approach." Expert systems with applications 23.2 (2002): 103-112.

[7] Tsai, Chih-Fong, and Yu-Hsin Lu. "Customer churn prediction by hybrid neural networks." Expert Systems with Applications 36.10 (2009): 12547-12553.

[8] Ngai, Eric WT, Li Xiu, and Dorothy CK Chau. "Application of data mining techniques in customer relationship management: A literature review and classification." Expert systems with applications 36.2 (2009): 2592-2602.

[9] Umayaparvathi, V., and K. Iyakutti. "Applications of data mining techniques in telecom churn prediction." International Journal of Computer Applications 42.20 (2012): 5-9.

[10] Quinlan, J. Ross. "Induction of decision trees." Machine learning 1.1 (1986): 81-106.

[11] Joachims, Thorsten. "Text categorization with support vector machines: Learning with many relevant features." Machine learning: ECML-98 (1998): 137-142.

[12] Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.

[13] Corpet, Florence. "Multiple sequence alignment with hierarchical clustering." Nucleic acids research 16.22 (1988): 10881-10890.

[14] X. Zue, I. Davidson, "Knowledge discovery and data mining," In-formation Science Reference, 2007.

Bu çalışma ASYU-2017 Konferansında sunulmuş ve sadece özet olarak
Bildiri Özetleri Kitabı (ISBN: 978-605-84722-8-0)'nda yayımlanmıştır.