All Sciences Proceedings
http://as-proceeding.com/

2nd International Conference on Engineering, Natural and Social Sciences
April 4-6, 2023 : Konya, Turkey

https://www.icensos.com/

YAPAY BAĞIŞIKLIK SİSTEMİ TABANLI YENİ BİR AZ ÖRNEKLEME YÖNTEMİ

Kübranur Gümüşlü[*1], Ayşe Merve ACILAR[2]

[1]Bilgisayar Mühendisliği / Fen Bilimleri Enstitüsü, Necmettin Erbakan Üniversitesi, Türkiye
[2]Bilgisayar Mühendisliği / Mühendislik Fakültesi, Necmettin Erbakan Üniversitesi, Türkiye

[*](gumuslukubranur@gmail.com)

Özet **–** Verilerden bilgi çıkarımı ve bu çıkarımlar baz alınarak yapılan çalışmalar gün geçtikçe yaygınlaşmaktadır. Veri setleri üzerinde öğrenme, sınıflandırma, kümeleme gibi işlemler uygulanması ve sağlıklı modellerin oluşturulması büyük önem taşımaktadır. Günümüzde kullanılan gerçek dünya verilerindeki en önemli sorunlardan biri dengesiz veri setleridir. Dengesiz veri seti, örneklem içerisindeki sınıf dağılımları arasındaki dengesizlikten doğar. Bir veri setinde sınıf etiketlerinin dağılımının birbirine yakın değerler olmaması durumunda bu veri seti dengesiz bir veri seti (Imbalanced Dataset) olarak kabul edilir. Bu çalışmada, veri kümesindeki bu dengesizliği gidermek için literatürde önerilen çözümlerden çoğunluk sınıfına uygulanan Az Örnekleme (Undersampling) tekniği baz alınmıştır. Az örnekleme işlemi için yapay bağışıklık algoritmalarından aiNet algoritmasının bu konuda yeni bir yöntem olacağı öngörülmüştür. Yapay bağışıklık sistemi; vücuda giren antijenler, vücutta üretilen antikorlar ve bağışıklık kazanıldıktan sonra oluşturulan hafıza hücreleri gibi temel özellikleri baz alınarak tasarlanmıştır. Veri setindeki çoğunluk sınıfına aiNet algoritması uygulandığında elde edilen hafıza matrisi, veri setindeki çoğunluk sınıfının yapısal organizasyonu temsil etmiştir. Önerilen yöntem tiroit veri seti üzerine uygulanmıştır. KNN sınıflandırma algoritması hem orijinal veri setine hem de azaltılmış veri setine uygulanmıştır. Başarı ölçütü olarak doğruluk, kesinlik, duyarlılık ve F1 skoru değerleri her iki durum içinde ayrı ayrı hesaplanmıştır. Sonuçlar göstermiştir ki, aiNet algoritması veriyi temsil yeteneğini kaybetmeden başarılı bir şekilde indirgemiştir.

Anahtar Kelimeler – Ainet Algoritması, Az Örnekleme, Dengesiz Veri Seti, Sınıflandırma, Yapay Bağışıklık Sistemi

I. GİRİŞ

Teknoloji ilerledikçe verilerden bilgi çıkarımı ve bu çıkarımlar baz alınarak yapılan çalışmalar oldukça yaygınlaşmaktadır. Alan fark etmeksizin çalışmalarda kullanılan veri setleri üzerinde öğrenme, sınıflandırma, kümeleme gibi işlemler uygulanması ve sağlıklı modellerin oluşturulması büyük önem taşımaktadır. Veri setleri üzerinde uygulanan model seçimi kadar kullanılan veri setinin yapısı da sonucu etkilemektedir. Günümüzde kullanılan gerçek dünya verilerindeki en önemli sorunlardan biri dengesiz veri setleridir. Dengesiz veri seti, örneklem içerisindeki sınıf dağılımları arasındaki dengesizlikten doğar. Bir veri setinde iki sınıf olduğu varsayımında, sınıf etiketlerinin dağılımının birbirine yakın değerler olmaması durumunda (%90-%10 veya %80-%20 gibi) bu veri setinin dengesiz bir veri seti (Imbalanced Dataset) olduğu kabul edilir. Literatürde (Imbalanced Ratio) veri dengesizliği oranı farklı olan veriler ile ilgili çalışmalar bulunmaktadır. IR, çoğunluk sınıf sayısının azınlık sınıf sayısına oranı olarak tanımlanır [1].

Dengesiz sınıflandırma problemini çözmek için Veri Düzeyinde ve Algoritmik düzeyde olmak üzere iki ana çözüm türü vardır [1]. Algoritmik düzeyde yöntemlerde genellikle yeni sınıflandırma algoritma tasarımı kullanılır veya dengesiz veriler tarafından üretilen önyargıların üstesinden gelmek için mevcut algoritmaların geliştirilmesi ele alınır [2].

Algoritmik düzeydeki çözümlere örnek olarak sınıf başına düşen maliyet değişimi, karar ağacında olasılık tahminlerini ayarlamak verilebilir. Veri düzeyindeki çözümler ise azınlık sınıfına uygulanan Aşırı Örnekleme (Oversampling) ve çoğunluk sınıfına uygulanan Az Örnekleme (Undersampling) tekniklerdir. Aşırı örnekleme yöntemleri, dengesiz sınıflardan az olanını artırarak veri setini dengelerken; Az örnekleme yöntemleri çoğunlukta olan sınıf verilerini azaltarak denge sağlamayı hedefler (Şekil-1). Aşırı ve Az Örnekleme yöntemlerinin bir arada kullanıldığı Hibrit metotlar da literatürde yer almaktadır [3]. Daha fazla veri toplama, verileri yeniden örnekleme veya birleştirme yöntemi kullanma gibi yöntemler de dengesiz veri setlerini dengelemek için kullanılan yaklaşımlardandır.

Şekil 1: Az Örnekleme ve Aşırı Örnekleme

Son dönemlerdeki literatür incelendiğinde dengesiz veri seti probleminin oldukça popüler bir araştırma konusu olduğu gözlemlenmiştir. Kimya, biyomedikal mühendisliği, finansal yönetim, güvenlik yönetimi vb. alanlarda yaygın olarak kullanılan çeşitli veri ön işleme teknikleri, sınıflandırma algoritmaları, model değerlendirme yöntemleri ve dengesiz sınıflandırma problemlerinin teknikleri, uygulamaları ile ilgili olarak 527 makale bir incelemiştir. Sonuçlarını maddeler halinde özetlemiş her çalışma alanının ve veri kümesinin özel yöntemlerle değerlendirilmesi gerektiği belirtilmiştir [9]. Diğer bir çalışmada veri setlerinde bulunan dengesiz örneklerin veri özelliklerinde adaptif bir metot kullanılmıştır. Adaptif Çoklu Sınıflayıcı (Adaptive Multiple Classifier System- AMCS) ile çeşitli örnekler için başarılı sonuçlar elde edilmiştir [10]. Dengesiz sınıflandırmada çoğunluk sınıfındaki gürültü ve aykırı örneklerin azaltılması için BNF, OBN ve DBSCAN algoritmaları ile hibrit yöntem sunulmuştur. Çalışma sonucunda gürültü kaldırma ve az örneklem yöntemi olarak RUS kullanılmıştır [6]. Sağlam ve ark. veri setlerindeki dengesizliği ortadan kaldırmak için kullanılan standart yöntem RUS ile sınıflandırma modeli oluştururken genetik algoritma, yapay arı kolonisi, parçacık sürü optimizasyonu algoritmalarının kullanılmasını karşılaştırılan bir çalışma yapmıştır. Az örneklem sonucunda sınıflandırma modelinin yapay arı kolonisi ile oluşturulduğunda daha iyi performans gösterdiği görülmüştür [11]. Xiaoying Xie ve arkadaşları tarafından az örnekleme prensibinde yaygın karşılaşılan, veri kümesinden kaç örneklem ve hangi örnekler çıkarılmalı sorununa çözüm bulabilmek için PUMD (Progressive Undersampling Method with Density) yöntemi önerilmiştir. Önerilen yöntem, diğer standart yöntemlerle 40 dataset üzerinde çalıştırılarak karşılaştırılmış ve başarılı sonuçlar vermiştir [12].

Yapay bağışıklık sistemleri, canlılardaki bağışıklık sisteminden esinlenerek oluşturulmuş biyolojik tabanlı algoritmalardır. Sistem; vücuda giren antijenler, vücutta üretilen antikorlar ve bağışıklık kazanıldıktan sonra oluşturulan hafıza hücreleri gibi temel özellikleri baz alınarak tasarlanmıştır. Bir yapay bağışıklık algoritması olan aiNet öğrenme algoritmasının amacı, veriyi tanıyan ve onun yapısal organizasyonunu temsil eden bir hafıza seti oluşturmaktır. Elde edilen bu hafıza setinden verinin özünü temsil eden herhangi bir grup veya alt grup var mı? Varsa kaç tane var? Bu verinin (grupların) yapısı veya görüntü dağılımı nasıldır? sorularının cevapları öğrenilebilir [4].

Bu çalışmada, tiroit veri setindeki çoğunluk sınıfına aiNet algoritması uygulanmış ve elde edilen hafıza matrisi, veri setindeki çoğunluk sınıfının yapısal organizasyonu temsil edebildiği görülmüştür.

II. MATERYAL VE YÖNTEM

Dengesiz sınıflandırma problemi sınıflar arasındaki örneklem sayılarının farkından doğan bir problemdir. Sınıflar gözlem sayısına göre azınlık ve çoğunluk sınıfı olarak adlandırılmıştır [5]. Sınıf dengesizlik oranı (IR) denklem-1 kullanılarak hesaplanır. IR değerinin bire yaklaşık olması beklenir.

$$IR = \frac{|Çoğunluk\_sınıfı|}{|Azınlık\_Sınıfı|} \quad (1)$$

Çoğunluk sınıfının sayısının azaltılarak azınlık sınıfına yaklaştırılması işlemine en genel ifadesiyle az örneklem (Undersampling) denilmektedir ve IR değerini düşürmesi sağlanır.

aiNet, Castro ve Von Zuben (2001) tarafından sunulmuş veri analizi, tanıma, sınıflandırma için kullanılabilen ağ tabanlı bir yapay bağışıklık algoritmasıdır. aiNet modeli antikor adı verilen ve birbirleri ile bağlantılı hücrelerden meydana gelir.

Bu antikorların sisteme girebilecek antijenlerin dâhili bir görüntüsünü oluşturduğu varsayılır. aiNet öğrenme algoritmasının amacı, veriyi tanıyan ve onun yapısal organizasyonunu temsil eden bir hafıza seti oluşturmaktır.

A. *Önerilen Yöntem*

Bu çalışmada, aiNet algoritmasına giriş olarak verilen çoğunluk sınıfı yerine, algoritma çıkışında elde edilen hafıza matrisinin kullanılması önerilmiştir. Bu yönteme AiNUS (aiNet under sampling) ismi verilmiştir. Az örnekleme yönteminin çoğunluk sınıfının azaltılması olarak genellendiği varsayılırsa, çalışmada bu azaltma yöntemi önerilen aiNUS ile gerçekleştirilmiştir. AiNUS ile azaltılan sınıf, azınlık sınıfı ile birleştirilerek sınıflandırıcıya verilir. Bu çalışmada sınıflandırıcı olarak (K-en yakın komşu) sınıflandırıcısı seçilmiştir. Model hem orijinal eğitim seti hem de az örnekleme yapılmış eğitim seti ile eğitilerek oluşturulmuştur. Oluşturulan iki model test sınıfı ile test edilmiştir.

B. *Kullanılan Veri seti*

Keel.es [8] sitesindeki dengesizlik oranı 1,5-9 arası olan veri setlerinden "new-thyroid" seçilmiştir. Seçilen veri setinin dengesizlik oranı IR=5.14'dir. Veri setinin 5 öznitelik (T3resin, Thyroxin, Triiodothyronine, Thyroidstimulating, TSH_value) ve sınıf (pozitif, negatif) etiketi bulunan 215 örneklemden oluşmaktadır. Eğitim seti(172) ve test seti(43) olarak ayrılmıştır. Deneysel çalışmalarda 5 kat çapraz doğrulama (5-fold cv) yöntemi kullanılmıştır.

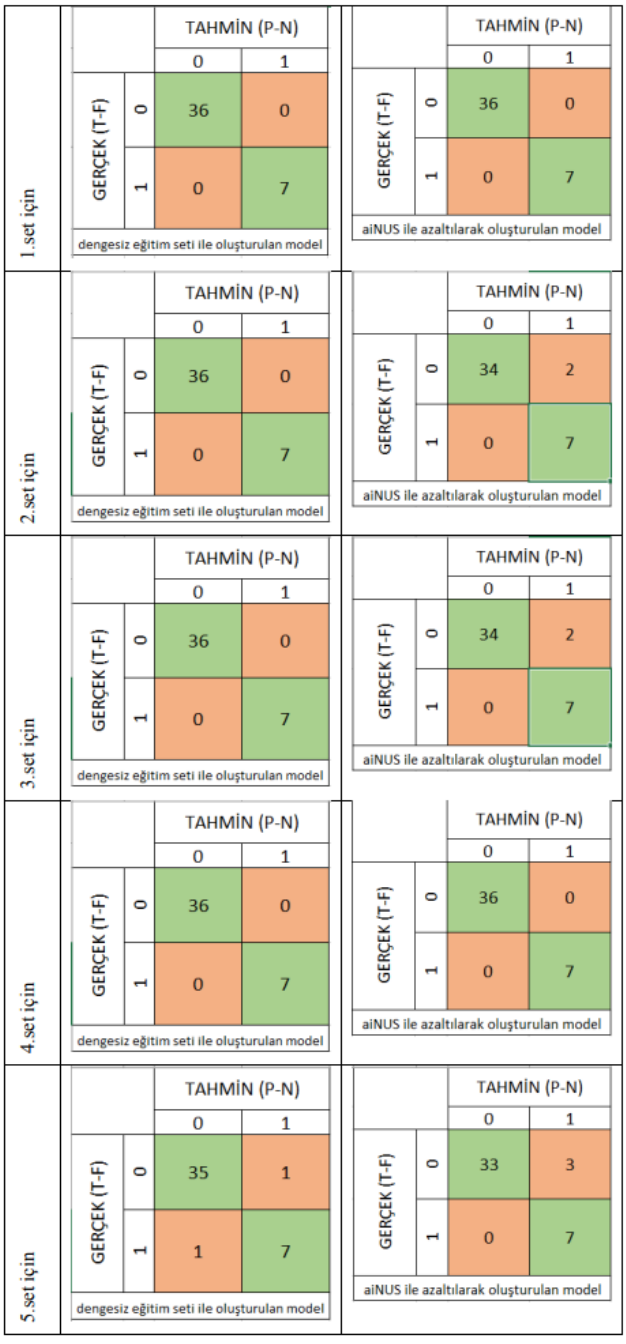
III. BULGULAR

Önerilen yöntem aiNUS ile yeniden örneklendirme yapılmış, dengesiz veri kümesi daha dengeli hale getirilmiştir. Dengesizlik oranı 5.14 olan new_thyroid1 veri kümesinin çoğunluk sınıfı aiNet algoritması ile azaltılmıştır. Azaltılmış örneklemin boyutu diğer bir değişle elde edilen hafıza matrisinin boyutu aiNet algoritmanın bir hiper parametresi olan baskılama eşik (supression threshold - *ts*) değerlerine göre değişmektedir. Yapılan deneysel çalışmalarda *ts* 0.02 ile 0.7 arasında değerler ile çalıştırılmış ve en iyi *IR* değeri *ts*=0.1 iken elde edilmiştir (Tablo-1).

Tablo1. ts değerlerine göre IR_{aiNus} değerleri

Baskılama Eşiği (ts) değerleri	**aiNus ile Azaltılan sınıftaki veri sayısı = $\lvert Azaltılmış\_Sınıf \rvert$**	$IR_{aiNus} = \frac{\lvert Azaltılmış\_sınıf \rvert}{\lvert Azınlık\_Sınıfı \rvert}$
0.02	113	4.04
0.07	44	1.57
0.1	30	1.07
0.2	11	0.39
0.4	5	0.18
0.7	2	0.07

*Azınlık sınıfındaki veri sayısı = $\lvert Azınlık_{sınıfı} \rvert = 28$,
Çoğunluk sınıfındaki veri sayısı = $\lvert Çoğunluk_{sınıfı} \rvert$=144,
$IR_{orjinal} = 144 / 28 = 5.14$



Şekil 2. 5-fold için konfüzyon Matrisi (Test Kümeleri için)

5-kat çapraz doğrulama şeklinde kullanılan veri seti için tüm işlemler her eğitim ve test seti için ayrı ayrı

çalıştırılmıştır. KNN sınıflandırıcı model hem orijinal eğitim seti için hem de aiNUS ile azaltılmış veri olarak oluşturulmuştur. Tablo 2'de sınıflandırma modelinin performans hesaplamaları verilmiştir. Dengesiz sınıflandırma problemlerinde Doğruluk metriği tek başına yeterli olmadığı için Duyarlılık, Kesinlik ve F1 Skoru da ölçümlere eklenmiştir. Şekil 2'de sınıflandırma modellerinin test edilmesi sonrasındaki konfüzyon matrisleri verilmiştir.

Tablo 2. Sınıflandırıcı modeller için Doğruluk, Duyarlılık, Kesinlik, F1 Skoru

	Orijinal Eğitim Seti ile oluşturulan Model için Performans Hesapları				aiNet ile azaltılmış eğitim seti ile oluşturulan model için performans hesapları			
	Doğruluk	Duyarlılık	Kesinlik	F1 Skoru	Doğruluk	Duyarlılık	Kesinlik	F1 Skoru
Train1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Test1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Train2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Test2	1.00	1.00	1.00	1.00	0.95	1.00	0.77	0.87
Train3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Test3	1.00	1.00	1.00	1.00	0.95	1.00	0.77	0.87
Train4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Test4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Train5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Test5	0.95	0.85	0.85	0.90	0.93	1.00	0.7	0.82

IV. TARTIŞMA

Çalışmada dengesizlik oranı 5.14 olan "new-thyroid1" veri seti kullanılmıştır. Amacımız aiNet ile indirgenen çoğunluk sınıfının temsil yeteneğini koruyup korumadığının araştırılmasıdır. Diğer bir deyişle azaltılmış sınıf, veri kaybı olmasına rağmen çoğunluk sınıfını temsil etmeyi sürdürebilmiş midir? Bu sebeple her iki veri seti de Knn (K-en yakın komşu) algoritması ile sınıflandırılmış ve bulgular Tablo2'de sunulmuştur. Bu bulgulara ait konfüzyon matrisleri şekil2'de verilmiştir. Sonuçlar incelendiğinde, nerdeyse aynı başarıyı yakaladığı gözlemlenmektedir.

V. SONUÇLAR

Konusu fark etmeksizin verilerden bilgi çıkarımı ve bu çıkarımlar baz alınarak yapılan çalışmalarda kullanılan veri setleri üzerinde öğrenme, sınıflandırma, kümeleme gibi işlemler uygulanmakta ve modellerin sağlıklı oluşturulması büyük önem taşımaktadır. Bu modellerde kullanılan gerçek dünya verilerindeki önemli sorunlardan biri dengesiz veri setleridir. Veri kümesindeki dengesizliğin ortadan kaldırılması bu çalışmada veri düzeyindeki çözümlerden, çoğunluk sınıfına uygulanan az örnekleme (undersampling) tekniği temel alınmıştır. Çoğunluk sınıfın örneklem sayısının indirgenmesi için yapay bağışıklık sistemi algoritması olan aiNet algoritması seçilmiştir. Bu algoritma sonucunda elde edilen hafıza matrisi, veri setindeki çoğunluk sınıfının yapısal organizasyonu temsil etmiştir.

Dengesiz sınıflandırma problemine örnek çalışma için, Keel.es sitesinde Imbalanced Dataset kategorisinde hazır olarak bulunan dengesizlik oranı (IR) 5.14 olan newthyroid1 veri seti seçilmiştir. Sınıflandırma algoritması olarak ise KNN (K-en yakın komşu) seçilmiştir. Dengesiz sınıflandırma problemlerinde sınıflandırıcı modellerin performans ölçütlerinden Doğruluk metriğinin tek başına yeterli olmadığı literatürde vurgulandığı için, Kesinlik, Duyarlılık ve F1 Skoru da hesaplanarak başarı ölçümü yapılmıştır. Sonuç olarak, aiNus ile indirgenen veri setinin, orijinal veri seti ile neredeyse aynı temsil başarısını gösterdiği görülmüştür.

KAYNAKLAR

[1] Fernández, A., García, S., del Jesus, M. J., & Herrera, F. (2008). A study of the behaviour of linguistic fuzzy rule based classification systems in the framework of imbalanced data-sets. Fuzzy Sets and Systems, 159(18), 2378–2398.

[2] Spelmen, V. S., & Porkodi, R. (2018). A Review on Handling Imbalanced Data. Proceedings of the 2018 International Conference on Current Trends towards Converging Technologies, ICCTCT 2018, 1–11.

[3] Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2012). A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews, 42(4), 463–484.

[4] Castro, L. N. de, & Zuben, F. J. Von. (2001). aiNet: An Artificial Immune Network for Data Analysis. http://www.dca.fee.unicamp.br/~lnunesftp://ftp.dca.fee.unicamp.br/pub/docs/vonzuben/lnunes/DMHA.pdf

[5] He, H., & Garcia, E. (2009). Learning from imbalanced data. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 21(9), 1263–1284.

[6] Peng, C. Y., & Park, Y. J. (2022). A New Hybrid Under-sampling Approach to Imbalanced Classification Problems. In Applied Artificial Intelligence (Vol. 36, Issue 1).

[7] ie, X., Liu, H., Zeng, S., Lin, L., & Li, W. (2021). A novel progressively undersampling method based on the density peaks sequence for imbalanced data. Knowledge-Based Systems, 213, 106689.

[8] http://www.keel.es/

[9] Öztürk, Ş., & Özkaya, U. (2020). Skin lesion segmentation with improved convolutional neural network. Journal of digital imaging, 33, 958-970.

[10] Yijing, L., Haixiang, G., Xiao, L., Yanan, L., & Jinling, L. (2016). Adapted ensemble classification algorithm based on multiple classifier system and feature selection

for classifying multi-class imbalanced data. Knowledge-Based Systems, 94, 88–104. https://doi.org/10.1016/J.KNOSYS.2015.11.013

[11] Sağlam, F. (2021). Optimization Based Undersampling for Imbalanced Classes. Adıyaman University Journal of Science, 385–409. https://doi.org/10.37094/adyujsci.884120

[12] Xie, X., Liu, H., Zeng, S., Lin, L., & Li, W. (2021). A novel progressively undersampling method based on the density peaks sequence for imbalanced data. Knowledge-Based Systems, 213, 106689. https://doi.org/10.1016/j.knosys.2020.106689