BugHunter Veri Setinde Özellik Azaltımı ve Başarı Metriklerinin Kapsamlı İncelenmesi

Abdullah Veli Özcan

Yazılım Mühendisliği, Manisa Celal Bayar Üniversitesi

192803015@ogr.cbu.edu.tr

**Özetçe** — Bu çalışmada, BugHunter metrik veri seti üzerinde özellik seçimi ve azaltma yöntemlerinin başarı metriklerine etkisi incelenmiştir. Araştırmanın temel amacı, veri setinden en etkili özellikleri seçerek sınıflandırma başarısını artırmak ve veri işleme süreçlerini optimize etmektir. İlk olarak, OrientDB veritabanında, 98 özelliği 85'e indirgeyerek Logistic Regression, Multilayer Perceptron ve çeşitli sınıflandırıcıları içeren bir Voting Classifier modeli uygulanmıştır. Daha sonra, bu 85 özelliğe PCA uygulanarak özellik sayısı 21'e düşürülmüş ve aynı modeller tekrar test edilmiştir. Ayrıca, 85 özelliğin yüksek korelasyona sahip 55'ine PCA uygulanmadan ve uygulandıktan sonra (14 özelliğe indirgenmiş) yine aynı sınıflandırıcılar kullanılmıştır. İkinci yaklaşım olarak MapDB verisi üzerinde, tüm özellikler ve yüksek korelasyona sahip özellikler ile yapılan modellemelerin doğruluk oranları karşılaştırılmıştır. Her iki yaklaşımın sonuçları karşılaştırılarak, özellik azaltmanın model performansına etkileri değerlendirilmiştir.

Anahtar Kelimeler— BugHunter Veri Seti, Özellik Seçimi, Principal Component Analysis (PCA), Voting Classifier, Dummy Classifier, Logistic Regression, Multilayer Perceptron, Yazılım Kriterleri, Doğruluk Oranı

1. GİRİŞ

Veri setlerinin analizi ve özellik seçimi, makine öğrenimi ve veri madenciliği alanlarında başarı için kritik bir öneme sahiptir. Özellikle, yazılım mühendisliği alanında, büyük ve karmaşık veri setlerinin işlenmesi, modelin performansını doğrudan etkileyebilir. Bu çalışmanın amacı, özellik seçimi ve azaltmanın sınıflandırma modelleri üzerindeki etkilerini incelemektir. Bu kapsamda, BugHunter metrik veri seti içerisinde bulunan 98 adet özellik üzerinde incelemeler yapılmıştır.

Yazılım metrikleri, yazılım geliştirme süreçlerinin ve ürünlerinin çeşitli yönlerini ölçmek amacıyla kullanılan nicel verilerdir. Bu metrikler, yazılımın kalitesini, performansını, verimliliğini ve diğer önemli özelliklerini değerlendirmek için kullanılmaktadır. Çalışmada,veri setinden bulunan metriklerin nasıl seçileceği ve azaltılacağı üzerine odaklanılmaktadır. Özellikle, veri setinin boyutunun azaltılması ve bu işlemin sınıflandırma modellerinin doğruluğu üzerindeki etkileri ele alınmaktadır.

OrientDB ve MapDB üzerinde yapılan çalışmalarda elde edilen metrikler ve hata sayıları kullanılarak,çeşitli sınıflandırma modelleri test edilmiştir. Bu modeller, özellik azaltma teknikleriyle birleştirilerek, model performansları üzerindeki etkisini değerlendirmek için kullanılmıştır.

1. İLGİLİ ÇALIŞMALAR

Özellik seçimi konusunda model tabanlı yaklaşımlar, model performansıyla daha sıkı bir şekilde ilişkilidir ve öngörücüler arasındaki korelasyon yapısını önem hesaplamasına dahil edebilir. Bu alandaki çalışmalar, Random Forests, Recursive Feature Elimination (RFE) ve Boruta gibi algoritmaları içermektedir. RFE, özelliklerin SVM modelindeki işbirliğini sıralamak için kullanılmakta ve ilgisiz özellikleri modelden çıkarmaktadır [1].

Özellik seçiminin yazılım sistemlerindeki hata sayısını tahmin etme üzerindeki etkisini inceleyen önemli bir araştırma, Osman, Ghafari ve Nierstrasz tarafından gerçekleştirilmiştir. "Hata Sayısını Tahmin Etmek Üzerinde Özellik Seçiminin Etkisi" başlıklı çalışmada, beş farklı açık kaynak Java yazılım sisteminde beş yaygın kullanılan tahmin modelinin özellik seçimi filtre ve sarıcı yöntemleri kullanılarak hata sayısını tahmin etme performansı değerlendirilmiştir. Bulgular, sarıcı yöntemlerin CFS filtresinden daha üstün olduğunu ve tahmin doğruluğunu önemli ölçüde artırabildiğini göstermektedir *[2].*

1. KULLANILAN YAKLAŞIM VE ARAÇLAR
2. Kullanılan Yaklaşım

Bu çalışmada, OrientDB ve MapDB veritabanları üzerinde yapılan analizler için Python programlama dili kullanılmıştır. Veri işleme ve analizi, Kaggle Notebook ortamında gerçekleştirilmiştir. Özellik seçimi ve sınıflandırma modellemesi için çeşitli Python kütüphaneleri ve algoritmaları uygulanmıştır. Bu kapsamda, Logistic Regression, Multilayer Perceptron ve Voting Classifier gibi makine öğrenimi modelleri ile PCA (Principal Component Analysis) gibi özellik azaltma teknikleri kullanılmıştır. Bu çalışmanın amacı, farklı veritabanlarındaki özelliklerin sınıflandırma başarısına olan etkisini değerlendirmek ve bu özelliklerin seçim sürecinin optimize edilmesidir.

1. Kullanılan Araçlar

Bu çalışmada, Kaggle Notebook ortamında veri analizi ve modelleme süreçleri gerçekleştirilmiştir. Veri işleme ve makine öğrenimi uygulamaları için Python programlama dili tercih edilmiştir. Modellemelerde, DummyClassifier basit tahminler için kullanılmış, çeşitli sınıflandırıcıların tahminlerini birleştiren Voting Classifier, derin öğrenme temelli Multilayer Perceptron (MLP) ve ikili sınıflandırma problemleri için Logistic Regression kullanılmıştır. Ayrıca, farklı makine öğrenimi modellerinin performanslarını hızlıca değerlendirmek amacıyla LazyPredict kütüphanesi de kullanılmıştır. Bu araçların entegrasyonu, veri seti üzerindeki çeşitli sınıflandırma modellerinin etkinliğinin değerlendirilmesine olanak sağlamıştır.

1. VERİ

Bu çalışmada kullanılan BugHunter Veri Seti*[3]*, yazılım sistemlerindeki hata tahminine yönelik detaylı bilgiler içerir. Veri seti, çeşitli yazılım metrikleri ve hata bilgilerini kapsar. Dosyalar, sınıflar ve metodlar gibi farklı yazılım elemanlarına ait geniş kapsamlı kod metrikleri ve hata detayları sunar. Makine öğrenimi modellerinin eğitiminde kullanılmak üzere tasarlanmış olan bu veri seti, hata tahmin modellerinin geliştirilmesi ve değerlendirilmesi için temel bir kaynak olarak hizmet eder.

metin, ekran görüntüsü, diyagram, daire içeren bir resim

Açıklama otomatik olarak oluşturuldu

**Şekil. 1** Projelere Göre Commit Sayısı

Şekil 1, projelerdeki commit sayısının oranını göstererek, projelerin aktivitesini ve büyüklüğünü ortaya koymaktadır. Bu çalışma için Şekil 1’de yer alan projelerden OrientDB ve MapDB kullanılmıştır.

TABLO I  
VERİ SETİNİN YAPISAL ÖZELLİKLERİ

|  |  |  |
| --- | --- | --- |
| Veri Seti | Satır Sayısı | Sütun Sayısı |
| OrientDB | 3802 | 98 |
| MapDB | 816 | 98 |

Tablo 1'de belirtildiği gibi, çalışmada kullanılan OrientDB veri setinde 3802 satır ve 98 sütun bulunmaktadır. MapDB veri setinde ise 816 satır ve 98 sütun mevcuttur. OrientDB ve MapDB veri setlerindeki özellikler birbiriyle aynıdır. Her iki veri seti de aynı 98 özellik üzerinden incelenmiştir. Çalışmada kullanılan veri setinde eksik veri bulunmamaktadır.

TABLO 2  
VERİ SETİNDEKİ ÖZELLİK TİPLERİ

|  |  |  |
| --- | --- | --- |
| Veri Seti | Kategorik Özellik | Numerik Özellik |
| OrientDB | 2 | 96 |
| MapDB | 2 | 96 |

Tablo 2’de, her iki veri setine de ait olan özelliklerin tipi belirtilmiştir.

metin, diyagram, çizgi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Açıklama otomatik olarak oluşturuldu

**Şekil. 2** OrientDB Veri Setinde Hedef Sınıfının Dağılımı

Şekil 2’de OrientDB veri setindeki hedef sınıf olan “Number of Bugs” sınıfının dağılımı gösterilmiştir.

diyagram, metin, çizgi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Açıklama otomatik olarak oluşturuldu

**Şekil. 3** MapDB Veri Setinde Hedef Sınıfının Dağılımı

Şekil 3’de MapDB veri setindeki hedef sınıf olan “Number of Bugs” sınıfının dağılımı gösterilmiştir.

Veri setinin hedef sınıfına ait dağılımın dengesiz olduğu, Şekil 2 ve Şekil 3'ün incelenmesiyle belirtilebilir. Bu durum, modellemelerin performansını olumsuz yönde etkilemektedir.

1. DEĞERLENDİRME

Bu çalışmada, OrientDB ve MapDB veri setleri üzerinde üç aşamalı bir modelleme işlemi gerçekleştirilmiştir. İlk aşamada her iki veri setinde 85 özellik kullanılarak Logistic Regression ve MLP modelleri uygulanmıştır. İkinci aşamada, OrientDB veri setine PCA uygulanarak özellik sayısı 85'ten 21'e düşürülmüş ve bu sete Logistic Regression, MLP ve bir Voting Classifier modeli uygulanmıştır. Üçüncü aşamada, düşük korelasyona sahip özellikler çıkarılarak OrientDB için 55, MapDB için 6 özellikle modeller tekrar uygulanmıştır. MapDB için DummyClassifier, OrientDB için ise PCA sonrası 14 özelliğe indirgenmiş set üzerinde tekrar modeller test edilmiştir.

OrientDB ve MapDB veri setlerinden "Hash" ve "LongName" gibi model başarısına etki etmeyen sütunlar çıkarılmıştır. Ayrıca, tüm değerleri sıfır olan 13 numerik sütun da benzer sebeplerle kaldırılmıştır. Her iki veri seti, bağımsız değişkenler (X) ve hedef değişken ('Number of Bugs' yani Y) olarak ayrıldı. Test boyutu %20 olarak ayarlanarak her iki set için eğitim ve test verileri oluşturulmuştur. Logistic Regression ve MLP modelleri uygulanmıştır. Elde edilen sonuçlar Tablo 3’de gösterilmiştir:

TABLO 3  
85 ÖZELLİKLE MODEL DOĞRULUK DEĞERLERİ

|  |  |  |
| --- | --- | --- |
| Veri Seti | Logistic Regression | Multilayer Perceptron |
| OrientDB | %55.98 | %43.36 |
| MapDB | %50.61 | %42.07 |

OrientDB veri setindeki özellik sayısını 85'ten 21'e indirgemeyi sağlayan bir PCA (Principal Component Analysis) yöntemi kullanılmıştır. Bu indirgeme, veri setindeki %95 kümülatif varyansı koruyacak şekilde gerçekleştirilmiştir. Bu yaklaşım, veri setinin boyutunu azaltırken aynı zamanda önemli bilgilerin büyük bir kısmını muhafaza etmeyi sağlar, böylece modelin etkinliği ve hesaplama verimliliği artırılmıştır.

Şekil 4’te OrientDB veri setinde %95 kümülatif varyansa ulaşma hızı gösterilmiştir.

metin, ekran görüntüsü, çizgi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Açıklama otomatik olarak oluşturuldu

**Şekil. 4** OrientDB Veri Setinde Kümülatif Varyans Grafiği

TABLO 4  
PCA UYGULANMIŞ SETTE DOĞRULUK DEĞERLERİ

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Veri Seti | Özellik Sayısı | Logistic Regression | Multilayer Perceptron | Voting Classifier |
| OrientDB | 21 | %56.90 | %47.96 | %55.72 |

Tablo 4’te gösterilen %55.72 doğruluk değerini sağlamış olan Voting Classifier içerisinde Adaboost Classifier, Ridge Classifier, Logistic Regression, SVC ve Dummy Classifier bulunmaktadır. Bu modeller LazyPredict kütüphanesinin kullanımı sonrasında en yüksek doğruluk değerlerine sahip olan 5 modeldir.

Çalışmada, OrientDB veri setinden korelasyonu 0.1'den düşük olan özellikler çıkarılarak, 55 özellik içeren bir set oluşturulmuştur. Bu set üzerinde Logistic Regression, MLP ve Voting Classifier modelleri test edilmiştir. MapDB veri seti için ise, 0.19'dan düşük korelasyonlu özellikler atılarak 6 özellikli bir set hazırlanmış ve bu sete Logistic Regression ve DummyClassifier modelleri uygulanmıştır. Elde edilen sonuçlar, Tablo 5 ve 6'da detaylı olarak verilmiştir. Bu sonuçlar, farklı özellik setlerinin ve uygulanan modellerin her iki veri setindeki performansını karşılaştırmaktadır.

TABLO 5  
YÜKSEK KORELASYONLU ÖZELLİKLERİN DOĞRULUK DEĞERLERİ

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Veri Seti | Özellik Sayısı | Logistic Regression | Multilayer Perceptron | Voting Classifier |
| OrientDB | 55 | %55.71 | %52.04 | %56.37 |

TABLO 6  
YÜKSEK KORELASYONLU ÖZELLİKLERİN DOĞRULUK DEĞERLERİ

|  |  |  |  |
| --- | --- | --- | --- |
| Veri Seti | Özellik Sayısı | Logistic Regression | Dummy Classifier |
| MapDB | 6 | %52.41 | %62.19 |

OrientDB’ye ait olan yüksek korelasyonlu özellikleri bulunduran veri setindeki 55 özelliğe PCA yöntemi uygulanmıştır. %95 kümülatif varyansı koruyacak şekilde gerçekleştirilen bu işlemde 55 olan özellik sayısı 14’e indirgenmiştir. Şekil 5’te OrientDB veri setinde %95 kümülatif varyansa ulaşma hızı gösterilmiştir.

metin, ekran görüntüsü, çizgi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Açıklama otomatik olarak oluşturuldu

**Şekil. 5** OrientDB Yüksek Korelasyonlu Sette Kümülatif Varyans Grafiği

PCA uygulanarak indirgenmiş olan 14 özellik ile modelleme tekrar yapılmış, elde edilen doğruluk değerleri Tablo 7’de gösterilmiştir.

TABLO 7  
PCA UYGULANMIŞ SETTE DOĞRULUK DEĞERLERİ

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Veri Seti | Özellik Sayısı | Logistic Regression | Multilayer Perceptron | Voting Classifier |
| OrientDB | 14 | %56.11 | %53.35 | %58.16 |

1. SONUÇ

Çalışmanın sonuçlarında, OrientDB ve MapDB veri setlerine uygulanan modelleme tekniklerinin etkinliği değerlendirilmiştir. İlk olarak, 85 özellik kullanılarak elde edilen doğruluk oranları; OrientDB için Logistic Regression %55.98, MLP %43.36 ve MapDB için Logistic Regression %50.61, MLP %42.07 olarak saptanmıştır. PCA ile 21 özelliğe indirgenmiş OrientDB setinde Logistic Regression %56.90, MLP %47.96 ve Voting Classifier %55.72 doğruluk oranlarına ulaşılmıştır. 55 özellikli OrientDB setinde Logistic Regression %55.71, MLP %52.04 ve Voting Classifier %56.37; 6 özellikli MapDB setinde Logistic Regression %52.41 ve Dummy Classifier %62.19 doğruluk oranları gözlemlenmiştir. 14 özelliğe indirgenmiş OrientDB setinde ise Logistic Regression %56.11, MLP %53.35 ve Voting Classifier %58.16 oranları elde edilmiştir.

Bu çalışmanın sonucunda, OrientDB ve MapDB veri setlerine uygulanan çeşitli modelleme tekniklerinin değerlendirilmesinde, ilk aşamada, her iki veri setinde 85 özellik kullanılarak elde edilen doğruluk oranları, PCA uygulaması ve yüksek korelasyonlu özelliklerle yapılan analizler model performansını artırmıştır. Özellikle, PCA ile azaltılmış özellik setlerinde ve yüksek korelasyonlu özelliklerle yapılan modellemelerde gözlenen doğruluk artışları, veri setlerinin doğru şekilde işlenmesinin önemini vurgulamaktadır.

kaynakça

1. Saeed, F., Paul, A., & Ahmad, A. (2020). Selecting critical features for data classification based on machine learning methods. Journal of Big Data, 7(1). <https://doi.org/10.1186/s40537-020-00327-4>
2. Osman, H., Ghafari, M., & Nierstrasz, O. (2018). The Impact of Feature Selection on Predicting the Number of Bugs. Retrieved from <https://ar5iv.org/abs/1807.04486>
3. Ferenc, R., Gyimesi, P., & Gyimesi, G. (2020). BugHunter Dataset [Veri seti]. Mendeley Data. <https://doi.org/10.17632/8tx7kjbkg4.2>
4. Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), 20150202. <https://doi.org/10.1098/rsta.2015.0202>