Hibrid Sınıflandırma Yöntemleriyle Kredi Risk Analizi: Alman ve Türk Kredi Verisetleri Üzerinde Vaka Çalışmaları

Credit Risk Analysis based on Hybrid Classification: Case Studies on German and Turkish Credit Datasets

Erkan Çetiner ve Prof. Dr. Taşkın Koçak
Fen Bilimleri Enstitüsü
Bahçeşehir Üniversitesi
İstanbul, Türkiye
ecetiner87@hotmail.com, taskin.kocak@eng.bau.edu.tr

Doç. Dr. V. Çağrı Güngör
Bilgisayar Mühendisliği
Abdullah Gül Üniversitesi
Kayseri, Türkiye
cagri.gungor@agu.edu.tr

***Özetçe* — Kredi risk analizi, karar verme süreçleri açısından finans sektöründe önemli bir rol oynamaktadır. Bankalar ve finansal kuruluşlar, müşterilerinden büyük ölçeklerde ham veri toplamaktadırlar. Veri madenciliği teknikleri, bu ham veri içerisinden kullanışlı bilgiler edinmek amacıyla kullanılabilir. Destek-vektörleri, yapay sinir ağları ve bayesian yaklaşımı bu alanda hali hazırda kullanılan sınıflandırma yöntemleridir. Bu çalışmada, farklı tekil sınıflandırma yöntemlerinin bir araya getirilerek hibrid bir yaklaşımla, sınıflandırma sonuçlarının doğruluğunun arttırılması hedeflenmiştir. Farklı kombinasyonlar ayrıca sınıflandırma yetkinliği açısından performans karşılaştırılmasına tabi tutulmuştur. Hem Alman kredi veriseti hem de ulusal bir bankadan alınan veriseti üzerinde ilgili yaklaşım çalıştırılmış ve yöntemin genelleştirilebilme özelliğinin görülmesi de amaçlanmıştır. Deney sonuçları, özellik seçiminin sınıflandırma başarımı ve hesaplama zamanı açısından çok önemli olduğunu, hibrid yaklaşımın tekil sınıflandırma yöntemlerine göre sınıflandırma doğruluğu açısından daha iyi sonuçlar verdiğini ve son olarak radial-basis fonksiyonu ile birlikte kullanıldığında destek-karar vektörlerinin hem tekil hem hibrid yaklaşımlar içerisinde en iyi sınıflandırma başarımına sahip olduğunu göstermiştir.**

Anahtar Kelimeler — kredi riski, hibrid-sınıflandırma, özellik seçimi

***Abstract* — In finance sector, credit risk analysis plays a major role in decision process. Banks and finance institutions gather large amounts of raw data from their customers. Data mining techniques can be employed to obtain useful information from this raw data. Several data mining techniques, such as support-vector machines (SVM), neural networks, naive-bayes, have already been used to classify customers. In this paper, we propose hybrid classification approaches, which try to combine several classifiers and ensemble learners to boost accuracy on classification results. Furthermore, we compare these approaches' performance with respect to their classification accuracy. We work with two diverse datasets; namely, German credit dataset and Turkish bank dataset. The goal of using such diverse dataset is to show generalization capabality of our approaches. Experimental results provide three important consequences. First, feature selection stage has a major role both on result accuracy and calculation complexity. Second, hybrid approaches have better generalability over single classifiers. Third, using SVM-Radial Basis Function (RBF) as the base classifier and a hybrid model member gives the best accuracy and type-1 accuracy results among others.**

Keywords — credit risk, hybrid-classifier,feature selection

I. GİRİŞ

Kredi riski, bir finansal kuruluşun müşterisine kredi açtığı andan itibaren geçerli olan risktir. Planlı bir şekilde yönetilmeyen krediler hem müşteriler hem finansal kuruluşlar açısından sonu krizleri de beraberinde getirebilecek durumlara yol açabilir [1]. Bazı müşterilerin aldıkları krediyi geri ödeyebilecek yetkinliği olmadığından, finansal kuruluşlar kendilerine gelen başvuruları değerlendirme aşamasında kredi risk analizi yöntemlerini kullanırlar [2]. Bu yöntemlerin kullanılması ile gelecekteki olası kayıpların minimize edilmesi ve finansal açıdan gelirlerin planlanan seviyede tutulması hedeflenir.

Kredi risk analizine yardımcı olacak şekilde veri madenciliği yöntemleri yaygın bir şekilde kullanılmaktadır. Gallo'nun çalışmaları yapay sinir ağlarının ekonomik modellemede oldukça kullanışlı olduğunu göstermiştir [3]. Shachmurove, yapay sinir ağlarının kompleks modelleri hızlı ve yüksek doğrulukla analiz ettiğini deneylerinde ortaya çıkarmıştır [4]. Satchidananda karar-ağaçlarının ve lojistik regresyon analizinin verimliliğini karşılaştırmış ve karar-ağaçlarının daha iyi sonuçlar verdiğini belirtmiştir [5]. Yu, Huang, Cai ve Hu karşılaştırmalı bir çalışmayla lojistik regresyon, destek-vektör makinaları ve karar-ağaçlarının performanslarını incelemiş [6], benzer bir çalışmayı füzyon mantığınıda içerecek şekilde Wong, Wang ve Lai de sunmuştur [7]. Doumpos ve Zopounidis ise farklı sınıflandırma tekniklerini istifleme mantığıyla bir araya getirmiş ve daha iyi genellemelere varmışlardır [8]. Bu bildirideki çalışmanın da altyapısını oluşturan bir diğer çalışmada, Kotsiantis hybrid bir model sunmuş ve farklı hibridleme yöntemleriyle kredi riski analizi çalışmaları yapmıştır [9].

Güncel çalışmalar toplu öğrenme yöntemleriyle doğruluk başarımını arttırma hedefindedir ve hibrid yaklaşımlarla uygulanıp tekil sınıflandırma yöntemlerine göre daha iyi sonuçlar vermektedir [10]. Ayrıca özellik seçimi yöntemlerinin sınıflandırma performansına katkısı yadsınamaz şekilde kabul edilmiştir.

Bu bildiride anlatılan çalışmada, farklı hibridleme yöntemlerinin sınıflandırma üzerindeki etkilerini görmek ve doğruluk başarımlarının karşılaştırılması sonucu performans açısından kredi risk analizi alanında kullanılabilecek olası modelleri tespit etmek amacı güdülmüştür.

978-1-5386-1501-0/18/$31.00 ©2018 IEEE

II. ÖN ÇALIŞMA VE MODELLEME

A. Özellik Seçimi Algoritmaları

Özellik seçimi bir sınıflandırma probleminde, veriseti içerisinde yer alan ve karar aşamasına etki eden en iyi özellik altsetini bulmayı heeflemektedir. Yapılan çalışmalarda özellik seçimi sayesinde veriseti üzerindeki hesaplama zamanının azaltıldığı, doğruluk oranının yükseldiği ve aşırı-uygunluk problemlerinin önüne geçildiği görülmüştür. *InfoGain*, *GainRatio* ve *Chisquared* algoirtmaları çalışmada kullanılan verisetleri üzerinde uygulanmış ve doğruluk ölçütü açısından en iyi sonucu veren algoritmanın sunduğu altset çalışmanın geri kalanında kullanlmıştır.

B. Sınıflandırma ve Kümeleme Yöntemleri

Özellik seçimi çalışması sonrası elde edilen altset üzerinde çeşitli sınıflandırma ve kümeleme yöntemleri hem tekil hem de hibrid yaklaşımlarla uygulanmıştır. Hibrid yaklaşımların detaylı bilgisi bir sonraki bölümde anlatılmıştır. Çalışmada kullanılan sınıflandırma ve kümeleme yöntemleri aşağıda sıralanmıştır:

1) Sınıflandırma Yöntemleri:

a) Temel Sınıflandırma Yöntemleri: Yapay sinir ağları (ANN), Destek-Vektör Makinaları (SVM), Naive – Bayes yaklaşımı, Gizli Markov Modellemesi (HMM), Öz Organize Haritalar (SOM), Karar Ağaçları –ADTree, J48, Random Forest-

b) Toplu Öğrenme Yöntemleri: Arttırım(Boosting), Çantalama(Bagging), Kümeleme ile Sınıflandırma(ClassificationviaClustering).

2) Kümeleme Yöntemleri:

a) K-Means, EM, DBSCAN.

C. Verisetleri ve Performans Karşılaştırma Ölçütleri

1) Verisetleri:

Alman kredi veriseti genel kullanıma açık bir kredi veribankasından elde edilmiş ve üzerinde deneyler yapılmıştır. İlgili verisetinde 1000 müşterinin 20 farklı özelliği bulunmaktadır ve bu müşterilerden 700 tanesi "iyi" –kredi geri ödemelerini zamanında gerçekleştiren- 300 tanesi ise "kötü" –kredi geri ödemelerini zamanında gerçekleştiremeyen- olarak etikeketlenmiştir. Veriseti özellikleri arasında müşterinin kredi geçmişi, hesap bilgileri, taşınmaz malları, hali hazırdaki kredileri, kredi vade bilgisi gibi finansal özellikler yanı sıra, evlilik durumu, sahip olduğu ev/araba sayısı, yaşı, çocuk sahibi olup olmadığı şeklinde demografik özellikleri de bulunmaktadır [11].

Ulusal bankadan elde edilen ve müşterilerin farklı özelliklerini tutan verisetinde ise 15740 müşterinin 96 özellik içerecek şekilde bilgisi tutulmaktadır [12]. Bu müşterilerin bilgileri kişisel verilerin korunumu göz önüne alınarak anonimleştirilmiş ve veriseti üzerindeki çalışmalar bu anonim veri üzerinde gerçekleştirilmiştir. İlgili veriseti büyük ölçekli olduğundan, sınıflandırma başarımının yüksek doğrulukla sonuçlanması amacıyla üzerinde deneyler yapılmadan önce; *eksik-değer içeren kayıtlar çıkarılmış*, *bazı özellikler bir araya getirilmiş(data integration)* ve *sürekli(continuous) değer içeren özellikler kategorik hale getirilmiştir*. Daha sonra modelleme kısmında anlatılan yaklaşım her iki verisetine de uygulanmıştır.

2) Performans Karşılaştırma Ölçütleri:

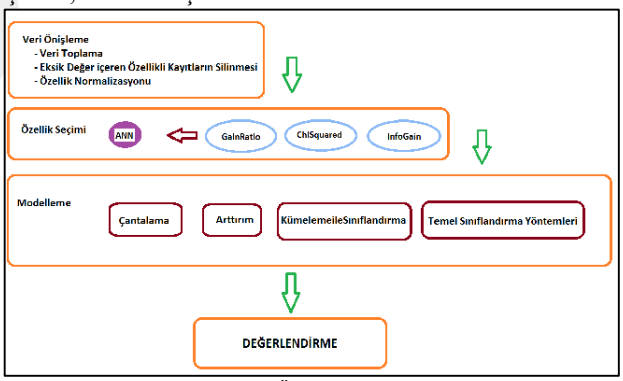
Deneyler sonunda elde edilen sonuçların, performans değerlendirmeleri açısından anlaşılabilir bir biçimde yorumlanabilmesi için; performans karşılaştırma metriklerinin düzgün bir şekilde belirlenmesi ve tanımlanması gereklidir. Bu çalışmada aşağıdaki performans metrikleri kullanılmıştır:

- **Doğru Pozitif (TP)**: **Good** olarak doğru etiketlenmiş müşterilerin tahmin edilme sayısı.
- **Yanlış Negatif (FN): Bad** olarak yanlış etiketlenen müşterilerin tahmin edilme sayısı. (Gerçekte Good olan müşterinin Bad olarak etiketlenme durumu.)
- **Yanlış Pozitif (FP): Good** olarak yanlış etiketlenen müşterilerin tahmin edilme sayısı. (Gerçekte Bad olan müşterinin Good olarak etiketlenme durumu.)
- **Doğru Negatif (TN): Bad** olarak doğru etiketlenmiş müşterilerin tahmin edilme sayısı.
- İlgili performans metrikleri ile aşağıdaki karşılaştırma ölçütleri her bir test için hesaplanmıştır:
- **Doğruluk** = (TP + TN)/(Toplam Kayıt Sayısı) – Bir algoritmanın olası müşterileri hangi oranda doğru sınıflandıracağı,
- **TYPE-1 Doğruluk** = FP/(Toplam Kayıt Sayısı) – Bir müşterinin normalde "kötü-bad" olmasına ragmen "iyi-good" olarak sınıflandırılma yüzdesi.

Deneyler yapılırken, doğruluk oranını yükseltme amacıyla çapraz-onaylama(cross validation) ve uzatma (holdout) metodları da kullanılmıştır.

D. Önerilen Model

Aşağıda yer alan şekil-1'de görüldüğü üzere önerilen modelde 3 ana aşama bulunmaktadır. Bunlar sırasıyla veri önişleme, özellik seçimi ve modelleme kısımlarıdır.



Şekil. 1.Önerilen Model

İlk olarak veri önişleme teknikleri kullanılarak ham haldeki veriseti daha kapsamlı, temiz ve kullanışlı hale getirilmiştir. İkinci aşamada, özellik seçimi algoritmaları yapay sinir ağları ile birlikte uygulanmış ve ilgili verisetinden sınıflandırma

amacına en uygun altset elde edilmiştir. Son aşamada ise, toplu öğrenme yöntemleri tekil sınıflandırma yöntemleri ile birlikte uygulanmış ve sonuçlarda tekil sınıflandırma yöntemlerinin "Doğruluk" ve "TYPE-1 Doğruluk" performanslarının aşılması hedeflenmiştir.

Toplu öğrenme tekniklerinin temel sınıflandırma yöntemleriyle hibrid şekilde uygulamasından elde edilen sonuçlar karşılaştırılmış, en iyi 4 sonucu veren deneylerde yer alan yöntemler aynı şekilde Alman verisetine de uygulanmıştır. Böylece önerilen hibrid yaklaşımın genellenebilirliği de test edilmiştir. Ayrıca hesaplama kompleksitesi de göz önüne alınmış ve her bir denemede belirtilmiştir.

III. DENEYLER VE SONUÇLAR

A. *Özellik Seçimi Sonuçları*

Özellik seçimi testleri, Ulusal bankadan alınan verisetinde uygulanmış, WEKA üzerinde InfoGain, GainRatio ve Chisquared özellik seçimi yöntemleri yapay sinir ağları ile birlikte test edilmiştir. Bu testler sonucunda büyük ölçekli verisetinden, sınıflandırmaya direkt etkisi olan özellikleri içeren altset elde edilmiştir.

TABLO I. INFOGAIN SONUÇLARI

Model – Yapay Sinir Ağları	Özellik Sayısı	Doğruluk(%)
1	10	74
2	15	74.2
3	20	74.6
4	25	74.6
5	30	74.6
6	40	74.9
7	50	74.9

TABLO II. GAINRATIO SONUÇLARI

Model – Yapay Sinir Ağları	Özellik Sayısı	Doğruluk(%)
1	10	81.1
2	15	81.2
3	20	81.2
4	25	81.2
5	30	81.9
6	40	81.9
7	50	81.9

TABLO III. INFOGAIN SONUÇLARI

Model – Yapay Sinir Ağları	Özellik Sayısı	Doğruluk(%)
1	10	78.9
2	15	78.9
3	20	77.9
4	25	77.9
5	30	77.8
6	40	77.8
7	50	77.5

Doğruluk performans ölçütü göz önüne alınarak, en iyi sonuç GainRatio'da elde edilmiş ve 96 özellikten oluşan veriseti, 40 özellikli bir altsete dönüştürülerek, çalışmanın geri kalanında kullanılmıştır.

B. *Toplu Öğrenme Yöntemleri – Hibrid Deney Sonuçları*

1) *Arttırım(Boosting) Sonuçları:*

Tablo-4'te yer alan sonuçlara bakıldığında, destek-vektör makinalarının diğer sınıflandırma algoritmalarını her iki performans ölçütü açısından da geride bıraktığı görülmektedir. Ayrıca karar-ağaçları, diğer temel sınıflandırma algoritmalarına göre daha düşük sonuçlar elde etmiştir.

TABLO IV. ARTTIRIM YÖNTEMİ İLE HİBRİD DENEY SONUÇLARI – ULUSAL BANKA VERİSETİ

Model	Toplu Öğrenme Yöntemi	Performans Ölçütleri		Hesaplama Kompleksitesi (dk)
		Doğruluk	Type I Doğruluk	
1	ANN	81.9	0.079	58
2	SVM-RBF	83.7	0.082	56,3
3	NaiveBayes	80.1	0.074	12,4
4	RandomForest	79.7	0.072	16,7
5	ADTree	79.7	0.072	16,7
6	J48	80	0.074	15,3

2) *Çantalama(Bagging) Sonuçları:*

Arttırım toplu öğrenme yöntemine göre çok az da olsa gelişme kaydetmiş ve yine kendi içerisinde, destek-vektör makinaları en iyi sonucu vermiştir. Karar-ağaçları ve diğer temel sınıflandırma algoritmaları açısından da arttırım sonuçlarına benzer davranışlar elde edilmiştir.

TABLO V. ÇANTALAMA YÖNTEMİ İLE HİBRİD DENEY SONUÇLARI – ULUSAL BANKA VERİSETİ

Model	Toplu Öğrenme Yöntemi	Performans Ölçütleri		Hesaplama Kompleksitesi (dk)
		Doğruluk	Type I Doğruluk	
1	ANN	82	0.079	58
2	SVM-RBF	83.9	0.082	56,3
3	NaiveBayes	80.1	0.074	12,4
4	RandomForest	79.9	0.072	16,7
5	ADTree	79.9	0.072	16,7
6	J48	80.1	0.074	15,3

3) *KümelemeileSınıflandırma(ClassificationviaClustering) Sonuçları:*

Kendi içinde K-means algoritmasının en iyi doğruluk sonuçlarını verdiğini, fakat daha önceki toplu öğrenme yöntemleri kadar sınıflandırma başarımı elde edemediğini söylemek mümkündür.

TABLO VI. KÜMELEMEİLESINIFLANDIRMA YÖNTEMİ İLE HİBRİD DENEY SONUÇLARI – ULUSAL BANKA VERİSETİ

Model	Toplu Öğrenme Yöntemi	Performans Ölçütleri		Hesaplama Kompleksitesi (dk)
		Doğruluk	Type I Doğruluk	
1	EM	81	0.078	23,4

2	K-Means	81.4	0.078	23,4
3	DBSCAN	78.5	0.076	25,6

4) Tekil Sınıflandırma Yöntemlerinin Sonuçları:

Destek-vektör makinalarının yine en iyi sonucu verdiği deneylerde, toplu öğrenim yöntemleriyle birlikte uygulanan modellere göre doğruluk açısından daha düşük sınıflandırma başarımı elde edilmiştir. Ayrıca Gizli Markov Modelinin diğer tekil sınıflandırma yöntemlerine göre en az destek-vektör makinaları kadar daha başarılı olduğu görülmüştür.

TABLO VII. KÜMELEMEİLESINIFLANDIRMA YÖNTEMİ İLE HİBRİD DENEY SONUÇLARI – ULUSAL BANKA VERİSETİ

Model	Tekil Sınıflandırma Algoritması	Performans Ölçütleri		Hesaplama Kompleksitesi (dk)
		Doğruluk	Type I Doğruluk	
1	SVM-RBF	83.4	0.081	48,5
2	ANN	81.9	0.079	54,4
3	NaiveBayes	79.9	0.078	10,4
4	HMM	82.8	0.08	12,5
5	SOM	80	0.079	35,3

5) Alman Veriseti Üzerindeki Sınıflandırma Sonuçları:

Yapılan testlerde en iyi sonuçları veren 3 hibrid model ve tekil sınıflandırma yöntemlerinde en iyi sonucu veren SVM sınıflandırma algoritması, Alman kredi veriseti üzerinde uygulanmış ve Ulusal bankadan alınan verisetine uygulanan testlerin sonuçlarına oldukça benzer davranışlar gözlemlenmiştir.

TABLO VIII. EN İYİ 4 SINIFLANDIRMA YÖNTEMİNİN ALMAN VERİSETİ ÜZERİNDEKİ SONUÇLARI

Model	Temel Sınıflandırma Algoritması	Performans Ölçütleri		Hesaplama Kompleksitesi (dk)
		Doğruluk	Type I Doğruluk	
Bagging	ANN	73.1	0.072	4 min
Bagging	SVM-RBF	74.1	0.074	3 min
Boosting	SVM-RBF	72.6	0.071	3 min
Tekil Sınıflandırma	SVM-RBF	73.5	0.073	3 min

IV. SONUÇ

Yapılan bu çalışmada birden fazla sınıflandırma yöntemini, toplu öğrenme yöntemleriyle bir araya getirerek, sınıflandırma problemine daha kesin ve doğruluk açısından yüksek sonuç bulma oranı hedefleyen çözümler getirilmeye çalışılmıştır.

Sınıflandırma gücüne direkt etkisi olduğu bilinen özellik seçimi algoritmalarıyla, elimizdeki ulusal banka verisetinin özellik sayısı daraltılmış ve karar mekanizmasına en çok etki eden özelliklerden meydana gelen bir altset elde edilmiştir. Ayrıca hesaplama için gereken zamanı da önemli ölçüde düşürme hedefi gerçekleşmiştir.Toplu öğrenme yöntemlerinin tekil sınıflandırma yöntemleriyle birlikte kullanıldığında; sınıflandırma doğruluk oranını pozitif etkilediği yapılan deneyler sonucu ortaya çıkmıştır. Kendi içlerinde karşılaştırıldığında en iyi sonuçların; destek-vektör makinaları ve çantalama yöntemlerinin birlikte kullanıldığında elde edildiği gözlenmiştir.

Deneylerde karşılaştırılan bir diğer ölçüt ise hesaplama kompleksitesidir. Naive bayes yönteminin çok hızlı sonuçlar ürettiği görülmüş, karar-ağaçlarının da yine daha iyi sonuçlar verdiği gözlenmiştir. Yapay sinir ağları ise içerisinde bulunan geri-yayılım (back-propogation) modellemesi gereği en fazla zamana ihtiyaç duyan sınıflandırma modeli olarak kaydedilmiştir.

İlgili testler Alman kredi veriseti üzerinde de tekrar edilerek; hibrid modellerin tekil sınıflandırma algoritmalarına nazaran daha iyi sonuçlar verdiği tekrar onaylanmış ve ilgili modellerin kredi risk analizi alanında genellenebilirliği konusunda güvenilir çıkarımlar yapılmıştır.

Bu çalışmanın sonuçları, önümüzde hedeflediğimiz yeni çalışmada optimum bilgi füzyon yaklaşımı ve destek/seçim kombinasyon kural yöntemleriyle birlikte; çoklu sınıflandırma yöntemi oluşturma amacıyla kullanılacaktır.

KAYNAKLAR

[1] Credit Risk Analysis:A Tryst with Strategic Prudence" , C. Joseph, Chapter- 1 ,ISBN: 0070581363 , 2006.

[2] Credit Risk Analysis:A Tryst with Strategic Prudence" , C. Joseph, Chapter- 2 ,ISBN: 0070581363 , 2006.

[3] Gallo, C., Letizia, C., Stasio, G. "Artificial Neural Networks in Financial Modelling", Research Gate, 1-21 , 2006.

[4] Shachmurave , Y. "Applying Artificial Neural Networks to Business Economics & Finance " , 2002.

[5] Sotchidananda S. Sogala, Ph.D. , "Comparing the Efficacy of the Decision Trees with Logistic Regression for Credit Risk Analysis", Head Risk Solutions & Research, HP India .

[6] Hong Yu, Xiadei Huang, Xiarong Hu, Hongwon Cai."A Comparative Study on Data Mining Algorithms for Individual Credit Risk Evaluation", Department of Information Science, Nanchang Teachers College , China, 2010.

[7] Yangqioo Wang, Shauyang Wang, K. K. Lei . "A New Fuzzy SVM to Evaluate Credit Risk" , 6 December 2005.

[8] Michael Doumpos, Constantin Zopounidis. "Model combination for credit risk assessment:A stacked generalization approach", University of Crete, Greece , 2006.

[9] S. Kotsiantis. "Credit Risk Analysis Using a Hybrid Data Mining Model", University of Peloponnose , Tripolis , Greece.

[10] Koutanaie, F. ''A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring', Journal of Retailing and Consumer Services 27, pg:11-23, 2015

[11] German Dataset,Statlog Project Databases, ftp://ftp.ics.uci.edu/pub/machine-learning databases/statlog/german

[12] Turkish Bank Dataset, Anonymized.