ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH

DOI: 10.5336/biostatic.2019-66816

Sınıf Dengesizliği Varlığında Hastalık Tanısı için Kolektif Öğrenme Yöntemlerinin Karşılaştırılması: Diyabet Tanısı Örneği

Comparison of Ensemble Learning Methods for Disease Diagnosis in Presence of Class Unbalanced: Case of Diabetes

Sultan TURHAN[a], Yüksel ÖZKAN[b], B. Sarer YÜREKLİ[c], Aslı SUNER[b], Eralp DOĞU[a]

[a]Muğla Sıtkı Koçman Üniversitesi Fen Fakültesi, İstatistik Bölümü, Muğla, TÜRKİYE
[b]Ege Üniversitesi Tıp Fakültesi, Biyoistatistik ve Tıbbi Bilişim AD, İzmir, TÜRKİYE
[c]Ege Üniversitesi Tıp Fakültesi, Endokrinoloji BD, İzmir, TÜRKİYE

ÖZET Amaç: Günümüzde makine öğrenmesi yöntemleri hastalık tanısının konulmasında yaygın olarak kullanılmaktadır. Ancak sağlık verisinin büyük hacimli, çok boyutlu ve karmaşık olması nedeniyle dengesiz sınıf problemi ile karşılaşılması durumunda bu yöntemlerin doğrudan kullanımı performans düşüşüne neden olmaktadır. Bu çalışmada diyabet hastalarına ilişkin dengesiz yapıdaki bir veri seti kullanılarak çeşitli yeniden örnekleme yöntemleri dengesizlik probleminin giderilmesinde kullanılmış ve kolektif (ensemble) öğrenme algoritmalarına entegre edilerek diyabet tanısı üzerinden sınıflandırma performansları karşılaştırılmıştır. **Gereç Yöntemler:** Kullanılan veriler Haziran – Eylül 2013 tarihleri arasında, İzmir Bozkaya Eğitim ve Araştırma Hastanesi, Endokrinoloji ve Metabolizma Hastalıkları polikliniğine başvuran, 18 yaşından büyük 185 hastadan elde edilmiştir. Diyabet tanısının sınıflandırmasına yönelik sınıf dengesizliği problemini gidermek amacıyla alt örnekleme (under sampling), aşırı örnekleme (over sampling) ve sentetik azınlık aşırı örnekleme (SMOTE) yöntemleri kullanılmıştır. Sınıflandırma performansı üzerindeki etkiler, torbalama (bagging) ve arttırma (boosting) temelli kolektif öğrenme yöntemlerine entegre edilmesiyle karşılaştırılmıştır. Algoritmaların doğru sınıflandırma performanslarının karşılaştırılmasında doğruluk, Kappa istatistiği, duyarlılık ve seçicilik ölçütleri kullanılmıştır. Tüm istatistiksel analizler, açık kaynak kodlu bir yazılım olan R programlama dilinde yapılmıştır. **Bulgular:** Dengesiz veri setinde ham veri ile yapılan diyabet tanısı sınıflandırma başarısı oldukça düşüktür. Aşırı örnekleme yöntemi ile yapılan sınıflandırmaların, orijinal dengesiz veri seti, alt örnekleme ve sentetik azınlık aşırı örnekleme yöntemi ile yapılan sınıflandırmalardan çok daha başarılı tahmin gücüne sahip olduğu tespit edilmiştir. **Sonuç:** Sınıf dengesizliği varlığında veri setlerini yeniden örnekleme yöntemlerine tabi tutarak veriyi dengeledikten sonra sınıflandırma algoritmalarının kullanılması önerilmektedir.

Anahtar Kelimeler: Kolektif öğrenme; sınıflandırma; dengesiz veri; hastalık tanısı; diyabet

ABSTRACT Objective: Recently, machine learning methods have been widely used in disease diognosis. However, due to the large volume, multidimensional and complexity of the information, an unbalanced data problem arises. In this study, it is aimed to eliminate problem of imbalance by using re-sampling methods in an unbalanced data set related to diabetes patients, to classify diagnosis of diabetes with ensemble learning algorithms and to compare correct classification performances of algorithms. **Material and Methods:** The data were collected from 185 patients older than 18 years of age who were admitted to Izmir Bozkaya Training and Research Hospital, Endocrinology and Metabolism Diseases outpatient clinic between June and September 2013. Under-sampling, over-sampling and synthetic minority over-sampling methods were used to eliminate unbalanced class problem for diagnosis of diabetes. The effects on classification performance were compared by integrating bagging and boosting methods into ensemble learning methods. Accuracy, Kappa statistics, sensitivity and specificity were used to compare correct classification performance of algorithms. All statistical analyzes were made in the R programming language, an open source software. **Results:** The success rate of diabetes diagnosis with raw data is very low in the unbalanced data set. It is determined that classifications made with over-sampling method have much more successful estimation power than classifications made with original unbalanced data set, under-sampling and synthetic minority over-sampling method. **Conclusion:** It is recommended to use classification algorithms after balancing the data by subjecting the data sets to resampling methods in the presence of class imbalance.

Keywords: Ensemble learning; classification; unbalanced data; disease diagnosis; diabetes

Correspondence: Eralp DOĞU
Muğla Sıtkı Koçman Üniversitesi Fen Fakültesi, İstatistik Bölümü, Muğla, TÜRKİYE/TURKEY
E-mail: eralp.dogu@mu.edu.tr

Peer review under responsibility of Turkiye Klinikleri Journal of Biostatistics.

Received: 19 Aug 2019 ***Received in revised form:*** 24 Oct 2019 ***Accepted:*** 19 Nov 2019 ***Available online:*** 10 Dec 2019

Dünyada görülme sıklığı giderek artan diyabet (şeker hastalığı), her yaş grubundaki insan sağlığını ciddi anlamda tehdit eden kronik bir hastalıktır.[1] Diyabet, pankreastan salgılanan insülin hormonunun etkinliğinin azalması sonucu kandaki şeker miktarının artması ile ortaya çıkmaktadır. Gizli şeker olarak da adlandırılan prediyabet ise, kan şekeri düzeyinin normal değerden yüksek olmasına karşı, diyabet tanısı koyacak kadar yeterli yükseklikte olmamasıdır.[1,2] Diyabet, beraberinde birçok hastalığı getirdiğinden, organ kayıplarına yol açmakta ve hastalığın etkileri hayat boyu sürmektedir. Değişen yaşam koşullarına ve beslenme alışkanlıklarına bağlı olarak diyabet çağın hastalığı haline gelmektedir. Uluslararası Diyabet Federasyonu'nun yayınladığı Diyabet Atlası 2017 yılı tahminlerine göre, Dünya'da her 11 yetişkinden 1'inin diyabet hastası (415 milyon) olduğu ve 2 diyabetli yetişkinden 1'ine (%46,5) ise teşhis konulamadığı şeklindedir. Aynı zamanda, küresel sağlık harcamalarının %12'sinin (673 milyar ABD Doları) diyabet hastalığı için harcandığı bilinmektedir.1 Bunun yanı sıra her 7 doğumdan 1'i gebelik diyabetinden etkilenmektedir. Diyabet hastalarının dörtte üçü (%75) düşük ve orta gelir düzeyindeki ülkelerde yaşamaktadır. Federasyonun 2040 yılı tahminlerine göre, her 10 yetişkinden 1'i diyabet hastası olabileceği (642 milyon) ve diyabet ile ilişkili hastalıkların sağlık harcamaları 802 milyon ABD Dolarını aşabileceği yönündedir.[3]

İnsan sağlığını bu derece tehdit eden diyabetin, erken tanı ve teşhisinin konulmasında kullanılacak sınıflandırma yaklaşımları da önem kazanmaktadır. Sağlık araştırmaları ve karar verme süreçlerine ilişkin uygulamaların sayısının artışına bağlı olarak, bu alanda üretilen bilginin de hacmi hızlı bir şekilde artış göstermektedir. Buna paralel olarak, makine öğrenmenin tıbbi tanıdaki uygulamaları, gelişmiş ve karmaşık veri analizini mümkün kılan algoritmaların, sistemlerin ve metodolojilerin ortaya çıktığına işaret etmektedir.[1] Yakın bir gelecekte, akıllı veri analizinin bir parçası olan makine öğrenmesi tekniklerinin, modern teknoloji tarafından üretilen ve depolanan büyük miktarda bilginin işlenmesinde kullanılacağı yaygın olarak öngörülmektedir.[3] Halen mevcut makine öğrenmesi yöntemleri, biyomedikal alandaki verilere ait tanımlanmamış ilişkilerin ve gizli örüntülerin ortaya çıkarılmasında oldukça önemli bir potansiyele sahiptir. Aynı zamanda, makine öğrenmesi, sağlık uzmanlarının hassas tıp (*precision medicine*) olarak bilinen kişiselleştirilmiş bakım hizmetine geçmelerini sağlayan büyük veri setlerini otomatik olarak anlamlı ve yorumlanabilir hale getirmelerine yardımcı olabilmektedir.[2]

Literatürde medikal tanı ile ilgili yapılan çalışmalardan olan Jutel (2011) çalışmasında, sınıflandırmanın tıbbi tanı pratiğine rehberlik ettiğini ifade etmiştir.[4] Bu çalışmada, sınıflandırmayı anlamanın, tanı sonuçlarını daha iyi anlamanın bir parçası olması gerektiği savunulmuştur.[4] Buna yönelik olarak doğru tanının konulması için sınıf dengesizliği problemiyle baş edilmesine yönelik geliştirilen algoritmalar ile mevcut makine öğrenme algoritmalarının entegre bir şekilde kullanılması gerekmektedir. Yapılan çalışmalardan yola çıkarak sınıf dengesizliği probleminin sağlık alanında sınıflandırma problemlerinde sıklıkla ortaya çıktığı söylenebilir.[5-8] Khalilia (2011) çalışmasında, aşırı derecede dengesiz verilerden kaynaklanan hastalık risklerini tahmin etmeyi amaçlamıştır. Aralarında diyabetin de bulunduğu toplam sekiz kronik hastalığın tanısının sınıflandırılmasında destek vektör makinesi, torbalama, artırma ve rastgele orman algoritmalarının performansları karşılaştırılmıştır.[9] Bu veri setleri aşırı derece sınıf dengesizliğine sahip oldukları için tekrarlı rastgele alt örnekleme (*repeated random sub-sampling*) yöntemi entegre edilerek probleme çözüm getirmişlerdir. Tüm bu çalışmalar ışığında, direkt orijinal veri setine mevcut makine öğrenme algoritmalarını uygulamak yerine öncelikle sınıf dengesizliği problemiyle baş edilmesi sonrasında mevcut makine yöntemlerinin sınıflandırma performanslarının karşılaştırılmasına yönelik bir çalışma yapılması amaçlanmıştır.

Bu çalışmada, diyabet tanısının sınıflandırmasına yönelik sınıf dengesizliği problemini gidermek için alt örnekleme ve aşırı örnekleme yöntemleri kullanılarak sınıflandırma performansı üzerindeki etkiler, dengeli verilerin torbalama ve artırma temelli kolektif öğrenme yöntemlerine (rastgele orman, ağırlıklı alt uzay rastgele orman, eklemeli lojistik regresyon ve stokastik gradyan artırma) entegre edilmesiyle karşılaştırılmıştır.

GEREÇ VE YÖNTEMLER

VERİ SETİ

Kullanılan veri seti Haziran – Eylül 2013 tarihleri arasında, İzmir Bozkaya Eğitim ve Araştırma Hastanesi, Endokrinoloji ve Metabolizma Hastalıkları polikliniğine başvuran, 18 yaşından büyük hastalardan elde edilmiştir. Çalışmanın etik kurul onayı (İzmir Bozyaka Eğitim ve Araştırma Hastanesi Klinik Araştırmalar Etik Kurulu-24.12.2013-Karar no:8) alınmış ve Helsinki bildirgesi temel alınarak çalışmaya başlanmıştır. Veri setinde; 23'ü prediyabet, 111'i diyabet, 51'i ise normal hasta olmak üzere üç ayrı grupta toplam 185 kişi ve 31 değişken mevcuttur ve değişken açıklamaları Tablo 1'de verilmiştir. Veri seti, %75'i eğitim ve %25'i test olacak şekilde ikiye ayrılmıştır.

TABLO 1: Değişkenler ve açıklamaları.

Değişkenler Kısaltmaları	Değişken Açıklamaları
Yas	Hastanın Yaşı
bmp_4	Kemik Morfojenik Proteini
Mgp	Matriks Gla-Protrini
sLOX-1	Reseptör
lipokalin_2	Lipkolin
Vki	Vücut Kitle İndeksi
Abi	Ayak Bileği Kol Basınç İndeksi
Aks	Açlık Kan Şekeri
Ggt	Açlık Plazma Glokuzu
Kre	Kreatinin
Alt	Alanin Transaminaz
Alp	Alkalin Fosfatoz
Mpv	Ortalama Trombosit Hacmi
Rdw	Eritrosit Dağılım Genişliği
Noggin	Nogging
Hb	Hemoglobin
Plt	Trombosit
Tkol	Toplam Kolestrol
Hdl	Yüksek Yoğunluklu Lipoprotein Kolestrol
Ldl	Düşük Yoğunluklu Lipoprotein Kolestrol
Aip	Plazma Aterojenik İndeksi
Crp	C-Reaktif Protein
Fibrinojen	Fibrinojen
Tsh	Trioid Uyarıcı Hormon
fT3	Serbest Triodotironin
fT4	Serbest Tiroksin
Ca	Kalsiyum
Plt	Fosfor
Ürik	Ürik Asit

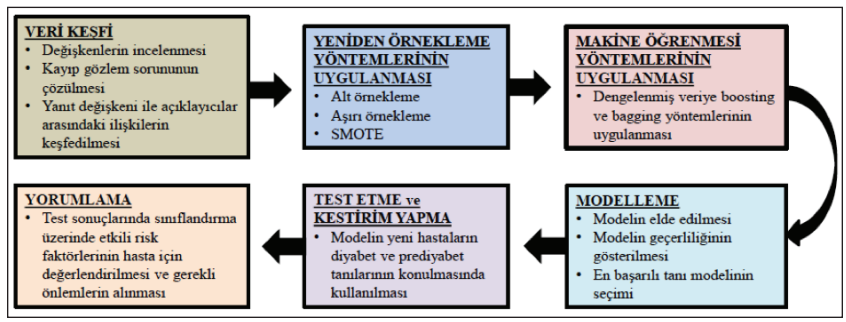
SINIF DENGESİZLİĞİ VARLIĞINDA MAKİNE ÖĞRENMESİ

Sınıf dengesizliği problemi, belirli bir sınıf (örneğin, hastalık sınıfı) diğerlerine kıyasla orantılı olarak daha az temsil edildiğinde ortaya çıkmaktadır. Bu nedenle makine öğrenmesi yaklaşımları kullanılırken, çoğunluğa yönelik bir öğrenme yanlılığı ve yanlış tanı konulması olasıdır. Yanlılık genellikle azınlık sınıfını görmezden gelme eğiliminden kaynaklanan yanlış sınıflandırma problemi olarak tanımlandığından, bu problem tanısal hataya yol açmaktadır.[5,6] Sınıflandırma algoritmalarının sağlıkta sıkça rastlanan bu tür bir probleme sahip verilerle maksimum verimle çalışabilmesi için çeşitli algoritmaların ve/veya yaklaşımların kullanılması gerekmektedir. Böylelikle, hasta güvenliği/kalite geliştirme, karar verme ve problem çözme gibi çeşitli alanlarda çalışan araştırmacılar, tıbbi tanıyı daha güvenilir olarak elde edebilmektedir. Dengesiz veriler söz konusu olduğunda Şekil 1'deki iş akışı takip edilmektedir.

Bu çalışmada veri setine yeniden örnekleme yöntemleri uygulanarak dengesizlik giderilmiş ve sonrasında çeşitli makine öğrenmesi algoritmaları ile diyabet sınıflandırması gerçekleştirilmiştir. Algoritmaların doğru sınıflandırma performanslarının karşılaştırılmasında dengeli doğruluk, Kappa istatistiği, F1 skoru, Matthews korelasyon katsayisi, duyarlılık ve seçicilik ölçütleri kullanılmıştır. Tüm istatistiksel analizler, açık kaynak kodlu bir yazılım olan R'da *caret, lime* ve *ggplot2* paketleri ile yapılmıştır.

YENİDEN ÖRNEKLEME YÖNTEMLERİ (RESAMPLING TECHNIQUES)

Sınıf dengesizliği problemini gidermek için literatürde birçok yöntem bulunmakla birlikte yeniden örnekleme yöntemleri sınıf dengesizliğini gidermek için en yaygın kullanılan teknikler olarak belirtilmiştir. Dengesiz veri setlerinin sınıflandırma üzerindeki etkisinin ortadan kaldırılması için en çok kullanılan yöntemler alt örnekleme (*under sampling*) ve aşırı örnekleme (*over sampling*) yöntemleridir.[10] Alt örnekleme, veri setinde örnek sayısının çok olduğu sınıfı (çoğunluk sınıf) örnek sayısının az olduğu sınıfa (azınlık sınıf) sayıca yaklaştırılmasıdır. Aşırı örnekleme ise, veri setinde örnek sayısının az olduğu sınıfı (azınlık sınıf) örnek sayısının çok olduğu sınıfa (çoğunluk sınıf) sayıca yaklaştırılmasını ifade etmektedir.[8,11] Alt örnekleme ve aşırı örneklemenin farklı avantaj ve dezavantajları bulunmaktadır. Bunlardan en dikkat çekenleri, alt örnekleme yöntemi, veri setindeki örneklem sayısını azaltarak bilgi kaybına yol açarken, aşırı örnekleme yöntemi, örneklem sayısını artırarak aşırı uyum (*overfitting*) problemine sebep olabilmektedir. Diğer yandan alt örnekleme yöntemi, veri setindeki yapılan bu örnek azaltma işlemi eğitim süresini önemli ölçüde azaltmakta ve bellek açısından önemli bir tasarruf sağlamaktadır.[12,13] Aşırı örnekleme yöntemi, veri setinin büyüklüğünü büyük oranda arttırdığı için eğitim süresinde artış gözlenmekte ve kullanılan hafıza oldukça büyük yer tutmaktadır.[8] Diğer yandan, aşırı örnekleme yönteminin çalışma prensibine dayanan SMOTE yöntemi de kullanılmaktadır. Bu yöntem, azınlık sınıfları örneklem sayısını genişletmek için aşırı örnekleme yöntemini temel alarak geliştirilmiştir. SMOTE, azınlık sınıfı örneklemi ile aynı sınıfta bulunan en yakın komşular arasında rastgele interpolasyonla sentetik örnekler oluşturulması mantığına dayanmaktadır.[14] SMOTE'ta azınlık sınıfı ile k-en yakın komşular arasındaki farklar hesaplanmaktadır. Bu farklar, 0 ile 1 arasında rasgele bir değerle çarpılır. Elde edilen sonuç değeri



ŞEKİL 1: İş Akışı: Veri seti veri ön işleme sürecinde, eğitim ve test olacak şekilde ikiye ayrılmış, alt örnekleme, aşırı örnekleme ve SMOTE yöntemleri ile veri seti dengeli bir hale getirilmiştir. Dengelenen veri setine kolektif öğrenme yöntemleri uygulanmış ve rastgele seçilen bir kişiye diyabet tanısı konmuştur.

kullanılarak üretilen örnek, eğitim veri setine eklenir. Sonuç olarak, SMOTE, var olan örnekleri, komşuluk ilişkileri bazında rastgele herhangi bir nokta ekleyerek çalışır. Böylelikle, SMOTE rastgele aşırı örneklemeye benzer şeklinde, azınlık sınıfındaki örnek sayısını arttırma yönündedir. Ancak, aynı örneği yeniden oluşturmaz. Mevcut örnekleri, uygun şekilde birleştirerek yeni bir örnek oluşturur, böylelikle aşırı örnekleme yöntemindeki aşırı uyum dezavantajını bir dereceye kadar önleyebilmektedir.[15]

HASTALIK TANISI SINIFLANDIRMAYA ENTEGRE EDİLEN MAKİNE ÖĞRENMESİ ALGORİTMALARI

Dengesiz veri setlerinin yukarıda belirtilen yöntemler ile dengelenmesinden sonra makine öğrenmesi algoritmaları güvenle kullanılabilmektedir. Bu çalışmada kolektif öğrenme algoritmalarının performansının incelenmesine odaklanılmıştır. Kolektif öğrenme, diğer öğrenme yöntemlerindeki gibi tek bir sınıflandırıcı kullanmak yerine çoklu sınıflandırıcıların kullanıldığı hassas bir yöntemdir.[16,17] Torbalama ve artırma yöntemleri ise iki önemli kolektif öğrenme yöntemi olarak belirtilmektedir.[18,19] Torbalama yöntemi Breiman'ın 1996 yılında önermiş olduğu bir yöntemdir.[20] Torbalama, orijinal veri setinden rastgele örnekleme yöntemi ile oluşturulan örneklemlerden birden çok sınıf tahminini çoğunluklu oylamaya tabi tutarak, en yüksek oyu alan sınıflandırıcıyı nihai sınıf olarak belirlemektedir. Bu yöntemde, her iterasyonda tüm örneklere eğitim kümesi için eşit seçilme şansı verilmekte ve yeniden örnekleme yapılırken bir önceki sınıflandırıcının performansı dikkate almamaktadır; bu da torbalama yönteminin bağımlı bir yöntem olmadığını göstermektedir. Artırma, Shapire tarafından 1996 yılında önerilen bir yöntem olmakla birlikte her örnek için eşit dağılım ile eğitime başlanmakta ve sınıflandırma performansına göre zayıf sınıflandırıcılara odaklanmaktadır.[19] Bu zayıf sınıflandırıcılar ağırlıklandırılarak model tekrar eğitilmektedir. Belirli bir iterasyon sonucunda zayıf sınıflandırıcılar bir araya getirilerek güçlü bir sınıflandırıcı oluşturulmaktadır. Artırma yönteminde örneklerin eğitim kümesine seçilme olasılıkları sürekli güncellenmekte ve sonradan oluşturulan sınıflandırıcılar önceden oluşturulan sınıflandırıcılara bağımlı olmaktadır.[21,22]

RASTGELE ORMAN (RO)

Rastgele Orman (random forest), Leo Breiman tarafından, Breiman'ın 1996 yılında kendi önermiş olduğu torbalama yöntemi ve Ho'nun 1998 yılında önermiş olduğu rastgele alt-uzay yönteminin birleştirilmesi sonucunda 2001 yılında önerilmiş bir yöntemdir.[23]

RO, bağımsız aynı dağılmış rastgele vektörlerden meydana gelen ağaç yapılı sınıflandırıcı topluluğundan oluşan bir yöntemdir. Burada çok sayıda ağaç oluşturulduktan sonra her bir ağaç herhangi bir sınıfa oy verir ve en çok oyu alan sınıf ise nihai karar olarak belirlenir.[24]

RO algoritmasını büyük ve küçük boyutlu veri setlerinde yüksek doğrulukta sonuçlar vermektedir.[25] RO algoritmasını diğer algoritmalardan ayıran en önemli özelliklerden biri her ağaçtaki değişkenlerin farklı olmasından dolayı aşırı uyum probleminin oluşmamasıdır. RO algoritması yapısı gereği hızlı ve istenildiği kadar ağaçla çalışabilmesi nedeniyle şu ana kadarki algoritmalar arasında yüksek bir doğruluğa sahip algoritma olarak tanımlanmaktadır.[23,26]

AĞIRLIKLI ALT UZAY RASTGELE ORMAN (AAURO)

Ağırlıklı Alt Uzay Rastgele Orman (weighted subspace random forests), Xu tarafından önerilmiş bir yöntemdir. Burada karar ağaçları oluşturulurken rastgele değişken örneklemenin geleneksel yaklaşımı yerine alt uzay seçimi için değişken ağırlıklandırma yöntemi önermişleridir.[27] Yöntemde, bir değişkenin sınıfa göre bilgilendirilmesi bir bilgi kazancı oranı ile ölçülür. Ölçü, ağaç oluşturma işlemi sırasında belirli bir düğümü bölerken değişken alt uzaya dâhil olmak üzere seçilen değişkenin olasılığı olarak kullanılır. Bu nedenle, ölçüye göre daha yüksek değerlere sahip değişkenlerin değişken seçimi sırasında aday olarak seçilmesi daha olasıdır ve daha güçlü bir ağaç oluşturulabilir.

Ayrıca AAURO yönteminde rastgele orman yönteminden farklı olarak ağaçlar CART yerine C4.5 yöntemine dayanmaktadır.[28]

EKLEMELİ LOJİSTİK REGRESYON (ELR)

Eklemeli Lojistik Regresyon(boosted logistic regression), Friedman tarafından 1998 yılında önerilmiş bir yöntemdir. ELR yöntemi çok gürültülü verilerden kaynaklı aşırı uyum problemini çözmek için geliştirilmiştir.[29] ELR yöntemi, bu sorunu çözmek için eğitim hatasını doğrusal bir biçimde azaltmaktadır.[30]

Sonuç olarak ELR lojistik kayıp fonksiyonunu kullanarak aşırı uyumluluk sorununa neden olan verilerin ağırlığını daha da arttırmakta ve artırma yöntemine göre aşırı uyum sorununun üstesinden daha iyi gelmektedir.[31]

STOKASTİK GRADYAN ARTIRMA (SGA)

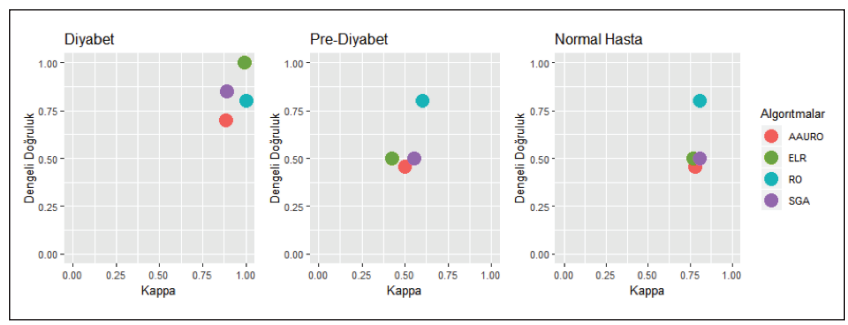
Stokastik Gradyan Artırma (Stochastic gradient boosting), Friedman tarafından gradyan artırma yöntemine rastgeleliliği dâhil ederek 1999 yılında önerdiği bir yöntemdir. SGA'da her yenilemede permütasyon örnekleme strateji ile rastgele alt örnek seçilir seçilen bu alt örnek, tüm öğrenicilerin yerine model güncellemesinin hesaplanması için ve ağaçlar arasındaki korelasyonu azalmak için kullanılır.[32]

BULGULAR

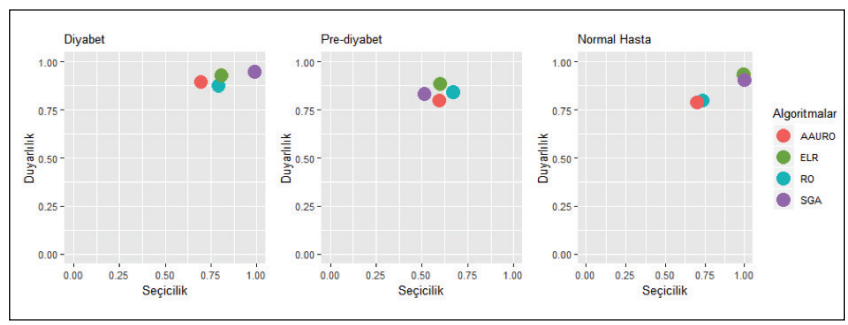
YENİDEN ÖRNEKLEME YÖNTEMLERİNİN SINIFLANDIRMA PERFORMANSINA ETKİLERİ

Diyabet veri setinin sınıflandırmasında daha önce söz edilen yeniden örnekleme yöntemleri ve makine öğrenmesi algoritmaları sırasıyla kullanılmış. Performans analizi için kullanılan girdilerde bir standartlaştırma yapılmamıştır. Veri setinin orijinal hali ile yapılan sınıflandırma sonucu kullanılan yöntemler için Şekil 2 incelendiğinde dengeli doğruluk ve Kappa değerlerinin yaklaşık 0.45'lere kadar düştüğü görülmüştür.

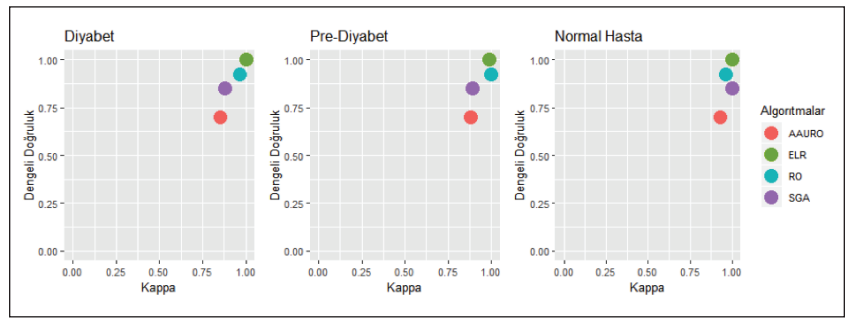
Şekil 2'deki değerlendirme ölçütlerine ek olarak duyarlılık ve seçicilik değerleri karşılaştırılmıştır (Şekil 3). Duyarlılık ölçütü göz önüne alındığında diyabet ve normal hasta sınıfları için SGA ve ELR oldukça başarı iken pre-diyabet sınıfı için ELR algoritması oldukça başarılıdır. Seçicilik ölçütü göz önüne alındığında ise diyabet sınıfı ve normal hasta sınıfı için SGA yöntemi oldukça başarı iken pre-diyabet sınıfı için çok yüksek bir başarı sonucuna ulaşamasada RO algoritması diğerlerine gore nispeten daha başarılı sonuçlar elde etmiştir (Şekil 3).



ŞEKİL 2: Dengesiz veri seti dengeli doğruluk ve Kappa ölçütleri.



ŞEKİL 3: Dengesiz veri seti torbalama ve artırma yöntemleri sınıflandırma başarı ölçütleri.



ŞEKİL 4: Alt örnekleme sonrası dengeli doğruluk ve Kappa ölçütleri .

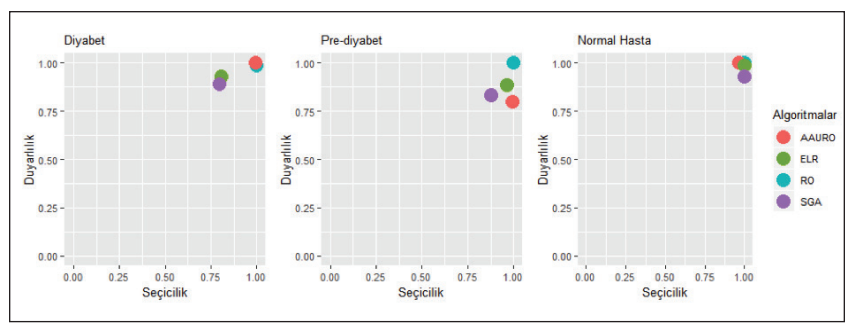
Genel olarak değerlendirildiğinde ham veri seti ile yapılan sınıflandırmada çok yüksek doğrulukta tahminler yapılamayacağı sonucuna varılmış ve yeniden örnekleme yönemleri uygulanmıştır. Alt örnekleme yönteminin uygulanmasından sonra dengeli doğruluk ölçütünün 0.70 ve 0.99 aralığında, Kappa değerlerinin yaklaşık 0.85 ve 0.99 aralığında değiştiği görülmüştür (Şekil 4).

Bunlara ek olarak duyarlılık ve seçicilik sınıflandırma başarı ölçütleri incelendiğinde ise sınıf içi en yüksek duyarlılık ve seçicilik değerine sahip yöntemin RO yöntemi olduğu görülmüştür (Şekil 5). Bu sonuçlar, alt örnekleme sonrasında elde edilen sınıflandırma sonuçlarının veri setinin dengesiz haline göre daha iyi bir sınıflandırma başarısına sahip olduğunu göstermektedir.

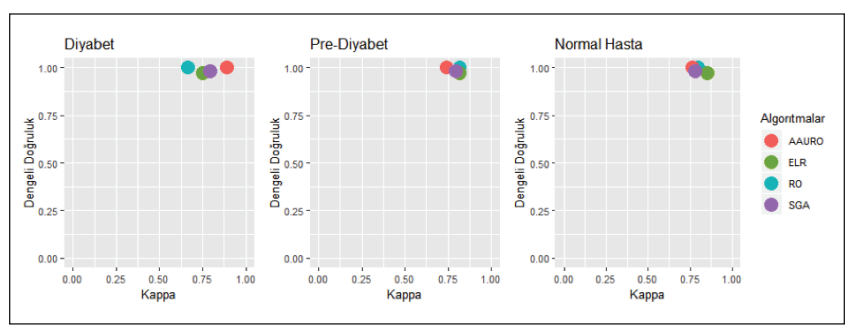
SMOTE sonrası Şekil 6 incelendiğinde ise dengeli doğruluk ölçütünün 0.95 ve 0.99 aralığında değiştiği, Kappa ölçütünün ise yaklaşık 0.60 ve 0.85 aralığında değiştiği görülmüştür.

Buna ek olarak duyarlılık ve seçicilik sınıflandırma başarı ölçütleri incelendiğinde ise torbalama ve artırma yöntemlerinin her birinin sınıf içi en yüksek duyarlılık ve seçicilik değerinde ulaştıkları görülmüştür (Şekil 7).

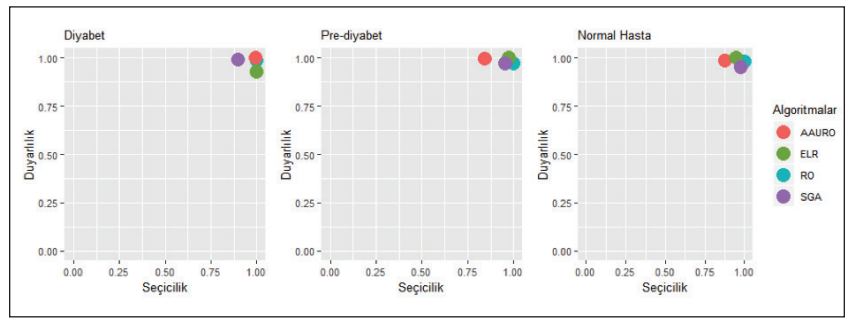
Aşırı örnekleme sonrası başarı ölçütleri incelendiğinde dengeli doğruluk ve Kappa değerlerinin yaklaşık 0.85 ve 0.99 aralığında değiştiği görülmüştür (Şekil 8).



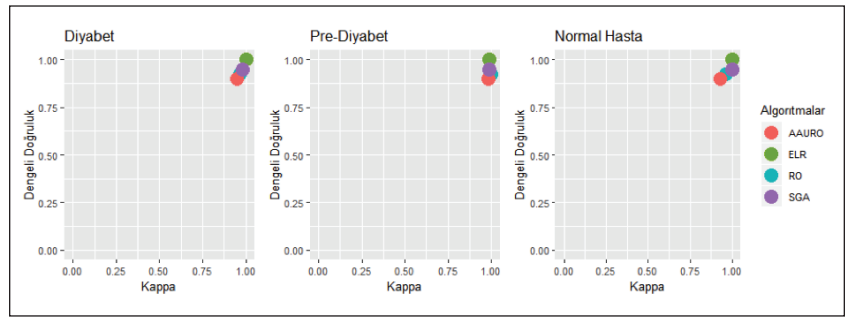
ŞEKİL 5: Alt örnekleme sonrası torbalama ve artırma yöntemleri sınıflandırma başarı ölçütleri.



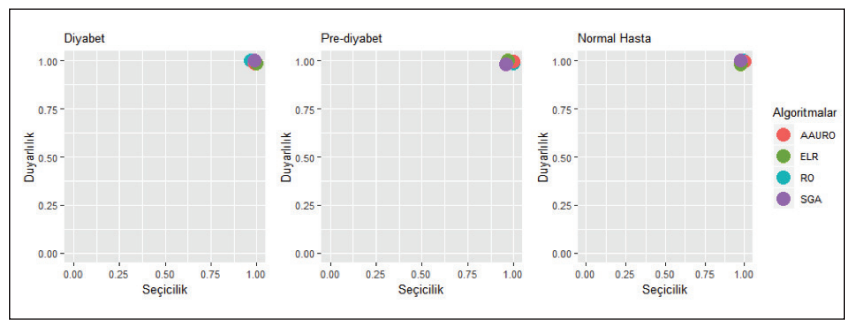
ŞEKİL 6: Sentetik azınlık aşırı örnekleme sonrası dengeli doğruluk ve Kappa ölçütleri.



ŞEKİL 7: Sentetik azınlık aşırı örnekleme sonrası torbalama ve artırma yöntemleri sınıflandırma başarı ölçütleri.



ŞEKİL 8: Aşırı sonrası dengeli doğruluk ve Kappa ölçütleri.



ŞEKİL 9: Aşırı örnekleme sonrası torbalama ve artırma yöntemleri sınıflandırma başarı ölçütleri.

Duyarlılık ve seçicilik sınıflandırma başarı ölçütleri incelendiğinde ise torbalama ve artırma yöntemlerinin birlikte sınıf içi en yüksek duyarlılık ve seçicilik değerlerine ulaştıkları görülmüştür (Şekil 9).

Bu bulguları desteklemek amacıyla dengesiz veri setleriyle sınıflandırma yapıldığında kullanılması önerilen F1 skoru ve Matthews korelasyon katsayısı ölçütleri de Tablo 2 ve Tablo 3'de verilmiştir. Dengesiz veri setleri ile sınıflandırma problemine hassas olan bu iki ölçüt ile de Şekil 2-4 de elde edilen bulgularla tutarlı sonuçlar elde edilmiştir.

Tüm bu sonuçlara bakıldığında alt örnekleme, aşırı örnekleme yöntemi ve SMOTE ile yapılan sınıflandırmaların, orijinal dengesiz veri seti ve alt örneklemeye göre çok daha başarılı tahmin gücüne sahip olduğu tespit edilmiştir. Bunun sebebi ise dengesiz veri setinde sınıflandırmada bu algoritmaların efektif olmamalarıdır.

MODELİN YENİ BİR HASTA İÇİN TEST EDİLMESİ

Son olarak, yapılan karşılaştırma sonuçlarından yola çıkarak diyabet tanısı için örnek bir uygulama yapılmıştır. Bu uygulama, makine öğrenmesi iş akışının son aşaması olan yorumlama aşamasının bir örneği olarak çalışmaya dahil edilmiştir. Öncelikle dengesiz veri seti performansı en yüksek yöntem olan aşırı örneklemeye tabi tutulmuş ve yapılan karşılaştırmalarda yüksek performans sonuçları veren rastgele orman yöntemi ile bir vaka çalışması yapılmıştır. Bu aşamada RO algoritması eğitildikten sonra performans değerlendirmesinde

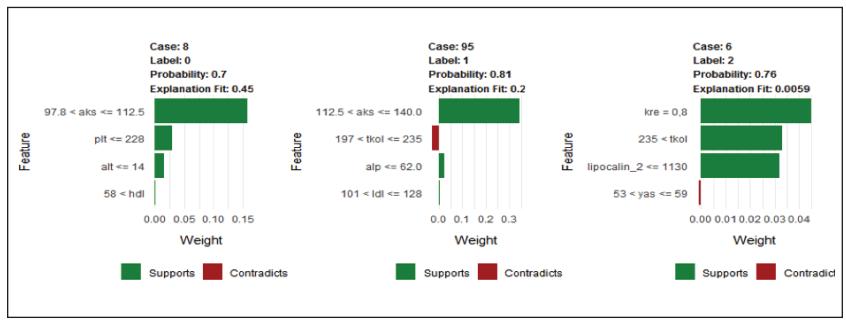
TABLO 2: Uygulanan yöntemler ve F1 skoru sonuçları.

Algoritma ve Sınıf	Dengesiz veri	SMOTE	Alt örnekleme	Aşırı Örnekleme
1 RO-Pre-diyabet	0.31	0.75	0.99	0.99
2 RO-Diyabet	0.79	0.61	0.92	0.99
3 RO-Normak hasta	0.76	0.71	0.92	0.99
4 AAURO-Pre-diyabet		0.77	0.60	0.99
5 AAURO-Diyabet	0.80	0.83	0.82	0.99
6 AAURO-Normal hasta	0.71	0.71	0.92	0.99
7 ELR-Pre-diyabet		0.81	0.99	0.99
8 ELR-Diyabet	0.83	0.79	0.99	0.99
9 ELR-Normal hasta	0.53	0.78	0.99	0.99
10 SGA-Pre-diyabet	0.18	0.74	0.86	0.99
11 SGA-Diyabet	0.80	0.85	0.87	0.99
112 SGA-Normal hasta	0.76	0.83	1.00	0.99

TABLO 3: Uygulanan yöntemler ve Matthews korelasyon katsayısı sonuçları.

Algoritma ve Sınıf	Dengesiz veri	SMOTE	Alt örnekleme	Aşırı Örnekleme
1 RO-Pre-diyabet	0.32	0.62	0.99	0.86
2 RO-Diyabet	0.32	0.62	0.99	0.86
3 RO-Normak hasta	0.32	0.62	0.99	0.86
4 AAURO-Pre-diyabet	0.00	0.56	0.59	0.86
5 AAURO-Diyabet	0.00	0.56	0.59	0.86
6 AAURO-Normal hasta	0.00	0.56	0.59	0.86
7 ELR-Pre-diyabet	0.00	0.56	0.99	0.90
8 ELR-Diyabet	0.00	0.56	0.99	0.90
9 ELR-Normal hasta	0.00	0.56	0.99	0.90
10 SGA-Pre-diyabet	0.09	0.26	0.71	0.99
11 SGA-Diyabet	0.09	0.26	0.71	0.99
112 SGA-Normal hasta	0.09	0.26	0.71	0.99

kullanılmamış olan yeni hastalar değerlendirilmiş ve önerilen modelleme yaklaşımı ile sınıflandırılmıştır. Yeni veri içinde 6., 8. ve 95. hastalar diyabet açısından sınıflandırılmıştır. Buna göre Şekil 10'da görüldüğü üzere 8. hastanın 0-etiketi ile verilen prediyabet sınıfına 0.7 olasılığı ile atandığı, 95. hastanın 1-etiketi ile verilen diyabet sınıfına 0.81 olasılığı ile atandığı ve 6. hastanın 2-etiketi ile verilen sağlıklı sınıfına 0.76 olasılığı ile atandığı söylenebilir. İlgili bulgular endokronoloji ve metabolizma hastalıkları uzmanı tarafından onaylanmıştır.



ŞEKİL 10: Test veri setindeki 6., 8. ve 95. hastalar için yapılan sınıflandırma işlemi sonuçları. Şekillerde y ekseni faktörleri ve karar kurallarını ve x ekseni ise bu faktörlerin sınıflandırma kararındaki ağırlıkları verilmektedir. O-etiketi: prediyabet, 1-etiketi: diyabet ve 2-etiketi: sağlıklı.

Kolektif öğrenme yöntemleri ile yapılan sınıflandırma çalışmalarında karar kuralları geleneksel yöntemlerde (karar ağacı gibi) olduğundan daha karmaşık yapıda olabilir. Yorumlamanın kolaylaşması amacıyla lime paketi ile yapılan analizler ile Şekil 10'da verildiği gibi her bir hastanın sınıflandırılmasına katkı sağlayan değişkenler ve bu değişkenler için karar kuralları görülebilir. Buna göre örneğin 8. hastanın prediyabet olarak sınıflandırılmasına en çok katkı sağlayan faktörün açlık kan şekeri (aks) olduğu ve bu faktörün değerinin 97.8 ve 112.5 arasında olması nedeniyle prediyabet sınıfına atandığı söylenebilir. Bunun yanında trombosit seviyesinin 228'den düşük olması ve alanin transaminaz seviyesinin 14'den düşük olması sınıflandırmaya katkı yapan diğer iki faktördür. Bu tür model agnostik analizleri risk faktörlerinin aralarindaki ilişkilerin ve karmaşık karar kurallarının değerlendirilmesi için kullanılabilir.

SONUÇ

Bu çalışmada, diyabet sınıflandırması için torbalama ve artırma temelli kolektif öğrenme yöntemlerinden RO, AAURO, SGA ve ELR yöntemleri kullanılmış ve yöntemlerin performansları karşılaştırılmıştır. Diyabet verisi sınıf dengesizliği problemine sahip olduğundan, öncelikle veri seti ön işlemeye tabi tutulmuştur. Ön işleme amacıyla alt örnekleme, aşırı örnekleme ve SMOTE yöntemleri sınıf dengesizliği problemini çözmek amacıyla kullanılmıştır. Böylelikle, veri ön işlemeden sonra bu yöntemlerin sınıflandırma performansı üzerindeki etkileri tartışılmıştır. Bulgular sonucunda, ham veri ile yapılan diyabet tanısına yönelik sınıflandırma başarısının oldukça düşük olması, dengesizliğin bir sonucu olarak yorumlanabilmektedir. Dengesiz veri seti ile diyabet sınıflandırmasında, aşırı örnekleme yöntemi ile yapılan sınıflandırmalar, alt örnekleme ve SMOTE yöntemleri ile yapılan sınıflandırmalardan daha yüksek doğruluğa sahiptir.

Ham veri setlerine doğrudan makine öğrenme yöntemlerinin uygulanması sınıflandırma başarıları üzerinde oldukça ciddi etkileri vardır. Bu çalışmada uygulandığı gibi, sınıf dengesizliği problemiyle baş edebilmek için ham veri setlerine yeniden örnekleme yöntemleri uygulanması ve veri setleri dengelendikten sonra sınıflandırma algoritmalarının kullanılması önerilmektedir. Böylelikle, doğru tanının konulması, tedavi yönteminin seçilmesinden ilaç kullanımına kadar pek çok noktada değişkenlik göstereceğinden, sağlık giderlerinde gereksiz harcamaların önüne geçilebilmesini sağlayabilecektir.

Finansal Kaynak

Bu çalışma sırasında, yapılan araştırma konusu ile ilgili doğrudan bağlantısı bulunan herhangi bir ilaç firmasından, tıbbi alet, gereç ve malzeme sağlayan ve/veya üreten bir firma veya herhangi bir ticari firmadan, çalışmanın değerlendirme sürecinde, çalışma ile ilgili verilecek kararı olumsuz etkileyebilecek maddi ve/veya manevi herhangi bir destek alınmamıştır.

Çıkar Çatışması

Yazarlar herhangi bir çıkar çatışması veya finansal destek bildirmemiştir.

Yazar Katkıları

Fikir/Kavram: *Sultan Turhan, Yüksel Özkan, B. Sarer Yürekli, Aslı Suner, Eralp Doğu;* ***Tasarım:*** *Sultan Turhan, Yüksel Özkan, B. Sarer Yürekli, Aslı Suner, Eralp Doğu;* ***Denetleme/Danışmanlık:*** *Aslı Suner, Eralp Doğu;* ***Veri Toplama Ve/Veya İşleme:*** *B. Sarer Yürekli;* ***Analiz Ve/Veya Yorum:*** *Sultan Turhan, Yüksel Özkan;* ***Kaynak Taraması:*** *Sultan Turhan, Yüksel Özkan;* ***Makalenin Yazımı:*** *Sultan Turhan, Yüksel Özkan, B. Sarer Yürekli, Aslı Suner, Eralp Doğu;* ***Eleştirel İnceleme:*** *Aslı Suner, Eralp Doğu;* ***Kaynaklar ve Fon Sağlama:*** *Aslı Suner, Eralp Doğu;* ***Malzemeler:*** *B. Sarer Yürekli.*

KAYNAKLAR

1. Goldenberg R, Punthakee Z. Definition, classification and diagnosis of diabetes, prediabetes and metabolic syndrome. Can J Diabetes. 2013;37(1):197-212. PMID: 24070969. [Crossref] [PubMed]
2. Laiteerapong N, Cifu AS. Screening for prediabetes and type 2 diabetes mellitus. JAMA. 2016;315(7):697-8. PMID: 26881373. [Crossref] [PubMed] [PMC]
3. Ogurtsova K, da Rocha Fernandes JD, Huang Y, Linnenkamp U, Guariguata L, Cho NH, et al. IDF Diabetes Atlas: global estimates for the prevalence of diabetes for 2015 and 2040. Diabetes Res Clin Pract. 2017;128:40-50. PMID: 28437734. [Crossref] [PubMed]

4. Jutel A. Classification, disease, and diagnosis. Perspect Biol Med. 2011;54(2):189-205. PMID: 21532133. [Crossref] [PubMed]
5. Liu Z, Tang D, Cai Y, Wang R, Chen F. A hybrid method based on ensemble WELM for handling multi class imbalance in cancer microarray data. Neurocomputing. 2017;266:641-50. [Crossref]
6. Wan S, Duan Y, Zou Q. HPSLPred: an ensemble multi-label classifier for human protein subcellular location prediction with imbalanced source. Proteomics. 2017;17(17-18). PMID: 28776938. [Crossref] [PubMed]
7. Zhang J, Cui X, Li J, Wang R. Imbalanced classification of mental workload using a cost-sensitive majority weighted minority oversampling strategy. Cogn Technol Work. 2017;19(4):633-53. [Crossref]
8. Wu Z, Lin W, Ji Y. An integrated ensemble learning model for imbalanced fault diagnostics and prognostics. IEEE Access. 2018;6:8394-02. [Crossref]
9. Khalilia M, Chakraborty S, Popescu M. Predicting disease risks from highly imbalanced data using random forest. BMC Med Inform Decis Mak. 2011;11(1):51. [Crossref] [PubMed] [PMC]
10. Zhou B, Li W, Hu J. A new segmented oversampling method for imbalanced data classification using quasi-linear SVM. IEEJ Trans Electr Electron Eng. 2017;12(6):891-8. [Crossref]
11. He H, Garcia EA. Learning from imbalanced data. IEEE Trans Knowl Data Eng. 2009;21(9):1263-84. [Crossref]
12. Alexander Yun-Chung Liu B. The Effect of Oversampling and Undersampling on Classifying Imbalanced Text Datasets. The University of Texas at Austin; 2004. http://fliphtml5.com/oefn/qjyp/basic/51-57
13. Lin WC, Tsai CF, Hu YH, Jhang JS. Clustering-based undersampling in class-imbalanced data. Inf Sci (Ny). 2017;(409-410):17-26. [Crossref]
14. Douzas G, Bacao F. Effective data generation for imbalanced learning using conditional generative adversarial networks. Expert Syst Appl. 2018;91:464-71. [Crossref]
15. Seo JH, Kim YH. Machine-learning approach to optimize SMOTE ratio in class imbalance dataset for intrusion detection. Comput Intell Neurosci. 2018;2018:9704672. PMID: 30515202. [Crossref] [PubMed] [PMC]
16. Zhou ZH. Ensemble Methods: Foundations and Algorithms. 1st ed. Cambridge: CRC Press; 2012. p.236. [Crossref]
17. Çolak MC, Çolak C, Erdil N, Arslan AK. Investigating optimal number of cross validation on the prediction of postoperative atrial fibrillation by voting ensemble strategy. Turkiye Klinikleri J Biostat. 2016;8(1):30-5. [Crossref]
18. Yang P, Yang JYH, Zhou B, Zomaya A. A review of ensemble methods in bioinformatics. Curr Bioinform. 2010;5(4):296-308. [Crossref]
19. Freund Y, Schapire RE. A desicion-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci. 1995;55(1):119-39. [Crossref]
20. Breiman L. Bagging predictors. Mach Learn. 1996;24(2):123-40. [Crossref] [Crossref]
21. Sewell M. Ensemble Methods. London: UCL; 2007. p.12.
22. Temel GO, Ankaralı H, Taşdelen B, Erdoğan S, Özge A. A comparison of boosting tree and gradient treeboost methods for carpal tunnel syndrome. Turkiye Klinikleri J Biostat. 2014;6(2):67-73.
23. Breiman L. Random forests. Mach Learn. 2001;45(1):5-32. [Crossref]
24. Akar Ö, Güngör O, Akar A. [Determination of land use area with random forest classifier]. Gebze: 3. Uzaktan Algılama ve Coğrafi Bilgi Sistemleri Sempozyumu; 2010. p.11-3.
25. Akman M, Genç Y, Ankaralı H. [Random forests methods and an application in health science]. Turkiye Klinikleri J Biostat. 2011;3(1):36-48.
26. Akşehirli ÖY, Ankaralı H, Aydın D, Saraçlı Ö. [An alternative approach in medical diagnosis: support vector machines]. Turkiye Klinikleri J Biostat. 2013;5(1):19-28.
27. Xu B, Huang JZ, Williams G, Wang Q, Ye Y. Classifying very high-dimensional data with random forests built from small subspaces. Int J Data Warehous Min. 2012;8(2):44-63. [Crossref]
28. Zhao H, Williams GJ, Huang JZ. wsrf: an R package for classification with scalable weighted subspace random forests. J Stat Softw. 2017;77(3):1-30. [Crossref]
29. Rätsch G, Onoda T, Müller KR. Soft margins for AdaBoost. Mach Learn. 2001;42(3):287-320. [Crossref]
30. Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting. Ann Stat. 2000;28(2):337-407. [Crossref] [Crossref]
31. Cai YD, Feng KY, Lu WC, Chou KC. Using LogitBoost classifier to predict protein structural classes. J Theor Biol. 2006;238(1):172-6. PMID: 16043193. [Crossref] [PubMed]
32. Friedman JH. Stochastic gradient boosting. Comput Stat Data Anal. 2002;38(4):367-78. [Crossref]