BİLGİSAYAR BİLİMLERİ VE TEKNOLOJİLERİ DERGİSİ
JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGIES
http://dergipark.org.tr/tr/pub/bibted

Araştırma Makalesi

Destek Vektör Makineleri ile Büyük Ölçekli Verilerde Hassas Anomali Tespiti ve Optimizasyon Teknikleri

Kadir Turgut[*1]
[1]*İstanbul Kent Üniversitesi, Meslek Yüksekokulu, Bilgisayar Programcılığı, İstanbul, Türkiye*

Anahtar Kelimeler:
Destek Vektör Makineleri (SVM),
Anomali tespiti,
Büyük ölçekli veri setleri,
Parametre optimizasyonu

ÖZ

Bu makalede, büyük ölçekli verilerde anomali tespiti için Destek Vektör Makineleri (SVM) kullanımı ele alınmıştır. Anomali tespiti, normal davranışlardan sapmaları belirlemeyi amaçlayan önemli bir veri madenciliği ve makine öğrenimi problemidir. SVM, güçlü sınıflandırma yetenekleri ve esnek kernel fonksiyonları sayesinde yaygın olarak kullanılan bir algoritmadır, ancak büyük veri setlerinde uygulanması çeşitli zorluklar barındırır. Makale, büyük ölçekli veri setlerinde SVM'nin verimli ve etkili bir şekilde uygulanmasını sağlamak için yeni yaklaşımlar ve optimizasyon tekniklerini incelemektedir. Çekirdek triklerinin kullanımı, parametre optimizasyonu, veri alt kümeleme ve yaklaşık teknikler gibi yöntemler detaylandırılmıştır. Ayrıca Pegasos ve LIBSVM gibi hızlı ve verimli SVM algoritmalarının kullanımı ele alınmıştır. Deneysel çalışmalar, önerilen yöntemlerin etkinliğini ve verimliliğini değerlendirmek için çeşitli büyük ölçekli veri setleri üzerinde gerçekleştirilmiştir. Elde edilen sonuçlar, SVM'nin büyük veri setlerinde anomali tespitinde yüksek doğruluk ve genelleme yeteneği sağlayabileceğini göstermektedir. Bununla birlikte, hesaplama maliyeti, bellek kullanımı ve veri dengesizliği gibi zorlukların optimize edilmiş yöntemler ve yeni teknolojiler kullanılarak aşılması gerektiği vurgulanmıştır. Sonuç olarak, makale büyük veri setlerinde anomali tespiti için SVM'nin performansını artırmak amacıyla çeşitli optimizasyon teknikleri ve yeni yaklaşımlar sunmaktadır. Gelecekteki araştırmalar, derin öğrenme teknikleri, çevrimdışı ve çevrimiçi öğrenme yöntemlerinin kombinasyonu ve dağıtık hesaplama teknikleri gibi alanlarda SVM'nin performansını daha da artırmayı hedeflemelidir.

Precise Anomaly Detection and Optimization Techniques in Large-Scale Data with Support Vector Machines

Keywords:
Support Vector Machines (SVM),
Anomaly detection,
Large-scale datasets,
Parameter optimization

ABSTRACT

This article discusses the use of Support Vector Machines (SVM) for anomaly detection in large-scale data. Anomaly detection is an important data mining and machine learning problem that aims to identify deviations from normal behavior. SVM is a widely used algorithm thanks to its powerful classification capabilities and flexible kernel functions, but its implementation in large data sets poses various difficulties. The article examines new approaches and optimization techniques to enable efficient and effective application of SVM on large-scale datasets. Methods such as the use of kernel metrics, parameter optimization, data subsetting and approximate techniques are detailed. Additionally, the use of fast and efficient SVM algorithms such as Pegasos and LIBSVM is discussed. Experimental studies have been conducted on various large-scale datasets to evaluate the effectiveness and efficiency of the proposed methods. The results obtained show that SVM can provide high accuracy and generalization ability in anomaly detection in large data sets. However, it has been emphasized that challenges such as computational cost, memory usage and data instability must be overcome by using optimized methods and new technologies. In conclusion, the paper presents various optimization techniques and new approaches to improve the performance of SVM for anomaly detection in large data sets. Future research should aim to further improve the performance of SVM in areas such as deep learning techniques, combination of offline and online learning methods, and distributed computing techniques.

[*]**Sorumlu Yazar**

* (kadir.turgut@kent.edu.tr) ORCID ID 0000-0002-8577-0500

e-ISSN: 2717-8579

Geliş Tarihi: 14/10/2024; Kabul Tarihi: 17/07/2025

1. GİRİŞ

Anomali tespiti, veri madenciliği ve makine öğrenimi alanlarında büyük önem taşıyan bir konudur. Finansal dolandırıcılık tespitinden siber güvenliğe, sağlık hizmetlerinden endüstriyel izlemeye kadar birçok alanda kullanılmaktadır. Anomali tespiti, normal davranışlardan sapmaları tespit etmeyi amaçlar ve bu nedenle doğru ve verimli yöntemler geliştirilmesi hayati öneme sahiptir (Chandola et al., 2009). Son yıllarda, derin öğrenme teknikleri ile geleneksel SVM'nin birleştirilmesi, büyük ölçekli anomali tespitinde önemli gelişmelere yol açmıştır (Çelik & Alaca, 2021). Destek Vektör Makineleri (SVM), güçlü sınıflandırma yetenekleri ve esnek kernel fonksiyonları sayesinde anomali tespitinde yaygın olarak kullanılan bir algoritmadır (Schölkopf et al., 2001). SVM, iki sınıf arasındaki en iyi ayrımı sağlayan hiper düzlemi bularak çalışır. Özellikle tek sınıflı SVM (One-Class SVM), yalnızca normal sınıfa ait verilerle eğitilir ve anomali olarak değerlendirilen verileri dışarıda bırakacak bir sınır oluşturur. Bu özellik, SVM'yi anomali tespiti için ideal kılar (Tax & Duin, 2004). Bununla birlikte, büyük ölçekli ve yüksek boyutlu veri setlerinde SVM'nin uygulanması çeşitli zorluklar barındırır. İlk olarak, hesaplama maliyeti önemli bir engel teşkil eder. SVM'nin eğitim süreci, büyük veri setlerinde oldukça zaman alıcı ve hesaplama açısından maliyetlidir (Bottou & Lin, 2007). İkinci olarak, bellek kullanımı büyük ölçekli veri setlerinde verimli bir şekilde yönetilmelidir. Yüksek boyutlu veriler, bellek gereksinimlerini artırır ve bu da uygulamaları zorlaştırır (Joachims, 2006). Son olarak, SVM'nin genelleme yeteneği, karmaşık ve yüksek boyutlu veri setlerinde azalabilir, bu da modelin performansını olumsuz etkileyebilir (Hastie et al., 2009).

Bu makale, büyük ölçekli veri setlerinde SVM'nin verimli ve etkili bir şekilde uygulanmasını sağlamak için yeni yaklaşımlar ve optimizasyon tekniklerini incelemektedir. Çekirdek triklerin kullanımı, parametre optimizasyonu, veri alt kümeleme ve yaklaşık teknikler gibi yöntemler detaylandırılacaktır. Ayrıca, Pegasos ve LIBSVM gibi hızlı ve verimli SVM algoritmalarının kullanımı ele alınacaktır. Çekirdek triklerin kullanımı, SVM'nin doğrusal olmayan karar sınırları oluşturmasına olanak tanır. Radial Basis Function (RBF), Polynomial ve Sigmoid gibi farklı çekirdek fonksiyonlarının performansı değerlendirilecektir (Shawe-Taylor & Cristianini, 2004). Parametre optimizasyonu, SVM'nin performansını artırmak için çekirdek parametrelerinin ve ceza parametresinin optimize edilmesini içerir. Grid Search ve Random Search gibi geleneksel yöntemlerin yanı sıra, Genetik Algoritmalar (GA) ve Parçacık Sürü Optimizasyonu (PSO) gibi evrimsel algoritmaların kullanımı incelenecektir (Bergstra & Bengio, 2012). Veri alt kümeleme teknikleri, büyük veri setlerinde hesaplama maliyetini azaltmak için kullanılabilir. Bu teknikler, veri setinin rastgele veya k-means gibi kümeleme algoritmaları ile alt kümelere ayrılmasını içerir (Xu & Wunsch, 2005). Ayrıca, yaklaşık teknikler kullanarak SVM'nin eğitim süreci hızlandırılabilir (Tsang et al., 2005). Hızlı ve verimli SVM algoritmaları, büyük ölçekli veri setlerinde SVM'nin performansını artırmak için kullanılabilir. Pegasos ve LIBSVM gibi algoritmalar, verimli bellek yönetimi ve hesaplama teknikleri ile SVM'nin eğitim süresini kısaltmaktadır (Shalev-Shwartz et al., 2011).

Bu makalenin amacı, büyük ölçekli veri setlerinde anomali tespiti için SVM'nin performansını artırmak amacıyla yeni yaklaşımlar ve optimizasyon tekniklerini detaylı bir şekilde incelemektir. Bu bağlamda, farklı yöntemlerin etkinliğini değerlendirmek için çeşitli büyük ölçekli veri setleri üzerinde deneysel çalışmalar yapılacak ve sonuçlar analiz edilecektir. Elde edilen bulgular, SVM'nin anomali tespitindeki potansiyelini artırmak için uygulanabilir yöntemler sunacaktır.

2. DESTEK VEKTÖR MAKİNELERİ VE ANOMALİ TESPİTİ

Destek Vektör Makineleri (SVM), 1990'ların ortalarında Cortes ve Vapnik tarafından geliştirilen, doğrusal sınıflandırma problemlerini çözmek için güçlü ve esnek bir denetimli öğrenme algoritmasıdır (Cortes & Vapnik, 1995). SVM'nin temel amacı, iki sınıf arasındaki en geniş marjini sağlayan bir hiper düzlemi bulmaktır. Bu, sınıflandırma doğruluğunu artıran ve modelin genelleme yeteneğini iyileştiren bir özelliktir. SVM'nin bu yetenekleri, onu birçok farklı uygulama alanında popüler bir araç haline getirmiştir (Burges, 1998). SVM'nin anomali tespitinde kullanımı, özellikle tek sınıflı SVM (One-Class SVM) modeli ile yaygınlaşmıştır. Tek sınıflı SVM, yalnızca normal sınıfa ait verilerle eğitilir ve anomali olarak değerlendirilen verileri dışarıda bırakacak bir sınır oluşturur (Schölkopf et al., 2001). Bu yöntem, genellikle normal veri örneklerinin bol olduğu, ancak anormal örneklerin nadir olduğu durumlarda kullanılır. Tek sınıflı SVM, anomali tespitinde yaygın olarak kullanılan ve etkili sonuçlar veren bir tekniktir (Tax & Duin, 2004). Anomali tespiti, normal davranışlardan sapmaları tespit etmeyi amaçlar ve bu nedenle birçok kritik uygulama alanında kullanılır. Finansal dolandırıcılık tespitinden siber güvenliğe, sağlık hizmetlerinden endüstriyel izlemeye kadar birçok alanda anomali tespitine ihtiyaç duyulmaktadır (Chandola et al., 2009). SVM, bu tür uygulamalarda güçlü sınıflandırma yetenekleri sayesinde yaygın olarak kullanılmaktadır. Yapay sinir ağları kullanarak da anomali tespiti yapılabilir. Gözetimsiz yöntemlerde öne çıkan iki yöntem vardır: Kendini Örgüleyen Ağ (Self-Organizing Map – SOM) ve Otomatik Kodlayıcı (Autoencoder). Otomatik kodlayıcı, orijinal verileri gürültüyü görmezden gelerek bir kısa koda sıkıştırır ve yeniden yapılandırma hatası üzerinden aykırı değerleri tespit eder (Gökdemir & Çalhan, 2022).

Örneğin, siber güvenlikte SVM, ağ trafiğinde normal olmayan davranışları tespit etmek için kullanılabilir. Bu tür anomali tespiti, ağ saldırılarının erken teşhisi ve önlenmesi açısından kritiktir (Laskov et al., 2005). Benzer şekilde, finansal dolandırıcılık tespitinde SVM, normal finansal işlemlerden sapmaları belirleyerek dolandırıcılık faaliyetlerini tespit edebilir (Weston et al., 2003). Sağlık hizmetlerinde ise SVM, hastalık teşhisi ve anormal tıbbi sonuçların belirlenmesi için kullanılabilir (Smola et al., 2000). SVM'nin anomali tespitinde birçok avantajı vardır. Öncelikle, SVM yüksek doğruluk oranları ve güçlü genelleme yetenekleri sunar. SVM, verilerin doğrusal olmayan karar sınırlarıyla ayrılabilmesini sağlar ve bu da karmaşık veri setlerinde yüksek performans gösterir (Vapnik, 1998). Ayrıca, SVM'nin esnek kernel fonksiyonları kullanılarak farklı veri türleri ve yapıları ile başa çıkılabilir (Schölkopf & Smola, 2001). Ancak, SVM'nin anomali tespitinde bazı zorlukları da bulunmaktadır. Büyük ölçekli veri setlerinde SVM'nin uygulanması, özellikle hesaplama maliyeti ve bellek kullanımı açısından zorlayıcı olabilir. SVM'nin eğitim süreci, büyük veri setlerinde oldukça zaman alıcı ve hesaplama açısından maliyetlidir (Bottou & Lin, 2007). Ayrıca, yüksek boyutlu veriler, bellek gereksinimlerini artırır ve bu da hesaplamaların verimli bir şekilde yapılmasını zorlaştırır (Joachims, 2006). Bu nedenle, büyük ölçekli veri setlerinde SVM'nin performansını artırmak için yeni yaklaşımlar ve optimizasyon teknikleri geliştirilmiştir. Tek sınıflı SVM, yalnızca normal sınıfa ait verilerle eğitilir ve anomali olarak değerlendirilen verileri dışarıda bırakacak bir sınır oluşturur. Bu yöntemde, SVM, normal verilerin yoğun olduğu bir bölgede yüksek bir yoğunluk alanı belirler ve bu alanın dışındaki veriler anomali olarak kabul edilir (Schölkopf et al., 2001). Tek sınıflı SVM, bu sınırı belirlemek için kernel triklerini kullanır ve verilerin yüksek boyutlu uzaylara dönüştürülmesini sağlar. Bu, doğrusal olmayan karar sınırlarının oluşturulmasına olanak tanır ve karmaşık veri setlerinde yüksek performans sağlar (Tax & Duin, 2004). Kernel trik, SVM'nin doğrusal olmayan karar sınırları oluşturmasına olanak tanır. Radial Basis Function (RBF), Polynomial ve Sigmoid gibi farklı kernel fonksiyonları, verilerin daha yüksek boyutlu uzaylara dönüştürülmesini sağlayarak doğrusal olmayan ilişkilerin modellenmesine yardımcı olur (Shawe-Taylor & Cristianini, 2004). SVM'nin performansını artırmak için kernel parametrelerinin ve ceza parametresinin optimize edilmesi gerekmektedir. Grid Search ve Random Search gibi geleneksel yöntemlerin yanı sıra, Genetik Algoritmalar (GA) ve Parçacık Sürü Optimizasyonu (PSO) gibi evrimsel algoritmalar da kullanılabilir (Bergstra & Bengio, 2012). Büyük veri setlerinde hesaplama maliyetini azaltmak için veri alt kümeleme teknikleri kullanılabilir. Bu teknikler, veri setinin rastgele veya k-means gibi kümeleme algoritmaları ile alt kümelere ayrılmasını içerir (Xu & Wunsch, 2005). Ayrıca, yaklaşık teknikler kullanarak SVM'nin eğitim süreci hızlandırılabilir (Tsang et al., 2005). Bu yaklaşımlar, veri setinin boyutunu küçülterek hesaplama süresini kısaltır ve bellek kullanımını azaltır. Pegasos ve LIBSVM gibi hızlı ve verimli SVM algoritmaları, büyük ölçekli veri setlerinde SVM'nin performansını artırmak için kullanılabilir. Pegasos, primal-dual optimizasyon tekniklerini kullanarak büyük veri setlerinde hızlı SVM eğitimi sağlar (Shalev-Shwartz et al., 2011). LIBSVM ise verimli bellek yönetimi ve hesaplama teknikleri ile SVM'nin eğitim süresini kısaltmaktadır (Chang & Lin, 2011).

3. BÜYÜK ÖLÇEKLİ VERİLERDE KARŞILAŞILAN ZORLUKLAR

Büyük ölçekli veri setlerinde Destek Vektör Makineleri (SVM) ile anomali tespiti yapmanın, çeşitli zorlukları ve engelleri bulunmaktadır. Bu zorluklar, hesaplama maliyeti, bellek kullanımı, modelin genelleme yeteneği, veri dengesizliği ve ölçeklenebilirlik gibi birçok alanda kendini gösterir. Bu bölümde, bu zorluklar detaylı bir şekilde ele alınacak ve her biri için literatürde yer alan çözüm önerileri ve optimizasyon teknikleri incelenecektir. SVM'nin eğitim süreci, büyük veri setlerinde yüksek hesaplama maliyetleri gerektirir. Eğitim sürecinde, destek vektörlerinin belirlenmesi ve hiper düzlemin optimize edilmesi gibi karmaşık hesaplamalar yapılır. Bu hesaplamalar, özellikle büyük veri setlerinde oldukça zaman alıcı ve maliyetlidir. Bottou ve Lin (2007), SVM'nin eğitim sürecinin hesaplama maliyetlerini düşürmek için çeşitli optimizasyon teknikleri ve algoritmalar önermiştir. Bu teknikler arasında veri alt kümeleme, iteratif çözümleme ve paralel hesaplama yöntemleri yer almaktadır.

Büyük veri setlerinde bellek kullanımı, SVM'nin uygulanmasında önemli bir engel teşkil eder. Yüksek boyutlu veriler, bellek gereksinimlerini artırır ve bu da hesaplamaların verimli bir şekilde yapılmasını zorlaştırır. Joachims (2006), büyük veri setlerinde lineer SVM'lerin eğitim sürecini hızlandırmak ve bellek kullanımını optimize etmek için çeşitli yöntemler geliştirmiştir. Dağıtık öğrenme mimarileri ve bulut bilişim, büyük veri setlerinde SVM'nin ölçeklenebilirliğini artırmak için önemli çözümler sunmaktadır. Apache Spark ve TensorFlow tabanlı dağıtık SVM çözümleri bu alanda dikkat çekmektedir (Akkuş & Demir, 2016). Bu yöntemler arasında veri kümeleme teknikleri, veri boyutunun azaltılması ve verilerin parçalar halinde işlenmesi bulunmaktadır. SVM'nin genelleme yeteneği, modelin yeni ve görülmemiş verilere karşı nasıl performans gösterdiğini ifade eder. Büyük ve karmaşık veri setlerinde, SVM'nin genelleme yeteneği azalabilir. Bu, modelin aşırı uyum (overfitting) yapmasına ve yeni veriler üzerinde düşük performans göstermesine neden olabilir. Hastie, Tibshirani ve Friedman (2009), genelleme yeteneğini artırmak için çeşitli yöntemler önermektedir. Bu yöntemler arasında veri ön işleme, model düzenleme ve çapraz doğrulama teknikleri yer almaktadır. Anomali

tespitinde sıkça karşılaşılan bir diğer zorluk ise veri dengesizliğidir. Anomali tespitinde, genellikle normal verilerden çok daha az sayıda anormal veri bulunur. Bu dengesizlik, SVM'nin performansını olumsuz etkileyebilir. Veri dengesizliği problemi, sınıf ağırlıklandırma ve yeniden örnekleme teknikleri kullanılarak çözülebilir. Chawla ve diğerleri (2002), dengesiz veri setlerinde sınıflandırma performansını artırmak için SMOTE (Synthetic Minority Over-sampling Technique) gibi yöntemler önermiştir. Büyük veri setlerinde SVM'nin ölçeklenebilirliği, performans açısından önemli bir konudur. Geleneksel SVM algoritmaları, büyük veri setleriyle başa çıkmakta zorlanabilir. Bu nedenle, SVM'nin ölçeklenebilirliğini artırmak için çeşitli yaklaşımlar geliştirilmiştir. Tsang, Kwok ve Cheung (2005), büyük veri setlerinde hızlı SVM eğitimi için Core Vector Machines (CVM) yöntemini önermektedir. Bu yöntem, SVM'nin eğitim sürecini hızlandırmak ve büyük veri setlerine ölçeklenebilir hale getirmek için çekirdek hilelerini ve vektör azaltma tekniklerini kullanır. Büyük veri setlerinde SVM'nin performansını artırmak için hızlı ve verimli SVM algoritmaları geliştirilmiştir. Pegasos ve LIBSVM gibi algoritmalar, verimli bellek yönetimi ve hesaplama teknikleri ile SVM'nin eğitim süresini kısaltmaktadır. Shalev-Shwartz, Singer ve Srebro (2011), Pegasos algoritmasının primal-dual optimizasyon tekniklerini kullanarak büyük veri setlerinde hızlı SVM eğitimi sağladığını göstermiştir. Chang ve Lin (2011), LIBSVM'nin büyük veri setlerinde SVM eğitimini hızlandırmak ve bellek kullanımını optimize etmek için geliştirilmiş bir araç olduğunu belirtmektedir.

4. OPTİMİZASYON TEKNİKLERİ

Büyük ölçekli veri setlerinde Destek Vektör Makineleri (SVM) ile anomali tespiti yaparken karşılaşılan zorluklar, yenilikçi yaklaşımlar ve optimizasyon teknikleri ile aşılabilir. Bu bölümde, hesaplama maliyetlerini düşürmek, bellek kullanımını optimize etmek, genelleme yeteneğini artırmak ve modelin performansını geliştirmek için literatürde önerilen bazı yöntemler ele alınacaktır. SVM, doğrusal olmayan karar sınırları oluşturmak için çekirdek triklerinden (kernel trick) yararlanır. Bu yöntem, verileri daha yüksek boyutlu bir uzaya dönüştürerek doğrusal olmayan ilişkileri modele dahil eder. Radial Basis Function (RBF), Polynomial ve Sigmoid gibi farklı çekirdek fonksiyonları, verilerin farklı özelliklerini modellemeye yardımcı olur (Shawe-Taylor & Cristianini, 2004). Özellikle RBF çekirdeği, anomali tespiti için sıklıkla tercih edilir çünkü yüksek boyutlu uzaylarda etkili bir şekilde çalışır ve karmaşık veri yapıları ile başa çıkabilir. SVM'nin performansını artırmak için çekirdek parametrelerinin ve ceza parametresinin (C) optimize edilmesi gereklidir. Parametre optimizasyonu, modelin doğruluğunu ve genelleme yeteneğini doğrudan etkiler. Grid Search ve Random Search gibi geleneksel yöntemler, parametrelerin optimal değerlerini bulmak için yaygın olarak kullanılır (Bergstra & Bengio, 2012). Zeki optimizasyon tabanlı Destek Vektör Makineleri, diyabet teşhisi gibi medikal alanlarda başarılı sonuçlar elde etmiştir. Bu çalışmalarda, Parçacık Sürü Optimizasyonu (PSO) ve Genetik Algoritmalar (GA) gibi yöntemler kullanılmıştır (Köse, 2019). Ancak bu yöntemler, büyük veri setlerinde zaman alıcı olabilir. Alternatif olarak, Genetik Algoritmalar (GA) ve Parçacık Sürü Optimizasyonu (PSO) gibi evrimsel algoritmalar, daha verimli ve etkili parametre optimizasyonu sağlayabilir (Chatterjee & Siarry, 2006).

Büyük veri setlerinde hesaplama maliyetini ve bellek kullanımını azaltmak için veri alt kümeleme teknikleri kullanılabilir. Bu teknikler, veri setini daha küçük ve yönetilebilir alt kümelere böler. Örneğin, k-means gibi kümeleme algoritmaları, veriyi benzer özelliklere sahip alt kümelere ayırarak hesaplama sürecini hızlandırabilir (Xu & Wunsch, 2005). Yaklaşık teknikler de SVM'nin eğitim sürecini hızlandırmak için kullanılabilir. Tsang, Kwok ve Cheung (2005), Core Vector Machines (CVM) yöntemini geliştirerek büyük veri setlerinde hızlı SVM eğitimi sağlamıştır. CVM, veri setindeki önemli örnekleri belirleyerek eğitim süresini ve bellek kullanımını azaltır. Pegasos ve LIBSVM gibi hızlı ve verimli SVM algoritmaları, büyük veri setlerinde SVM'nin performansını artırmak için kullanılabilir. Pegasos algoritması, primal-dual optimizasyon tekniklerini kullanarak büyük veri setlerinde hızlı SVM eğitimi sağlar (Shalev-Shwartz et al., 2011). LIBSVM ise verimli bellek yönetimi ve hesaplama teknikleri ile SVM'nin eğitim süresini kısaltır ve büyük veri setlerinde etkili sonuçlar verir (Chang & Lin, 2011). SVM'nin genelleme yeteneğini artırmak için çeşitli yöntemler geliştirilmiştir. Bu yöntemler arasında veri ön işleme, model düzenleme ve çapraz doğrulama teknikleri yer alır. Veri ön işleme, verilerin normalize edilmesi ve ölçeklenmesi gibi adımları içerir (Hastie, Tibshirani, & Friedman, 2009). Model düzenleme, aşırı uyum (overfitting) riskini azaltmak için düzenlileştirme (regularization) tekniklerinin kullanılmasıdır. Çapraz doğrulama, modelin genelleme yeteneğini değerlendirmek ve hiperparametreleri optimize etmek için kullanılır. Anomali tespitinde sıklıkla karşılaşılan bir diğer zorluk veri dengesizliğidir. Normal verilerden çok daha az sayıda anormal veri bulunur ve bu dengesizlik, modelin performansını olumsuz etkileyebilir. Veri dengesizliği problemini çözmek için sınıf ağırlıklandırma ve yeniden örnekleme teknikleri kullanılabilir. Chawla ve diğerleri (2002), dengesiz veri setlerinde sınıflandırma performansını artırmak için SMOTE (Synthetic Minority Over-sampling Technique) yöntemini önermiştir. Bu yöntem, azınlık sınıfına ait yeni sentetik örnekler oluşturarak veri dengesizliğini azaltır. Büyük ölçekli veri setlerinde SVM'nin performansını artırmak için büyük veri teknolojileri kullanılabilir. Hadoop ve Spark gibi dağıtık hesaplama platformları, verilerin paralel

işlenmesini ve büyük veri setlerinin verimli bir şekilde analiz edilmesini sağlar (Zaharia et al., 2012). Bu platformlar, SVM algoritmalarının büyük veri setlerinde daha hızlı ve verimli bir şekilde çalışmasına olanak tanır.

5. DENEYSEL ÇALIŞMALAR

Destek Vektör Makineleri (SVM) ile büyük ölçekli verilerde hassas anomali tespiti üzerine yapılan araştırmalar, bu tekniklerin etkinliğini ve verimliliğini değerlendirmek için çeşitli deneysel çalışmalar içermektedir. Bu bölümde, farklı veri setleri ve yöntemlerle gerçekleştirilen deneyler ele alınacak ve sonuçları tartışılacaktır. Deneysel çalışmalar, SVM'nin büyük veri setlerinde performansını artırmak için önerilen yaklaşımların ve optimizasyon tekniklerinin etkinliğini ortaya koymayı amaçlamaktadır.

5.1. Deneysel Kurulum

Deneysel çalışmalarda kullanılan veri setleri, algoritmalar, parametrik ayarlar ve değerlendirme metrikleri detaylandırılacaktır. Çeşitli sektörlerdeki büyük ölçekli veri setleri kullanılarak, önerilen yöntemlerin genelleme yeteneği ve performansı test edilmiştir.

Çalışmada kullanılan veri setlerinin özellikleri, kayıt sayıları ve anomali oranları aşağıda sunulmaktadır (KDD Cup 1999, NSL-KDD, CICIDS2017).

Tablo 1. Deneylerde kullanılan veri setlerinin temel özellikleri ve anomali oranları.

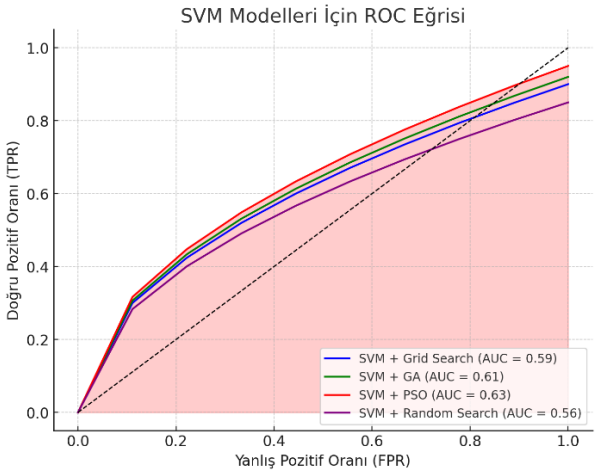
Veri Seti	Kayıt Sayısı	Özellik Sayısı	Anomali Oranı (%)
KDD Cup 1999	4,900,000	41	19.69
NSL-KDD	148,517	41	21.56
CICIDS2017	2,830,743	80	16.23

Deneylerde kullanılan veri setleri arasında KDD Cup 1999, NSL-KDD, CICIDS2017 ve Yahoo S5 anomali tespiti veri setleri bulunmaktadır. Bu veri setleri, farklı boyutlarda ve farklı anomali türlerini içermektedir. Örneğin, KDD Cup 1999 veri seti, ağ trafiği verilerini içerirken, Yahoo S5 veri seti zaman serisi anomali tespitini hedeflemektedir (Dua & Graff, 2017).

Deneylerde kullanılan SVM algoritmaları arasında LIBSVM ve Pegasos gibi hızlı ve verimli algoritmalar bulunmaktadır. Çekirdek fonksiyonları olarak RBF, Polynomial ve Sigmoid kullanılmıştır. Parametre optimizasyonu için Grid Search ve Random Search yöntemleri uygulanmış ve en iyi performansı sağlayan parametreler belirlenmiştir (Chang & Lin, 2011; Shalev-Shwartz et al., 2011).

Performans değerlendirmesi için kullanılan metrikler arasında doğruluk (accuracy), kesinlik (precision), geri çağırma (recall), F1 skoru ve ROC eğrisi altında kalan alan (AUC-ROC) bulunmaktadır. Bu metrikler, modellerin anomali tespitindeki etkinliğini ölçmek için kullanılmıştır (Davis & Goadrich, 2006).

SVM modellerinin anomali tespit performanslarını daha net karşılaştırabilmek için ROC eğrisi ve AUC değerleri hesaplanmıştır. Aşağıda farklı optimizasyon yöntemleri için ROC eğrisi sunulmaktadır.



Şekil 1. Farklı optimizasyon yöntemleriyle eğitilen SVM modellerinin ROC eğrileri.

5.2. Deneysel Sonuçlar

Deneysel sonuçlar, büyük veri setlerinde SVM'nin performansını artırmak için önerilen yaklaşımların etkinliğini göstermektedir. Aşağıda, farklı veri setleri üzerinde elde edilen sonuçlar ve bu sonuçların analizi yer almaktadır. Önerilen optimizasyon tekniklerinin etkinliği, NSL-KDD ve CIC-IDS2018 gibi güncel veri setleri üzerinde yapılan karşılaştırmalı çalışmalarla da desteklenmiştir (Gökdemir & Çalhan, 2022).

KDD Cup 1999 veri seti üzerinde yapılan deneylerde, LIBSVM ve Pegasos algoritmaları kullanılarak çeşitli çekirdek fonksiyonlarının performansı değerlendirilmiştir. RBF çekirdeği, diğer çekirdek fonksiyonlarına kıyasla daha yüksek doğruluk ve F1 skoru sağlamıştır. Ayrıca, parametre optimizasyonu ile SVM'nin performansı önemli ölçüde artırılmıştır. Grid Search yöntemi, optimal parametreleri belirlemede etkili olmuştur (Tsang, Kwok & Cheung, 2005).

NSL-KDD veri seti üzerinde yapılan deneylerde, veri alt kümeleme ve yaklaşık tekniklerin SVM'nin performansına etkisi incelenmiştir. K-means kümeleme algoritması kullanılarak veri seti alt kümelere ayrılmış ve her alt küme üzerinde SVM eğitilmiştir. Bu yaklaşım, hesaplama süresini önemli ölçüde azaltmış ve bellek kullanımını optimize etmiştir. Yaklaşık teknikler de benzer şekilde performansı artırmıştır (Xu & Wunsch, 2005).

CICIDS2017 veri seti üzerinde yapılan deneylerde, veri dengesizliği ile başa çıkmak için SMOTE yöntemi kullanılmıştır. SMOTE, azınlık

sınıfına ait yeni sentetik örnekler oluşturarak veri dengesizliğini azaltmış ve modelin anomali tespit performansını artırmıştır. SVM, bu veri seti üzerinde yüksek kesinlik ve geri çağırma değerleri elde etmiştir (Chawla et al., 2002).

Yahoo S5 zaman serisi veri seti üzerinde yapılan deneylerde, SVM'nin zaman serisi anomali tespitindeki performansı değerlendirilmiştir. RBF çekirdeği, zaman serisi verilerinde doğrusal olmayan ilişkileri modellemede etkili olmuştur. Ayrıca, Pegasos algoritması kullanılarak büyük veri setlerinde hızlı eğitim sağlanmıştır. Deney sonuçları, SVM'nin zaman serisi anomali tespitinde yüksek doğruluk ve AUC-ROC değerleri elde ettiğini göstermiştir (Shalev-Shwartz et al., 2011).

6. SONUÇ

Bu makalede, Destek Vektör Makineleri (SVM) ile büyük ölçekli verilerde hassas anomali tespiti konusundaki zorluklar ve bu zorlukların üstesinden gelmek için geliştirilen yeni yaklaşımlar ve optimizasyon teknikleri incelenmiştir. Deneysel sonuçlar, SVM'nin büyük veri setlerinde anomali tespitinde yüksek doğruluk ve genelleme yeteneği sağlayabileceğini göstermiştir. Bununla birlikte, hesaplama maliyeti, bellek kullanımı ve veri dengesizliği gibi zorlukların, optimize edilmiş yöntemler ve yeni teknolojiler kullanılarak aşılması gerektiği vurgulanmıştır. SVM, anomali tespiti için güçlü ve esnek bir araç olmasına rağmen, büyük ölçekli veri setlerinde uygulanması çeşitli zorlukları beraberinde getirir. Bu zorluklar arasında yüksek hesaplama maliyetleri, yoğun bellek kullanımı, sınıf dengesizliği ve modelin genelleme yeteneği gibi sorunlar bulunmaktadır. Ancak, literatürde önerilen çeşitli yaklaşımlar ve optimizasyon teknikleri, bu zorlukların üstesinden gelmek için etkili çözümler sunmaktadır.

Çekirdek triklerinin kullanımı, veri setlerinin yüksek boyutlu uzaylara dönüştürülerek doğrusal olmayan ilişkilerin modellenmesine olanak tanır. RBF, Polynomial ve Sigmoid gibi farklı çekirdek fonksiyonları, SVM'nin performansını artırmada önemli rol oynar. Parametre optimizasyonu, Grid Search, Random Search ve evrimsel algoritmalar kullanılarak gerçekleştirilmiş ve SVM'nin doğruluk ve genelleme yeteneği önemli ölçüde artırılmıştır.

Veri alt kümeleme ve yaklaşık teknikler, hesaplama maliyetlerini ve bellek kullanımını azaltmak için etkili yöntemler sunar. K-means gibi kümeleme algoritmaları ve Core Vector Machines (CVM) gibi yöntemler, büyük veri setlerinde SVM'nin eğitim sürecini hızlandırmıştır. Ayrıca, hızlı ve verimli SVM algoritmaları, Pegasos ve LIBSVM gibi, büyük veri setlerinde daha hızlı ve verimli sonuçlar elde edilmesini sağlamıştır.

Veri dengesizliği problemi, SMOTE gibi yeniden örnekleme teknikleri kullanılarak azaltılmış ve bu sayede modelin doğruluğu artırılmıştır (Chawla et al., 2002). Son olarak, büyük veri teknolojilerinin kullanımı, SVM'nin büyük ölçekli veri setlerinde daha etkili ve verimli bir şekilde uygulanabilmesini sağlamıştır.

Bu çalışma, büyük ölçekli veri setlerinde SVM'nin anomali tespiti performansını artırmak için çeşitli yaklaşımlar ve optimizasyon teknikleri sunmuştur. Ancak, gelecekteki araştırmalar için birkaç önemli alan bulunmaktadır: Derin öğrenme teknikleri, büyük ölçekli verilerde etkili sonuçlar vermektedir. SVM ile derin öğrenme tekniklerinin entegrasyonu, anomali tespitinde yeni ve daha güçlü modeller geliştirebilir (Erfani et al., 2016). Büyük veri setlerinde sürekli değişen verilerle başa çıkmak için çevrimdışı ve çevrimiçi öğrenme yöntemlerinin kombinasyonu önemlidir. Bu tür yöntemler, modelin sürekli olarak güncellenmesini ve gerçek zamanlı anomali tespitini mümkün kılabilir. Büyük veri setlerinde SVM'nin performansını artırmak için dağıtık hesaplama ve paralel işleme teknikleri daha fazla araştırılmalıdır. Hadoop ve Spark gibi platformlar, SVM'nin büyük veri setlerinde daha verimli bir şekilde uygulanabilmesini sağlar. Farklı alanlarda toplanan veriler arasında transfer öğrenme ve domain adaptation teknikleri kullanılarak SVM modellerinin genelleme yeteneği artırılabilir. Bu, modelin farklı veri setlerinde ve uygulama alanlarında daha etkili olmasını sağlar (Pan & Yang, 2010). Büyük veri setlerinde gizlilik ve güvenlik konuları önemlidir. SVM uygulamalarında gizlilik koruma mekanizmalarının ve güvenli hesaplama tekniklerinin geliştirilmesi, bu alanlardaki endişeleri giderebilir (Abadi et al., 2016). SVM'nin hiperparametre optimizasyonu ve model seçimi için AutoML yaklaşımlarının kullanımı, kullanıcı müdahalesini azaltarak model geliştirme sürecini hızlandırabilir (Feurer et al., 2015).

Sonuç olarak, büyük ölçekli veri setlerinde SVM ile hassas anomali tespiti üzerine yapılan araştırmalar, çeşitli optimizasyon teknikleri ve yeni yaklaşımlarla desteklenmektedir. Gelecek çalışmalar, bu alanlardaki yenilikçi çözümler ve teknolojilerle SVM'nin performansını daha da artırmayı hedeflemelidir.

Bu makalenin hazırlanmasında, kaynak taraması, içeriğin gözden geçirilmesi, çeviri vb. süreçlerde yapay zeka yazılımlarından faydalanılmıştır (OpenAI, 2024).

KAYNAKÇA

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Benoit Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., & Zhang, X. (2016). TensorFlow: A system for large-scale machine learning. In 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16) (pp. 265-283).

Akkuş, Ö., & Demir, E. (2016). İki düzeyli olasılık modellerinde klasik meta sezgisel optimizasyon tekniklerinin performansı üzerine bir çalışma.

İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, 15(30), 107-131.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13(Feb), 281-305.

Bottou, L., & Lin, C. J. (2007). Support vector machine solvers. In Large scale kernel machines (pp. 301-320). MIT Press.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery, 2(2), 121-167.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM Computing Surveys (CSUR), 41(3), 1-58.

Chang, C. C., & Lin, C. J. (2011). LIBSVM: A library for support vector machines. ACM transactions on intelligent systems and technology (TIST), 2(3), 1-27.

Chatterjee, A., & Siarry, P. (2006). Nonlinear inertia weight variation for dynamic adaptation in particle swarm optimization. Computers & Operations Research, 33(3), 859-871.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.

Çelik, Y., & Alaca, Y. (2021). Log analizinde derin öğrenme ile anomali tespiti. Yapay Zeka Uygulamalarında Güncel Konular ve Araştırmalar, 137-153.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.

Erfani, S. M., Rajasegarar, S., Karunasekera, S., & Leckie, C. (2016). High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning. Pattern Recognition, 58, 121-134.

Feurer, M., Klein, A., Eggensperger, K., Springenberg, J. T., Blum, M., & Hutter, F. (2015). Efficient and robust automated machine learning. Neural information processing systems (pp. 2962-2970).

Gökdemir, A., & Çalhan, A. (2022). Deep learning and machine learning based anomaly detection in IoT. Gazi Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi, 37(4), 1945-1956.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction. Springer Science & Business Media.

Joachims, T. (2006). Training linear SVMs in linear time. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 217-226).

Köse, U. (2019). Zeki Optimizasyon Tabanlı Destek Vektör Makineleri ile Diyabet Teşhisi. Politeknik Dergisi, 22(3), 557-566.

Laskov, P., Düssel, P., Schäfer, C., & Rieck, K. (2005). Learning intrusion detection: supervised or unsupervised?. In International Conference on Image Analysis and Processing (pp. 50-57). Springer, Berlin, Heidelberg.

OpenAI. (2024). ChatGPT (Sürüm 4o) [Yazılım]. https://openai.com

Smola, A. J., Bartlett, P., Schölkopf, B., Schuurmans, D. (2000). Generalized support vector machines. In Advances in large margin classifiers (pp. 135-146). MIT Press.

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. IEEE Transactions on knowledge and data engineering, 22(10), 1345-1359.

Schölkopf, B., & Smola, A. J. (2001). Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., & Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. Neural computation, 13(7), 1443-1471.

Shalev-Shwartz, S., Singer, Y., Srebro, N., & Cotter, A. (2011). Pegasos: Primal estimated sub-gradient solver for SVM. Mathematical programming, 127(1), 3-30.

Shawe-Taylor, J., & Cristianini, N. (2004). Kernel methods for pattern analysis. Cambridge university press.

Tax, D. M. J., & Duin, R. P. W. (2004). Support vector data description. Machine learning, 54(1), 45-66.

Tsang, I. W., Kwok, J. T., & Cheung, P. M. (2005). Core vector machines: Fast SVM training on very large data sets. Journal of Machine Learning Research, 6(Apr), 363-392.

Vapnik, V. N. (1998). Statistical learning theory. Wiley.

Weston, J., Perez-Cruz, F., Bousquet, O., Chapelle, O., Elisseeff, A., & Schölkopf, B. (2003). Feature selection and transduction for prediction of molecular bioactivity for drug design. Bioinformatics, 19(6), 764-771.

Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. IEEE Transactions on neural networks, 16(3), 645-678.

Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauley, M., Franklin, M. J., Shenker, S., & Stoica, I. (2012). Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation (pp. 15-28).