Available online at www.alphanumericjournal.com

alphanumeric journal

The Journal of Operations Research, Statistics, Econometrics and Management Information Systems

Volume 6, Issue 2, 2018

Received: August 09, 2017
Accepted: October 09, 2018
Published Online: December 30, 2018

AJ ID: 2018.06.02.MIS.03
DOI: 10.17093/alphanumeric.333785
Research Article

A Clustering Based Classifier Ensemble Approach to Corporate Bankruptcy Prediction

Aytuğ Onan, Ph.D.

Assoc. Prof., Department of Software Engineering, Faculty of Technology, Celal Bayar University, Manisa, Turkey, aytug.onan@cbu.edu.tr

* Manisa Celal Bayar Üniversitesi, Hasan Ferdi Turgutlu Teknoloji Fakültesi, Yazılım Mühendisliği Bölümü, 45400, Turgutlu, Manisa, Türkiye

ABSTRACT

Corporate bankruptcy prediction is an important research direction in finance. Building a robust prediction scheme for bankruptcy can be beneficial to several stakeholders, including management organizations, government and stockholders. Ensemble learning is a well-known technique to improve the predictive performance of classification algorithms by decreasing the generalization error and enhancing the classification accuracy. It has been a well-established technique in bankruptcy prediction to enhance the predictive performance. Diversity plays an essential role in constructing robust ensemble classification schemes. In this paper, a clustering based classifier ensemble approach is presented for corporate bankruptcy prediction. In this scheme, k-means algorithm is utilized to obtain diversified training subsets. Based on the subsets, each base learning algorithms are trained and the predictions of base learning algorithms are combined by a majority voting scheme. In the empirical analysis, four classification algorithms (namely, C4.5 algorithm, k-nearest neighbour algorithm, support vector machines and logistic regression) and three ensemble learning methods (Bagging, AdaBoost and Random Subspace) are evaluated.

Keywords: Corporate Bankruptcy Prediction, Ensemble Learning, Clustering, Diversity

Firma Başarısızlığının Tahmin Edilmesi İçin Kümelemeye Dayalı Bir Sınıflandırıcı Topluluğu Yaklaşımı

ÖZ

Firma başarısızlıklarının tahmin edilmesi, finansta önemli bir araştırma yönüdür. Güvenilir başarısızlık tahmin etme modellerinin geliştirilmesi, aralarında yönetim organizasyonlarının, devlet kurumlarının ve hisse senedi sahiplerinin de yer aldığı birçok farklı paydaş için oldukça yararlı olabilmektedir. Topluluk öğrenmesi yöntemi, genelleştirme hatasını azaltarak ve doğru sınıflandırma oranını artırarak, sınıflandırma algoritmalarının tahmin etme başarımını artıran önemli bir tekniktir. Topluluk öğrenmesi, firma başarısızlıklarının tahmin edilmesinde kullanılan yaygın kullanıma sahip bir yöntemdir. Yüksek başarımlı sınıflandırıcı topluluklarının oluşturulmasında çeşitlilik önemli bir rol oynamaktadır. Bu çalışmada, firma başarısızlıkların tahmin edilmesi için kümelemeye dayalı bir sınıflandırıcı topluluğu yaklaşımı sunulmaktadır. Önerilen tasarıda, k-ortalama algoritması kullanılarak, çeşitlendirilmiş eğitim alt kümeleri oluşturulmaktadır. Bu eğitim alt kümelerine dayalı olarak, sınıflandırıcı topluluğunda yer alan her bir temel öğrenme algoritması eğitilmekte ve temel öğrenme yöntemlerinin bireysel çıktıları çoğunluk oylaması aracılığıyla birleştirilmektedir. Deneysel analizlerde, dört sınıflandırma algoritması (C4.5 algoritması, k-en yakın komşu algoritması, destek vektör makineleri ve lojistik regresyon) ve üç topluluk öğrenmesi yöntemi (Bagging, AdaBoost ve rastgele alt uzay) değerlendirilmiştir.

Anahtar Kelimeler: Firma Başarısızlığının Tahmin Edilmesi, Topluluk Öğrenmesi, Kümeleme, Çeşitlilik

1. Giriş

Firma başarısızlıkların tahmin edilmesi, finans alanındaki önemli araştırma problemlerinden biridir. Firma başarısızlığı temel olarak, belirli bir firmanın borçlarını zamanında ödeyememesi, kar payında meydana gelen düşüş, acze düşme, iflas ve tasfiye gibi farklı durumları içerisine alan bir kavramdır (Blum, 1974; Lau, 1987). Firmaların finansal başarısızlıkları temel olarak ekonomik faktörler, yönetimsel başarısızlıklar, ödemede acze düşme ve iflas olmak üzere dört temel başlık altında incelenebilir (Brigham & Ehrhadt, 2013). Firma başarısızlıkları, hem ekonomik hem de sosyal etkileri olan bir problemdir. Firma başarısızlıkları, yalnızca ilgili firmayı ve firmanın iş yaptığı kurumlarla sınırlı olmayıp daha geniş bir alana ve ülke ekonomisine etkide bulunabilen önemli bir sorundur (Andreev, 2006). Firma başarısızlıklarının tahmin edilmesi, yönetim organizasyonları, devlet kurumları, hisse senedi sahipleri, kredi sağlayan kurumlar gibi birçok farklı paydaşı yakından ilgilendiren oldukça önemli bir finansal konudur (Barboza, Kimura & Altman, 2017).

Güvenilir başarısızlık tahmin etme yöntemlerinin geliştirilmesi, aralarında yönetim organizasyonlarının, devlet kurumlarının ve hisse senedi sahiplerinin de yer aldığı paydaşlar için oldukça yararlı olabilmektedir. Firma başarısızlıklarının tahmin edilmesi, finansal kurumlar ve firmalar için önemli bir karar desteği olarak işlev görmektedir. Firma başarısızlıklarının tahmin edilmesi için kullanılan istatistiksel ve makine öğrenmesine dayalı tekniklerin, uzman görüşüne dayalı tahminlere kıyasla daha güvenilir sonuçlar verdiği görülmektedir (Balcaen & Ooghe, 2006). Firma başarısızlıklarının tahmin edilmesinde kullanılan istatistiksel yöntemler arasında, regresyon analizi, diskriminant analizi ve lojistik regresyon gibi teknikler bulunmaktadır (Altman, 1968; Altman, Edward, Haldeman, & Narayanan,1977; Pantalone & Platt, 1987). İstatistiksel yöntemler, tahmin edici değişkenler arasında, doğrusal olma, normallik ve bağımsızlık gibi varsayımlara sahip olduğundan, firma başarısızlıklarına ilişkin gerçek dünya problemlerinde uygulanması kısmen kısıtlıdır (Kim & Kang, 2012). Firma başarısızlıklarının tahmin edilmesinde, makine öğrenmesine dayalı yöntemler başarıyla uygulanmaktadır (Barboza, Kimura & Altman, 2017). Makine öğrenmesine dayalı yöntemler arasında, karar ağaçları, yapay sinir ağları ve destek vektör makineleri gibi teknikler yer almaktadır (Kim & Kang, 2012).

Topluluk öğrenmesi yöntemi, tahmin etme modeli çıktısının tek bir temel öğrenme algoritması yerine, birden fazla öğrenme algoritması sonucu elde edilen çıktının birleştirilmesi ile elde edilmesini amaçlayan güncel bir makine öğrenmesi araştırma alanıdır (Kuncheva, 2004). Topluluk öğrenmesi yöntemleri, temel öğrenme algoritmalarının genelleştirme hatasını azaltarak ve doğru sınıflandırma oranını artırarak daha yüksek başarımlı tahmin etme modelleri oluşturulmasını amaçlar. Makine öğrenmesine dayalı firma başarısızlıklarının tahmini alanında yapılan çalışmalarda topluluk öğrenmesi yöntemleri kullanılarak daha yüksek tahmin etme başarımına sahip modellerin oluşturulması amaçlanmaktadır. Firma başarısızlıklarının tahmin edilmesinde kullanılan topluluk öğrenmesi yöntemleri arasında Bagging, Boosting, AdaBoost ve rastgele alt uzay gibi teknikler kullanılmaktadır (Tsai, Hsu & Yen, 2014; Nanni & Lumuni, 2009). Topluluk öğrenmesi yöntemlerinin genel olarak, temel öğrenme algoritmalarına kıyasla daha yüksek başarım elde etmesi beklenmektedir (Dietterich, 2000). Topluluk öğrenmesi, aralarında metin madenciliği, insan aktivite tanımlama ve biyoenformatik uygulamalarının da yer aldığı birçok farklı

alanda başarıyla uygulanmaktadır (Onan, 2016; Catal, Tufekci & Kocabag, 2015; Yang, Hwa, Zhou & Zomaya, 2010). Temel olarak, yüksek başarımlı sınıflandırıcı topluluklarının oluşturulması için gerekli iki önemli nokta vardır. Hem sınıflandırıcı topluluğunda yer alan temel öğrenme algoritmalarının olabildiğince yüksek tahmin etme başarımına sahip olması hem de temel öğrenme algoritmaları arasında çeşitlilik olması gerekmektedir (Zhou, 2012). Sınıflandırıcı topluluğunda yer alan temel öğrenme algoritmaları arasında çeşitlilik, veri seviyesi ya da model oluşturma seviyesinde gerçekleştirilebilmektedir (Onan, 2017; Mendes-Moreira, Soares, Jorge & De Sousa, 2012).

Bu çalışmada, firma başarısızlıklarının tahmin edilmesi için sınıflandırıcı topluluğuna dayalı bir yöntem önerinde bulunulmuştur. Geliştirilen yöntemde, k-ortalama algoritması kullanılarak, çeşitlendirilmiş eğitim alt kümeleri oluşturulmaktadır. Bu eğitim alt kümelerine dayalı olarak, sınıflandırıcı topluluğunda yer alan her bir temel öğrenme algoritması eğitilmekte ve temel öğrenme yöntemlerinin bireysel çıktıları çoğunluk oylaması aracılığıyla birleştirilmektedir. Deneysel analizlerde, dört sınıflandırma algoritması (C4.5 algoritması, k-en yakın komşu algoritması, destek vektör makineleri ve lojistik regresyon) ve üç topluluk öğrenmesi yöntemi (Bagging, AdaBoost ve rastgele alt uzay) değerlendirilmiştir.

Çalışmanın geri kalan kısmı şu şekilde yapılandırılmıştır: İkinci bölümde, literatür özeti, üçüncü bölümde çalışmada kullanılan temel yöntemler, dördüncü bölümde geliştirilen sınıflandırıcı topluluğu yöntemi, beşinci bölümde deneysel analiz ve sonuçlar ve altıncı bölümde çalışmanın temel sonuçları yer almaktadır.

2. Literatür Özeti

Literatürde firma başarısızlıklarının tahmin edilmesi için geliştirilmiş temel makine öğrenmesi sınıflandırıcılarına dayalı ve sınıflandırıcı topluluklarına dayalı birçok çalışma bulunmaktadır. Bu çalışmalardan bazıları bu bölümde özetlenmiştir. Örneğin, Olson vd. (2012) tarafından gerçekleştirilen çalışmada, yapay sinir ağları, karar ağaçları ve destek vektör makineleri sınıflandırıcılarının şirket iflaslarının tahmin edilmesindeki başarımları değerlendirilmiştir. Onan (2015) tarafından gerçekleştirilen çalışmada, şirket iflaslarının tahmin edilmesinde yedi farklı karar ağacı algoritmasının (C4.5 algoritması, decision stump algoritması, hoeffding tree algoritması, lojistik model ağacı algoritması, rastgele orman algoritması, rastgele ağaç algoritması ve RepTree algoritması) başarımları karşılaştırmalı olarak incelenmiştir.

Alfaro vd. (2008) çalışmalarında, firma başarısızlıklarının tahmin edilmesi için AdaBoost ve yapay sinir ağlarının başarımlarını incelemiş ve AdaBoost algoritmasının yapay sinir ağlarına kıyasla daha yüksek başarım elde ettiği sonucuna varmıştır. Hsieh ve Hung (2010) tarafından gerçekleştirilen çalışmada, kredi risk değerlendirme için veri güdümlü bir sınıflandırıcı topluluğu mimarisi sunulmuştur. Kim ve Kang (2012) çalışmalarında, genetik algoritma ve sınıflandırıcı topluluğuna dayalı bir yöntem önerisinde bulunmuş ve Kore'deki firmaların iflaslarının tahmin edilmesinde yöntemin başarımını, standart topluluk öğrenmesi yaklaşımları ile karşılaştırmıştır. Geliştirilen yöntemde, birbirleriyle yüksek korelasyona sahip temel öğrenme algoritmalarının, sınıflandırıcı topluluğunun başarımını düşürmesini engellemek amacıyla, genetik algoritmaya dayalı bir eniyileme aşaması uygulanmıştır. Marques vd. (2012) çalışmalarında, kredi derecelendirme için, yedi temel öğrenme algoritması (Naive

Bayes, k-en yakın komşu algoritması, çok katmanlı algılayıcı ağ, radyal tabanlı fonksiyon ağları, lojistik regresyon, destek vektör makineleri ve C4.5 algoritması) ile beş temel topluluk öğrenmesi yönteminin (Bagging, AdaBoost, rastgele alt uzay, DECORATE ve rotasyon ormanı yöntemleri) etkinliklerini değerlendirmiştir. Benzer şekilde, Tsai vd. (2014) tarafından gerçekleştirilen çalışmada, üç temel sınıflandırma algoritması (yapay sinir ağları, destek vektör makineleri ve karar ağaçları) ile iki temel topluluk öğrenmesi yönteminin (Bagging ve Boosting algoritması) etkinlikleri, firma başarısızlıklarının tahmin edilmesinde değerlendirilmiştir. Deneysel analizlerde, karar ağacı algoritmasının Boosting topluluk öğrenmesi algoritması ile birleştirilmesi sonucu en yüksek tahmin etme başarımına sahip modelin oluşturulduğu gözlenmiştir. Wang vd. (2014) tarafından gerçekleştirilen çalışmada ise firma başarısızlıklarının tahmin edilmesi için öznitelik seçimine dayalı bir sınıflandırıcı topluluğu mimarisi önerisinde bulunulmuştur. Benzer şekilde, Koutanaei vd. (2015) çalışmalarında, kredi değerlendirme için, farklı öznitelik setleri, temel öğrenme algoritmaları ve topluluk öğrenmesi yöntemlerinin başarımlarını değerlendirmiştir. Kim vd. (2016) tarafından gerçekleştirilen çalışmada, firma başarısızlıklarının tahmin edilmesi için kümelemeye ve yapay sinir ağlarına dayalı bir model önerisi sunulmuştur. Önerilen mimaride, kümeleme yöntemi, veri dengesizliğini ortadan kaldırmak amacıyla kullanılmıştır. Çoğunluk sınıfı üzerinde, kümeleme analizi uygulanarak, uygun veri nesnelerinin seçilmesi, veri dengesizliğinin kaldırılması ile doğru sınıflandırma başarımının artırılması amaçlanmıştır. Xia vd. (2016) tarafından gerçekleştirilen çalışmada ise kredi değerlendirme için, öğreticili kümeleme analizine dayalı bir topluluk öğrenmesi yaklaşımı geliştirilmiştir. Diğer bir çalışmada, Chou vd. (2017) tarafından bulanık c-ortalama kümeleme ve genetik algoritmaya dayalı melez bir firma başarısızlığı tahmin etme modeli geliştirilmiştir.

3. Metodoloji

Bu bölümde, geliştirilen yöntemde ve deneysel analizlerde kullanılan sınıflandırma algoritmaları, topluluk öğrenmesi yöntemleri ve kümeleme yöntemlerine ilişkin temel bilgilere yer verilmektedir.

3.1. Sınıflandırma Algoritmaları

Çalışma kapsamında dört temel sınıflandırma algoritması olan C4.5 algoritması, k-en yakın komşu algoritması, destek vektör makineleri ve lojistik regresyon yöntemleri kullanılmıştır.

C4.5 algoritması, sürekli ve kesikli değerler içeren veri setleri üzerinde çalışabilen temel karar ağacı yöntemlerinden biridir. Karar ağacına dayalı yöntemler, sınıflandırma ve tahmin etme problemlerinde sıklıkla uygulanmaktadır (Onan, 2015). C4.5 algoritmasında, bilgi kazancı ölçütü, öznitelik kullanışlılığının değerlendirilmesinde kullanılmaktadır. Hesaplanan bilgi kazancı değerine dayalı olarak, her bir öznitelik seti içerisi içerisinden en yüksek değere sahip küme seçilmektedir. C4.5 algoritmasında, karar ağacı oluşturulması ya da sonrasında bazı alt düğümlerin budanabilmesi ile aşırı uygunluk problemi ortadan kaldırılmaktadır.

K-en yakın komşu algoritması (KNN), örnek tabanlı bir sınıflandırma algoritmasıdır. KNN algoritmasında, sınıflandırılacak olan veri örneği, eğitim setinde yer alan mevcut veri örnekleri ile arasındaki benzerliğe göre sınıflandırma işlemine tabi tutulur. Burada,

eğitim setinde yer alan her bir örnek n boyutlu sayısal nitelikler ile belirtilir. Yeni gelen bir örnek, eğitim setinde yer alan ilgili örneğe en yakın k tane örneğin sınıf etiketlerinin çoğunluk oylamasına göre uygun sınıfa atanır (Taşçı & Onan, 2016; Han, Pei & Camber, 2011).

Destek vektör makineleri (SVM), hem doğrusal hem de doğrusal olmayan verilerin sınıflandırılmasında kullanılan temel sınıflandırma algoritmalarından biridir. Burada, doğrusal olmayan bir eşleme kullanılarak, başlangıçtaki veri daha üst bir boyuta dönüştürülerek, bu yeni boyutta, veriyi en uygun biçimde ayırabilecek bir üst düzlem bulunması amaçlanır. Destek vektör makineleri, genelleştirme yeteneklerinin yüksek olması, gürültülü ve aykırı değerler içeren verilere karşı dayanıklı olması ve yüksek başarımlarından ötürü, sınıflandırma problemlerinde başarıyla uygulanmaktadır (Han, Pei & Camber, 2011).

Lojistik regresyon algoritması (LR), istatistiksel bir sınıflandırma yöntemidir. Lojistik regresyon yönteminde, eğitim setinde yer alan örneklere dayalı bir sınıflandırma modeli oluşturulur. Yeni karşılaşılan sınıflandırılacak örnekler, hesaplanan en yüksek olasılık değerine dayalı olarak sınıflara atanır. Algoritmada, olasılık değerlerinin hesaplanması parametreler üzerinden gerçekleştirilir (Shatkay & Craven, 2012).

3.2. Topluluk Öğrenmesi Yöntemleri

Çalışma kapsamında üç temel topluluk öğrenmesi yöntemi olan AdaBoost algoritması, Bagging algoritması ve rastgele alt-uzay algoritması kullanılmıştır.

AdaBoost algoritması, en temel Boosting yöntemlerinden biridir. AdaBoost algoritmasında, sınıflandırılması zor olan örneklere daha fazla odaklanılarak, sınıflandırma modelinin başarımının artırılması amaçlanır (Freund & Schapire, 1996). Yöntemin her bir yinelemesinde, doğru sınıflandırılmayan örneklere ilişkin ağırlık değerleri artırılırken, doğru sınıflandırılan örneklere ilişkin ağırlık değerleri azaltılarak, temel öğrenme algoritmalarının eğitim setinde yer alan ve sınıflandırılması zor olan veri örneklerine daha fazla yineleme ayırması sağlanır. Buna ek olarak, AdaBoost algoritmasında sınıflandırma algoritmalarına da ağırlık değerleri atanarak, doğru sınıflandırma başarımı daha yüksek olan sınıflandırıcılar daha yüksek ağırlık değerleri ile temsil edilir (Rokach, 2010; Onan, 2016).

Bagging algoritması, eğitim setinin farklı örneklemleri üzerinde eğitilmiş temel öğrenme algoritmalarının birleştirilmesi ile sınıflandırıcı topluluğu oluşturulmasına yönelik bir topluluk öğrenmesi yöntemidir (Breiman, 1996). Bagging algoritmasının temel noktası, topluluğu oluşturan her bir temel öğrenme algoritmasının, farklı eğitim setleri üzerinde eğitilmesi ile çeşitliliğin sağlanması ilkesine dayanmaktadır. Burada, veri setinden farklı eğitim setlerinin oluşturulması amacıyla genellikle basit rastgele yerine koyarak örnekleme yöntemi uygulanmaktadır. Örneklem yöntemi ile elde edilen eğitim setleri ile eğitilen sınıflandırma yöntemlerinin çıktıları, çoğunluk oylaması aracılığıyla birleştirilmektedir.

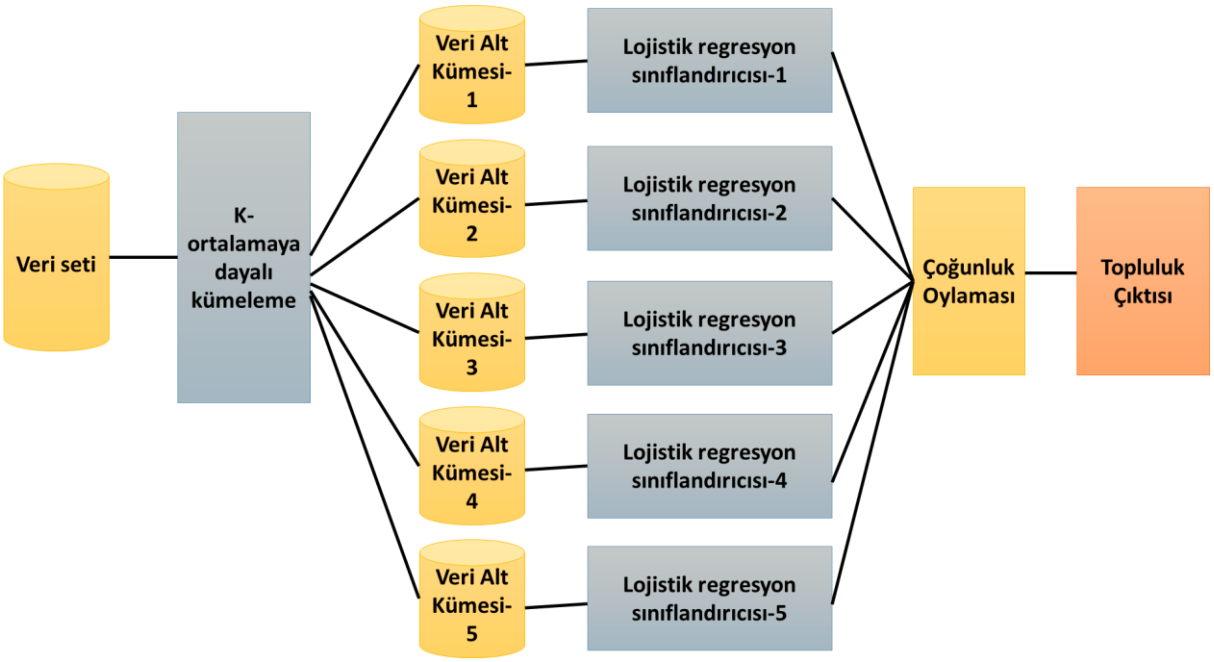
Rastgele alt-uzay yöntemi, öznitelik uzayı üzerinde değişiklik yapılarak, farklı eğitim setlerin elde edilir. Rastgele alt-uzay yöntemi ile mevcut eğitim seti küçük eğitim setlerine parçalanarak, her bir temel öğrenme algoritması daha küçük eğitim setleri üzerinde eğitilir (Ho, 1998).

3.3. K-ortalama Kümeleme Algoritması

K-ortalama algoritması, en temel kümeleme algoritmalarından biridir. K-ortalama algoritması, küme sayısını (k) girdi parametresi olarak alır. K-ortalama algoritması, rastgele olarak k tane veri nesnesinin seçilmesi ile başlar. Geriye kalan her bir nesne, kümelerdeki nesnelerin ortalama değerlerine göre kendilerine en yakın kümelere atanır. Ardından, her bir küme için nesnelerin ortalama değeri hesaplanarak küme ortalamaları güncellenir. Süreç, küme merkezleri değiştiği sürece sürdürülür (Onan, Bulut, Korukoğlu, 2017).

4. Geliştirilen Yöntem

Geliştirilen kümelemeye dayalı sınıflandırıcı topluluğu yaklaşımına (KM-LR) ilişkin genel mimari Şekil 1'de sunulmuştur. Geliştirilen sınıflandırıcı topluluğu yönteminde, öncelikle k-ortalama kümeleme algoritması kullanılarak, eğitim setinde yer alan veri örnekleri altkümelere parçalanmaktadır. Böylelikle, k-ortalama algoritması kullanılarak, çeşitlendirilmiş eğitim alt kümeleri oluşturulmaktadır. Çeşitlendirilmiş eğitim altkümeleri, k tane lojistik regresyon temel öğrenme algoritmasının eğitilmesi için kullanılmaktadır. Bu biçimde sınıflandırıcı topluluğunda yer alan her bir temel öğrenme algoritması (lojistik regresyon sınıflandırıcısı) eğitilmekte ve temel öğrenme algoritmalarının çıktıları çoğunluk oylaması kullanılarak birleştirilmektedir.



Şekil 1. Kümelemeye dayalı sınıflandırıcı topluluğu mimarisi.

Geliştirilen kümelemeye dayalı sınıflandırıcı topluluğu mimarisi üç temel aşamadan oluşmaktadır:

1. Aşama K-ortalamaya dayalı kümeleme aşamasıdır. Öncelikle veri setinde yer alan veri nesneleri k-ortalama algoritması (k=5) için işletilerek, beş tane farklı eğitim veri alt kümesi elde edilir. Sınıflandırıcı topluluğunun sonucu çoğunluk oylamasına dayalı olarak belirleneceğinden, toplulukta yer alacak öğrenme algoritması sayısının tek sayılı olarak alınması gerekmektedir. Farklı algoritma sayıları (k=3, k=5, k=7 ve k=9) için yapılan

deneysel analizlerde en yüksek başarım, k=5 için elde edildiğinden, geliştirilen mimaride beş veri altkümesi ve temel öğrenme algoritması kullanılmıştır.

2. Aşama Temel öğrenme algoritmaları (lojistik regresyon sınıflandırıcısı), çeşitlendirilmiş eğitim alt kümeleri kullanılarak eğitilir. Farklı sınıflandırma algoritmaları (C4.5, destek vektör makineleri, k-en yakın komşu algoritması ve lojistik regresyon sınıflandırıcısı) ile yapılan deneysel analizlerde, en yüksek başarım, lojistik regresyon algoritması ile elde edilmiştir. Bu nedenle, geliştirilen mimaride lojistik regresyon sınıflandırıcısı kullanılmaktadır.

3. Aşama Temel öğrenme algoritmalarının çıktıları, çoğunluk oylamasına dayalı olarak birleştirilerek sınıflandırıcı topluluğunun temel çıktısı elde edilmektedir.

5. Deneysel Analiz ve Sonuçlar

Bu bölümde, deneysel çalışmalarda kullanılan veri seti, deneysel analizlerde izlenen süreç ve deneysel sonuçlar sunulmaktadır.

5.1. Veri Seti

Çalışmada kullanılan veri seti, Polonya'daki üretim sektöründeki firmalara ilişkin finansal durumlar taranarak oluşturulmuştur (Zieba, Tomczak & Tomczak, 2016). Veri seti, toplam 10503 veri örneği içermektedir. Veri setinde firmaların finansal durumlarını temsil etmek üzere toplam 64 finansal ölçüt öznitelik olarak kullanılmıştır. Öznitelikler arasında, net karın toplam aktiflere oranı, dönen varlıkların kısa vadeli borçlara oranı, net karın toplam satışlara oranı gibi ölçütler yer almaktadır. Firmalara ilişkin veriler, tahmin etme periyoduna göre beş temel sınıfa bölünerek, beş farklı veri seti oluşturulmuştur (Zieba, Tomczak & Tomczak, 2016):

- Birinci yıl veri seti: Bir yıllık tahmin etme periyoduna ilişkin bilgileri içeren veri setidir. Veri setini oluşturan firmaların tahmin etme periyodunun birinci yılına ilişkin finansal veriler dikkate alınarak oluşturulmuştur. Sınıf etiketi, beş yıllık süreçte firmaların iflas edip etmediğine ilişkin bilgiyi içermektedir. Veri seti toplamda, 7027 örnek (firma) bilgisi içermektedir. Bu örneklerden 271 tanesi iflas ederken, geriye kalan firmalar ilgili tahmin etme periyodunda iflas etmemiştir.

- İkinci yıl veri seti: Firmaların tahmin etme periyodunun ikinci yılına ilişkin finansal veriler dikkate alınarak oluşturulmuştur. Sınıf etiketi, dört yıllık süreçte firmaların iflas edip etmediğine ilişkin bilgiyi içermektedir. Veri seti toplamda, 10173 örnek (firma) bilgisi içermektedir. Bu örneklerden, 400 tanesi iflas ederken, geriye kalan 9773 tanesi ilgili tahmin etme periyodunda iflas etmemiştir.

- Üçüncü yıl veri seti: Firmaların tahmin etme periyodunun üçüncü yılına ilişkin finansal veriler dikkate alınarak oluşturulmuştur. Sınıf etiketi, üç yıllık süreçte firmaların iflas edip etmediğine ilişkin bilgiyi içermektedir. Veri seti toplamda, 10503 örnek (firma) bilgisi içermektedir. Bu örneklerden, 495 tanesi iflas ederken, geriye kalan 10008 tanesi ilgili tahmin etme periyodunda iflas etmemiştir.

- Dördüncü yıl veri seti: Firmaların tahmin etme periyodunun dördüncü yılına ilişkin finansal veriler dikkate alınarak oluşturulmuştur. Sınıf etiketi, iki yıllık süreçte firmaların iflas edip etmediğine ilişkin bilgiyi içermektedir. Veri seti toplamda, 9792 örnek (firma) bilgisi içermektedir. Bu örneklerden, 515 tanesi iflas ederken, geriye kalan 9277 tanesi ilgili tahmin etme periyodunda iflas etmemiştir.
- Beşinci yıl veri seti: Firmaların tahmin etme periyodunun beşinci yılına ilişkin finansal veriler dikkate alınarak oluşturulmuştur. Sınıf etiketi, bir yıllık süreçte firmaların iflas edip etmediğine ilişkin bilgiyi içermektedir. Veri seti toplamda, 5910 örnek (firma) bilgisi içermektedir. Bu örneklerden, 410 tanesi iflas ederken, geriye kalan 5500 tanesi ilgili tahmin etme periyodunda iflas etmemiştir.

5.2. Deneysel Süreç

Deneysel analizlerde kullanılan sınıflandırma algoritmaları ve topluluk öğrenmesi yöntemleri, WEKA 3.9 kullanılarak gerçekleştirilmiştir. Temel öğrenme algoritmalarının ve topluluk öğrenmesi yöntemleri için WEKA yazılımında bulunan temel varsayılan parametre değerleri kullanılmıştır. Deneysel çalışmalarda, 10-kat çapraz geçerleme yöntemi kullanılarak, veri setleri 10 eşit parçaya ayrılmış ve her bir yinelemede parçalarda biri sınama amacıyla diğer parçalar eğitim amacıyla kullanılmıştır.

5.3. Değerlendirme Ölçütleri

Çalışma kapsamında, geliştirilen topluluk öğrenmesi yönteminin değerlendirilmesi için doğru sınıflandırma oranı kullanılmıştır.

Doğru sınıflandırma oranı, sınıflandırma algoritmalarının başarımını değerlendirmek için kullanılan en temel ölçütlerden biridir. Doğru sınıflandırma oranı (ACC), doğru pozitifler ve doğru negatifler toplamının, doğru pozitif, yanlış pozitif, yanlış negatif ve doğru negatiflerin toplamına oranlanması ile Eşitlik 1'e göre hesaplanır:

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

Burada, TN, TP, FP ve FN sırası ile doğru negatif, doğru pozitif, yanlış pozitif ve yanlış negatif sayılarını temsil etmektedir.

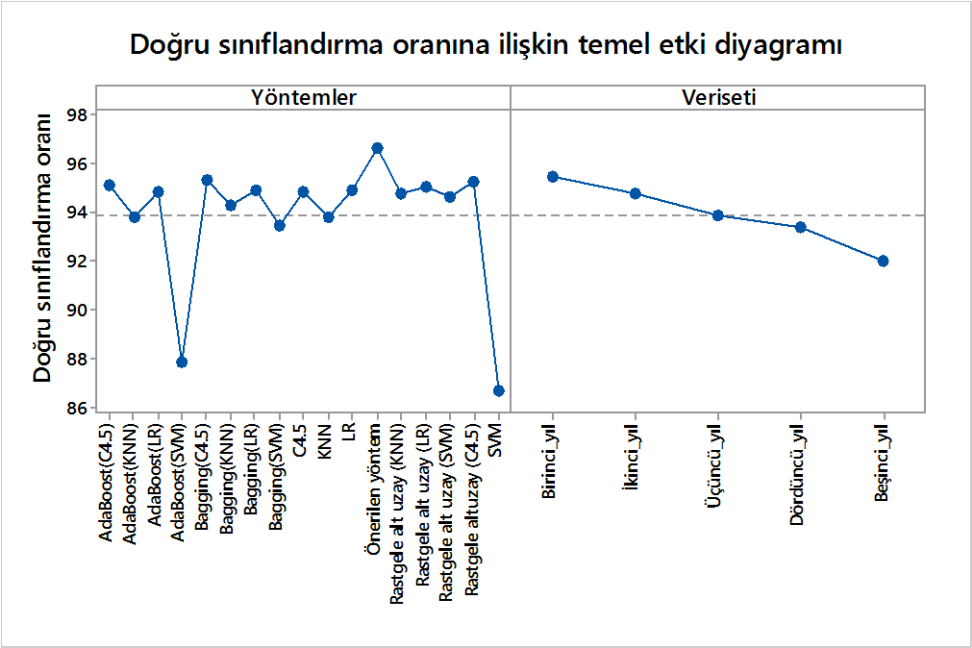
5.4. Deneysel Sonuçlar

Çalışma kapsamında, geliştirilen kümelemeye dayalı sınıflandırıcı topluluğu mimarisinin doğru sınıflandırma tahmin etme başarımı, dört temel makine öğrenmesi sınıflandırıcısı (C4.5 algoritması, k-en yakın komşu algoritması, destek vektör makineleri ve lojistik regresyon) ve üç topluluk öğrenmesi yöntemi (Bagging, AdaBoost ve rastgele alt-uzay) ile karşılaştırılmıştır.

Yöntemler	Birinci Yıl	İkinci Yıl	Üçüncü Yıl	Dördüncü Yıl	Beşinci Yıl
C4.5	95.76	95.95	95.01	94.41	93.29
AdaBoost(C4.5)	96.3	*96.13*	95.18	94.55	93.37
Bagging(C4.5)	*96.23*	96.21	95.43	94.96	*93.68*
Rastgele altuzay (C4.5)	96.2	96.13	*95.33*	*94.82*	93.87
SVM	90.28	86.14	85.35	85.83	85.88
AdaBoost(SVM)	92.69	86.89	91.43	85.09	83.24
Bagging(SVM)	94.66	94.43	93.95	93.09	91.29
Rastgele alt uzay (SVM)	95.97	95.83	94.86	94.32	92.22
KNN	95.67	95.26	92.05	93.65	92.3
AdaBoost(KNN)	95.67	95.26	92.05	93.65	92.3
Bagging(KNN)	95.96	95.75	92.76	94.25	92.7
Rastgele alt uzay (KNN)	95.9	95.79	94.87	94.41	92.8
LR	96.02	95.94	95.11	94.59	92.8
AdaBoost(LR)	96	95.82	95.07	94.57	92.73
Bagging(LR)	96.06	95.98	95.18	94.62	92.88
Rastgele alt uzay (LR)	96.14	96.02	95.23	94.71	93.02
Önerilen yöntem	97.43	97.24	96.74	96.11	95.79

Tablo 1. Makine öğrenmesi yöntemlerine ilişkin doğru sınıflandırma oranları.

Tablo 1.'de karşılaştırılan yöntemlere ilişkin doğru sınıflandırma oranları özetlenmektedir. Tablo 1.'de sunulan deneysel analiz sonuçları incelendiğinde, çalışmada kullanılan temel öğrenme algoritmaları içerisinde en yüksek başarımın lojistik regresyon sınıflandırıcısı ile en düşük doğru sınıflandırma oranının ise destek vektör makineleri ile elde edildiği gözlenmektedir. Temel öğrenme algoritmalarının, üç farklı sınıflandırıcı topluluğu yöntemi ile bir araya getirilmesi ile elde edilen farklı toplulukların, temel öğrenme algoritmalarının doğru sınıflandırma oranlarını genellikle iyileştirdiği görülmektedir. Deneysel analizlerde incelenen tüm makine öğrenmesi yöntemleri içerisinde, her bir veri seti için, en yüksek başarımın genellikle C4.5 algoritmasının, sınıflandırıcı topluluğu yöntemleri ile bir arada kullanılması ile elde edildiği görülmektedir. Çalışma kapsamında, önerilen kümelemeye dayalı sınıflandırıcı topluluğu mimarisi, deneysel analizlerde kullanılan firma başarısızlıklarının tahmin edilmesine ilişkin veri setleri için, dört temel öğrenme algoritmasına ve üç temel sınıflandırıcı topluluğunun farklı kombinasyonlarına kıyasla daha iyi sonuçlar vermektedir. Tüm analizler içerisinde en yüksek başarım (%97.43) doğru sınıflandırma oranı ile geliştirilen sınıflandırıcı topluluğu yöntemi ile elde edilmiştir.



Şekil 2. Doğru sınıflandırma oranına ilişkin temel etki diyagramı.

Şekil 2'de makine öğrenmesi yöntemlerinin doğru sınıflandırma oranları ve veri setlerinden elde edilen sonuçlara ilişkin temel etki diyagramı sunulmuştur. Deneysel analizlerde kullanılan veri setleri için, en yüksek doğru sınıflandırma oranın önerilen kümelemeye dayalı sınıflandırıcı topluluğu yöntemi ile elde edildiği görülmektedir. Firmalara ilişkin veriler tahmin etme periyoduna göre beş temel sınıfa bölünerek, beş veri setinin etkinlikleri incelenmiştir. Deneysel analizlere göre, farklı tahmin etme periyotları içerisinde en yüksek başarım, bir yıllık tahmin etme periyoduna ilişkin bilgi içeren veri seti ile elde edilirken, en düşük başarım ise firmaların tahmin etme periyodunun beşinci yılına ilişkin finansal veriler dikkate alındığında elde edilmiştir. Bir yıllık tahmin etme periyodundan, beş yıllık tahmin etme periyoduna kadar, makine öğrenmesi sınıflandırıcılarının doğru tahmin etme performansının giderek düştüğü gözlemlenmiştir. Bu çalışmada, etkin bir sınıflandırıcı topluluğu mimarisi oluşturulması için gerekli çeşitlilik, kümeleye dayalı çeşitlendirilmiş eğitim alt kümeleri oluşturularak gerçekleştirilmiştir. Deneysel analizlerden elde edilen sonuçlar, yöntemin başarımını destekler niteliktedir.

6. Sonuç

Bu çalışma kapsamında, firma başarısızlıkların tahmin edilmesi için, topluluk öğrenmesine dayalı bir yöntem önerisinde bulunulmuştur. Topluluk öğrenmesi yöntemi, firma başarısızlıklarının tahmin edilmesinde, yüksek doğru sınıflandırma başarımından ötürü, sıklıkla kullanılan yöntemler arasındadır. Etkin sınıflandırıcı topluluklarının oluşturulması için toplulukta yer alan temel öğrenme algoritmaları arasında çeşitlilik sağlanması önemlidir. Bu doğrultuda, bu çalışma kapsamında, k-ortalama kümeleme algoritması kullanılarak, çeşitlendirilmiş eğitim alt kümeleri oluşturularak, alt kümeler üzerinde eğitilen lojistik regresyon sınıflandırıcılarının çıktıları, çoğunluk oylaması yöntemi kullanılarak birleştirilmiştir. Geliştirilen yöntemin tahmin etme başarımı, dört sınıflandırma algoritması (C4.5 algoritması, k-en yakın komşu algoritması, destek vektör makineleri ve lojistik regresyon) ve üç topluluk öğrenmesi yöntemi (Bagging, AdaBoost ve rastgele alt uzay) değerlendirilmiştir.

Deneysel analizler, geliştirilen yöntemin, karşılaştırmada kullanılan sınıflandırma algoritmaları ve topluluk öğrenmesi yöntemlerine kıyasla daha yüksek tahmin etme başarımına sahip olduğunu göstermektedir.

Kaynakça

Alfaro, E., García, N., Gámez, M., & Elizondo, D. (2008). Bankruptcy forecasting: An empirical comparison of AdaBoost and neural networks. Decision Support Systems, 45(1), 110-122.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The journal of finance, 23(4), 589-609.

Altman, E. I., Haldeman, R. G., & Narayanan, P. (1977). ZETATM analysis A new model to identify bankruptcy risk of corporations. Journal of banking & finance, 1(1), 29-54.

Andreev, Y.A. (2006). Predicting financial distress of Spanish companies. Jornada De Pre-Comunicaciones A Congresos De Economia Y Administración De Empresas, 1-22.

Balcaen, S., & Ooghe, H. (2006). 35 years of studies on business failure: an overview of the classic statistical methodologies and their related problems. The British Accounting Review, 38(1), 63-93.

Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. Expert Systems with Applications, 83, 405-417.

Blum, M. (1974). Failing company discriminant analysis. Journal of accounting research, 1-25.

Breiman, L. (1996). Bagging predictors. Machine learning, 24(2), 123-140.

Brigham, E. F., & Ehrhardt, M. C. (2013). Financial management: Theory & practice. Cengage Learning.

Catal, C., Tufekci, S., Pirmit, E., & Kocabag, G. (2015). On the use of ensemble of classifiers for accelerometer-based activity recognition. Applied Soft Computing, 37, 1018-1022.

Chou, C. H., Hsieh, S. C., & Qiu, C. J. (2017). Hybrid genetic algorithm and fuzzy clustering for bankruptcy prediction. Applied Soft Computing, 56, 298-316.

Dietterich, T. G. (2000). Ensemble methods in machine learning. Multiple classifier systems, 1857, 1-15.

Freund, Y., & Schapire, R. E. (1996, July). Experiments with a new boosting algorithm. In Icml (Vol. 96, pp. 148-156).

Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.

Ho, T. K. (1998). The random subspace method for constructing decision forests. IEEE transactions on pattern analysis and machine intelligence, 20(8), 832-844.

Hsieh, N. C., & Hung, L. P. (2010). A data driven ensemble classifier for credit scoring analysis. Expert systems with Applications, 37(1), 534-545.

Kim, H. J., Jo, N. O., & Shin, K. S. (2016). Optimization of cluster-based evolutionary undersampling for the artificial neural networks in corporate bankruptcy prediction. Expert Systems with Applications, 59, 226-234.

Kim, M. J., & Kang, D. K. (2012). Classifiers selection in ensembles using genetic algorithms for bankruptcy prediction. Expert Systems with applications, 39(10), 9308-9314.

Koutanaei, F. N., Sajedi, H., & Khanbabaei, M. (2015). A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring. Journal of Retailing and Consumer Services, 27, 11-23.

Kuncheva, L. I. (2004). Combining pattern classifiers: methods and algorithms. John Wiley & Sons.

Lau, A. H. L. (1987). A five-state financial distress prediction model. Journal of accounting research, 127-138.

Marques, A. I., Garcia, V., & Sanchez, J. S. (2012). Exploring the behaviour of base classifiers in credit scoring ensembles. Expert Systems with Applications, 39(11), 10244-10250.

Mendes-Moreira, J., Soares, C., Jorge, A. M., & Sousa, J. F. D. (2012). Ensemble approaches for regression: A survey. ACM Computing Surveys (CSUR), 45(1), 10.

Nanni, L., & Lumini, A. (2009). An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring. Expert systems with applications, 36(2), 3028-3033.

Olson, D. L., Delen, D., & Meng, Y. (2012). Comparative analysis of data mining methods for bankruptcy prediction. Decision Support Systems, 52(2), 464-473.

Onan, A. (2015). Şirket iflaslarının tahmin edilmesinde karar ağacı algoritmalarının karşılaştırmalı başarım analizi. Bilişim Teknolojileri Dergisi, 8(1), 9-19.

Onan, A. (2016). Classifier and feature set ensembles for web page classification. Journal of Information Science, 42(2), 150-165.

Onan, A. (2017). Hybrid supervised clustering based ensemble scheme for text classification. Kybernetes, 46(2), 330-348.

Onan, A., Bulut, H., & Korukoglu, S. (2017). An improved ant algorithm with LDA-based representation for text document clustering. Journal of Information Science, 43(2), 275-292.

Pantalone, C. C., & Platt, M. B. (1987). Predicting commercial bank failure since deregulation. New England Economic Review, (Jul), 37-47.

Rokach, L. (2010). Ensemble-based classifiers. Artificial Intelligence Review, 33(1), 1-39.

Shatkay, H., & Craven, M. (2012). Mining the biomedical literature. MIT Press.

Tsai, C. F., Hsu, Y. F., & Yen, D. C. (2014). A comparative study of classifier ensembles for bankruptcy prediction. Applied Soft Computing, 24, 977-984.

Wang, G., Ma, J., & Yang, S. (2014). An improved boosting based on feature selection for corporate bankruptcy prediction. Expert Systems with Applications, 41(5), 2353-2361.

Xiao, H., Xiao, Z., & Wang, Y. (2016). Ensemble classification based on supervised clustering for credit scoring. Applied Soft Computing, 43, 73-86.

Yang, P., Hwa Yang, Y., B Zhou, B., & Y Zomaya, A. (2010). A review of ensemble methods in bioinformatics. Current Bioinformatics, 5(4), 296-308.

Zhou, Z-H. (2012), Ensemble methods: foundations and algorithms, Chapman and Hall, New York, NY.

Zięba, M., Tomczak, S. K., & Tomczak, J. M. (2016). Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. Expert Systems with Applications, 58, 93-101.