Sınıflandırmada Küçük ve Dengesiz Veri Kümesi Problemi

Small and Unbalanced Data Set Problem in Classification

Öznur Esra PAR
Bilgisayar Mühendisliği Bölümü
Hacettepe Üniversitesi
Ankara, Türkiye
par@hacettepe.edu.tr

Ebru AKÇAPINAR SEZER
Bilgisayar Mühendisliği Bölümü
Hacettepe Üniversitesi
Ankara, Türkiye
ebru@hacettepe.edu.tr

Hayri SEVER
Bilgisayar Mühendisliği Bölümü
Çankaya Üniversitesi
Ankara, Türkiye
sever@cankaya.edu.tr

***Özetçe*—Verilerinin sınıflandırılması, veri kümesinin küçük ve dengesiz olması durumunda zorlaşmakta ve sınıflama performansını direkt etkilemektedir. Veri setinin küçük olması ve/veya sınıflar arasında dengesizlik olması veri madenciliğinde büyük bir sorun haline gelmiştir. Sınıflama algoritmaları, veri setlerinin yeterli büyüklüğe sahip, dengeli olduğu varsayımı üzerine geliştirilmiştir. Bu algoritmaların çoğu, azınlık sınıfındaki örnekleri göz ardı ederken veya yanlış sınıflandırırken, çoğunluk sınıfa odaklanır. Medikal veri madenciliğinde bazı kısıtlardan dolayı küçük ve dengesiz veri seti problemi ile sıklıkla karşılaşılmaktadır. Çalışma kapsamında erişime açık hepatit veri seti, küçük veri setlerine bölünmüş, oluşturulan her bir veri seti uzaklık tabanlı yöntemlerle çoğaltılmıştır. Çoğaltılan veri setleri dört farklı makine öğrenmesi algoritması (Yapay Sinir Ağları, Destek Vektör Makineleri, Naive Bayes ve Karar Ağacı) kullanılarak sınıflandırılmış, elde edilen sınıflama sonuçları karşılaştırılmıştır.**

Anahtar Kelimeler — makine öğrenmesi; küçük veri seti; dengesiz veri seti; örneklem çoğaltma yöntemleri

***Abstract*— Classification of data is difficult in case of small and unbalanced data set and this problem directly affects the classification performance. Small and / or the imbalance dataset has become a major problem in data mining. Classification algorithms are developed based on the assumption that the data sets are balanced and large enough. The most of the algorithms ignore or misclassify examples of the minority class, focus on the majority class. Small and unbalanced data set problem is frequently encountered in medical data mining due to some limitations. Within the scope of the study, the public accessible data set, hepatitis, was divided into small and imblanced data subsets, each of the data subsets were oversampled by distance based data generation methods. The oversampled data sets were classified by using four different machine learning algorithms (Artificial Neural Networks, Support Vector Machines, Naive Bayes and Decision Tree) and the classification scores were compared.**

Keywords — machine learning; small data set; imbalanced data set; oversampling methods .

I. GİRİŞ

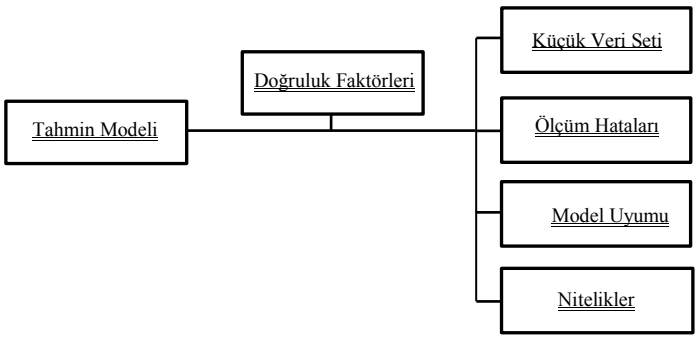
Veri madenciliğinin sağlık alanında kullanımı birçok açıdan yeni bir perspektif yaratmış ayrıca karar verme süreçlerini ivmelendirmeye başlamıştır[1].Veri madenciliğinin esas hedefi, veri yığınları içerisinden gelecek ile ilgili tahminde bulunurken katkı sağlayacağı düşünülen bağıntıların keşfedilmesidir. Başka bir deyişle, veri madenciliği, eldeki verilerden gizli kalmış, çok net olmayan, önceden bilinmeyen ancak potansiyel olarak kullanışlı bilginin çıkarılarak nitel modellerin oluşturulmasıdır[2]. Oluşturulan nitel modeller karar vermede etkin bir biçimde kullanılır.

Son yıllarda ortaya çıkan büyük veri çağı, daha hızlı ve daha iyi karar vermeyi gerektiren büyük ve karmaşık verileri de beraberinde getirmiş olmasına rağmen küçük veri kümesi problemleri de analiz ve karar vermenin zor olduğu alanlarda mücadele edilmesi gereken popüler bir konu olarak ortaya çıkmaktadır. Bir tahmin modeli oluşturmak için örneklem verisinin büyük olması gerekmektedir. Aksi durumda, küçük veri kümesi doğru bir tahmin modeli üretmek için yetersiz kalacaktır [3].

Medikal araştırmalar genellikle küçük ve dengesiz veri kümesi problemleriyle uğraşmaktadır. Hasta hastalık bulgu ilişkisinin karmaşıklığı ve / veya yüksek maliyetli tanı testleri, hastalık yaygınlığının düşük olması ve/veya hastalığa ait kayıtların eksik olması, numune seyrekliği, hasta verilerinin toplanmasının yüksek maliyetli olması medikal veri setlerini kısıtlamaktadır [4]. Bu nedenle, doğası gereği sağlık alanındaki çoğu hastalığa ait gözlem sayısı azdır.

Hastalık tahmin modellerinin oluşturulması birçok araştırmanın konusudur. Doğru bir tahmin modeli elde etmek için örnek veri setinin büyük olması beklenir. Aksi halde, oluşturulan model güvenilmez olarak kabul edilir [5].

Tahmin modelinin doğruluğunu etkileyen birçok faktör vardır [6].(Şekil 1)



Şekil 1: Tahmin modelinin doğruluğunu etkileyen faktörler

Şekil 1'de görüldüğü üzere, küçük veri kümesi, modelde eğitim veri kümesi olarak kullanılırsa, bilgi eksikliği nedeniyle tahmin belirsizliğine neden olabilir [7]. Küçük veri kümesinden elde edilen bilgi, öğrenme sistemi için güvenilmezdir ve kesin değildir [8]. Ayrıca, hesaplamalı öğrenme teorisi bağlamında, küçük veri seti, öğrenme performanslarına büyük bir yanlı etki sağlar.

II. Küçük ve Dengesiz Veri Seti ve Olasi Problemler

Tahmini değişken başına 10'dan az örnek olması, küçük veri seti için yeterli koşuldur [9].Küçük veri setine çoğunlukla, sınıflardaki gözlemlerin dengesiz dağılımı ile ortaya çıkan dengesiz veri seti problemi de eşlik eder. Veri madenciliğinde ortaya çıkan bazı problemler, örneklem verisinin boyutu küçük olduğundan daha da köpürebilir. Olası problemler aşağıda sıralanmıştır.

A. Aykırı Değerler

Birçok tahmin modeli için aykırı değerler, düşük oranda olmak koşuluyla önemlidir. Bu durum küçük veri kümesinde probleme neden olur. Çünkü birkaç aykırı değer bile, küçük veri kümesi içinde büyük oranlara sahip olabilir. Aykırı değerler, tahmin modelini büyük ölçüde değiştirerek doğruluktan uzaklaştıracaktır. [10]

B. Eğitim ve Test Verisi

Tahmin modeli oluşturulurken, verinin eğitim ve test verisi olarak ayrılması beklenir. Model, eğitim veri kümesi ile eğitilirken, test verisi olarak ayrılmış ve eğitimde kullanılmayan veriler üzerinde de performansı rapor edilir. Çapraz onaylama turu (cross validation round) ile veriler tümleyici alt kümelere bölünerek her bir alt küme için analiz işlemleri yapılır. Alt kümeler arasındaki değişkenliği azaltmak için, birden çok farklı alt kümeler kullanılarak çapraz onaylama turu gerçekleştirilir ve ardından sonuçların ortalaması alınır. Çapraz doğrulama, modelin performans tahmininde güçlü bir tekniktir[11]. Küçük veri kümelerinden tahmin modeli oluşturulmak istendiğinde, anlamlı performans değerlendirilmesi için gerekli gözlem sayısı çok az olabilir ve hatta çapraz doğrulama için gerekli miktarı barındırmayabilir.

C. Aşırı Uyum

Veri kümesi küçük olduğunda aşırı uyumun ortaya çıkması daha olasıdır. Aşırı uyumu elimine etmek için kullanılacak çapraz doğrulama yöntemi uygulama olasılığı da düşüktür[12].

D. Ölçüm Hataları ve Eksik Değerler

Veri kümesindeki her bir değer (nitelikler ve sınıf değişkeni) ölçüm hatası barındırabilir. Küçük veri kümesinde, bu tür hataların etkileri önem kazanmakta ve modeli olumsuz yönde etkilemektedir. Verilerdeki eksik değerler, ölçüm hataları ile aynı yönde fakat daha şiddetli etkiye sahiptir. Sınırlı sayıda gözlem, eksik değerlerin tamamlanmasında (impute) daha zorlayıcıdır. Sınıf değişkeninde eksiklikler varsa, gözlemin veri kümesinden çıkarılması, küçük veri kümesinin daha da küçülmesine neden olacaktır.

E. Yanlılık

Verilerin yanlı olması veya örneklem verisinin popülasyonun tamamını temsil etmemesi, küçük veri kümesinde sıklıkla karşılaşılır[13].

III. Küçük Veri Seti Problemi İçin Geleneksel Çözümler

A. Önyükleme (Bootstrapping)

Küçük veri kümelerindeki öğrenme problemini gidermek için sanal örnek oluşturma yöntemleri kullanılabilir. En kapsamlı olarak uygulanan sanal örnek oluşturma yöntemlerinden biri, belirli olasılıkla verileri örneklerden tekrar örnekleyerek yeni eğitim setleri oluşturan önyükleme (bootstrapping) yordamıdır. Bu yaklaşımın yararı, bir örneği en az iki kez eğitmesi ve böylece gerçek verilerin davranışlarını temsil etmelerine olanak sağlamasıdır[14]. Bunun yanında, yaklaşım aşırı öğrenme problemine neden olmaktadır.

B. Azınlık Sınıf için Uzaklık Tabanlı Örneklem Çoğaltma

Veri kümesinde eşit olmayan sınıf dağılımı olduğunda dengesiz sınıflandırmadan bahsedilmektedir. Çoğu sınıflandırma yöntemiyle yapılan tahminler çoğunluk sınıfı için oldukça doğru olmasına rağmen azınlık sınıfı için ise çok daha az doğrudur. Sınıflandırma doğruluğu, çoğunluk sınıfının lehine yanlı olduğundan, azınlık sınıfı çoğu zaman yanlış sınıflandırılır. Bu nedenle, dengesiz veri sorunu son on yılda yoğun bir şekilde araştırılmış ve ele alınması için çeşitli yöntemler önerilmiştir[15].

Örnek sayısının artırılması ya da azaltılmasında temel amaç, sınıflar arasındaki dengesizliği ortadan kaldırmaktır.

Azınlık sınıfındaki orijinal örneklerden farklı suni örnekler üretmek için uzaklık tabanlı örnekleme yöntemleri de kullanılır. Bu yöntemde genel olarak, orijinal azınlık sınıfı örneklerinin k-en yakın komşuları dikkate alınır. En yakın komşular arasındaki sürekli vektörler boyunca sentetik veriler üretir. Azınlık sınıfındaki bilgi boşlukları sentetik örneklerle dolmasına rağmen, azınlık sınıfındaki gerçek örneklerin alanı içinde dağıtılır[16].

Literatürde sıklıkla adı geçen uzaklık tabanlı veri üretme yöntemleri; ROSE, ADASYN, SMOTE, DBS, RSLS, SLS'dir.

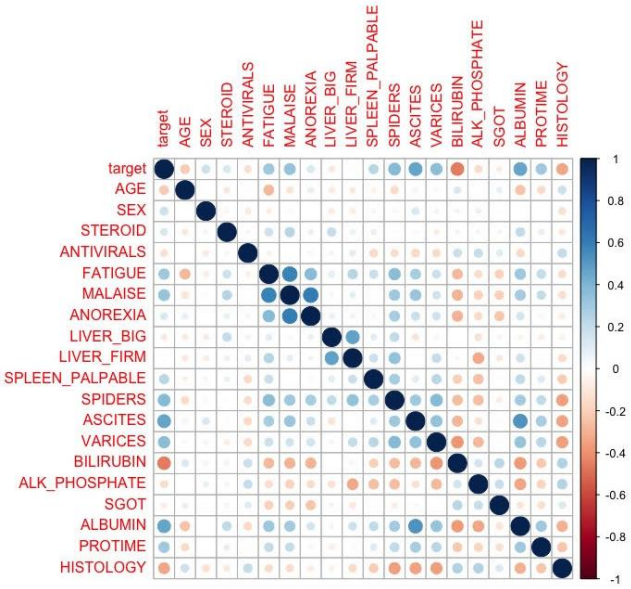
IV. DENEYSEL ÇALIŞMA

A. Veri Seti

UCI veri tabanından, açık erişimli hepatit veri seti deney kapsamında kullanılmıştır [17].
Veri setinin genel özellikleri TABLO I'de, korelasyon görseli Şekil 2'de gösterilmiştir.

TABLO I. KULLANILAN VERİ SETİNE AİT ÖZELLİKLER

Veri Seti Adı	Hepatit	
Gözlem sayısı	155	
Nitelik Sayısı	19	
Sınıf Sayısı	2	
Sınıf Dağılımı	32	123
Kıyaslama Doğruluğu	Max. %83	



Şekil 2: Korelasyon görseli

B. Deney Düzeneğinin Oluşturulması

1) 32-123 sınıf dağılımına sahip olan orijinal veri kümesinin eğitim ve test kümesi olarak %50 oranında ayrılması,

a) Örnek veri kümesi DS1 olarak anılmaktadır.

b) Eğitim veri setinde 78 (16-62), test veri setinde 77 (16-61) gözlem bulunmaktadır.

2) Azınlık sınıfındaki gözlem sayısı 12 olacak şekilde, 3 veri alt kümesinin oluşturulması,

a) Örnek veri kümesi DS2 olarak anılmaktadır.

b) Eğitim veri setinde 68 (6-62), test veri setinde 67 (6-61) gözlem bulunmaktadır.

3) Azınlık sınıfındaki gözlem sayısı 22 olacak şekilde, 3 veri alt kümesinin oluşturulması

a) Örnek veri kümesi DS3 olarak anılmaktadır.

b) Eğitim veri setinde 73 (11-62), test veri setinde 72 (11-61) gözlem bulunmaktadır.

4) Oluşturulan 3 veri seti üzerinde 4 sınıflama algoritması (Yapay Sinir Ağları, Destek Vektör Makineleri, Naive Bayes ve Karar Ağacı (C4.5))koşturularak sınıf içi kesinlik değerlerinin hesaplanması,

5) Her bir veri kümesi için, uzaklık tabanlı örneklem çoğaltma yöntemlerinden ROSE, ADASYN, SMOTE, DBS, RSLS, SLS kullanarak azınlık sınıfına ait gözlemlerin çoğaltılması,

6) Çoğaltılmış veri seti üzerinde 4 sınıflama algoritmasının (Yapay Sinir Ağları, Destek Vektör Makineleri, Naive Bayes ve Karar Ağacı (C4.5))koşturularak sınıf içi kesinlik değerlerinin hesaplanması ve karşılaştırılması.

C. Deney Sonuçları

Herhangi bir uzaklık tabanlı örneklem çoğaltma işlemi uygulanmamış ve önceki bölümde açıklandığı üzere oluşturulan veri kümelerinin azınlık sınıflarına ait kesinlik değerleri TABLO II'de sunulmuştur. Veri kümesinde dramatik bir azalma ve dengesizleşme var olduğunda azınlık sınıfa ait kesinlik değerleri de düşme eğilimi göstermektedir.
Azınlık sınıf için en yüksek sınıflandırma kesinliği, DS1 veri kümesi için Karar Ağacı algoritması ile elde edilmişken, görece dengesizleştirilmiş DS2 ve DS3 veri setlerinde ise Naive Bayes ile elde edilmiştir.

TABLO II. ÇOĞALTMA İŞLEMİ UYGULANMAMIŞ VERİ SETLERİNDE AZINLIK SINIFA AİT SINIF İÇİ KESİNLİK DEĞERLERİ

	DS1	DS2	DS3
ANN	0.7311	0.5573	0.5381
SVM	0.8571	0.000	0.7681
NB	0.7673	0.6572	0.8331
C4.5	0.9295	0.3909	0.000

Uzaklık tabanlı yöntemlere göre çoğaltılmış veri setlerinde şu sonuçlar elde edilmiştir.

- ROSE yöntemi ile üretilen verilerde, her bir veri kümesi için en iyi sonuç Naive Bayes algoritması ile elde edilmiştir.
- ADASYN yöntemi ile üretilen verilerde, her bir veri kümesi için en iyi sonuç Naive Bayes algoritması ile elde edilmiştir.
- SMOTE yöntemi ile üretilen verilerde, her bir veri kümesi için en iyi sonuç Naive Bayes algoritması ile elde edilmiştir.
- DBS yöntemi ile üretilen verilerde, her bir veri kümesi için en iyi sonuç Naive Bayes algoritması ile elde edilmiştir.
- RSLS yöntemi ile üretilen verilerde, DS1 ve DS3 veri kümeleri için en iyi sonuç Naive Bayes algoritması ile elde edilmiş, DS2 veri kümesi için en iyi sonuç Yapay Sinir Ağları ile elde edilmiştir.
- SLS yöntemi ile üretilen verilerde, DS1 ve DS3 veri kümeleri için en iyi sonuç Naive Bayes algoritması ile elde edilmiş, DS2 veri kümesi için en iyi sonuç Karar Ağaçları ile elde edilmiştir.

Sonuçlar TABLO III'te sunulmuştur.

TABLO III. ÇOĞALTILMIŞ VERİ SETLERİNDE AZINLIK SINIFA AİT SINIF İÇİ KESİNLİK DEĞERLERİ

DS1						
	ROSE	ADASYN	SMOTE	DBS	RSLS	SLS
ANN	0.3333	0.3871	0.3658	0.4642	0.5000	0.3871
SVM	0.5384	0.4062	0.5172	0.5454	0.4166	0.4090
NB	0.9375	0.8750	0.7500	0.5625	0.7333	0.7333
C4.5	0.4000	0.4210	0.2142	0.4285	0.4193	0.2727

DS2						
	ROSE	ADASYN	SMOTE	DBS	RSLS	SLS
ANN	0.1724	0.1593	0.1988	0.1166	0.5580	0.1666
SVM	0.1866	0.1468	0.2757	0.1000	0.1372	0.0000
NB	0.7222	0.5555	0.7340	0.5000	0.3068	0.2476
C4.5	0.1500	0.1637	0.1742	0.1839	0.3555	0.3317

DS3						
	ROSE	ADASYN	SMOTE	DBS	RSLS	SLS
ANN	0.3311	0.3765	0.3381	0.4260	0.5076	0.5541
SVM	0.3438	0.3237	0.3662	0.4387	0.5681	0.5222
NB	0.9090	0.7575	0.7878	0.6666	0.6381	0.6011
C4.5	0.2839	0.2559	0.2943	0.3361	0.3574	0.3474

V. SONUÇ

Küçük ve dengesiz veri seti ile sınıflandırma yapılırken sınıflandırma doğruluğu, çoğunluk sınıfının lehine yanlı olduğundan, azınlık sınıfı çoğu zaman yanlış sınıflandırılır. Örnek sayısının artırılması ya da azaltılmasında temel amaç, sınıflar arasındaki dengesizliği ortadan kaldırmaktır. Çalışma kapsamında, orijinal veri seti kademeli olarak küçültülüp ve dengesizleştirilerek uzaklık tabanlı azınlık örneklem çoğaltma yöntemleri uygulanarak kıyaslanmıştır.

Elde edilen deney sonuçlarında SLS çoğaltma yöntemi uygulanan DS2 veri seti SVM algoritmasında (Tablo III), çoğaltma yöntemi uygulanmadan evvel DS3 veri setinde C4.5 algoritması (Tablo II) azınlık sınıfında her hangi bir doğru sınıflandırma yapamamıştır. C4.5 algoritması, küçük ve dengesiz veri setleri için uygun bir sınıflandırma algoritması değildir. Çoğaltılmış DS2 veri seti için SVM algoritmasının SLS skorunun 0.0000 olması, diğer sonuçlarla kıyaslandığında şaşırtıcı bulunmamıştır.

ROSE yöntemi, Naïve Bayes sınıflandırmasında tüm veri kümeleri için azınlık sınıfta en yüksek kesinliği vermiş, SMOTE ise yine Naïve Bayes sınıflandırmasında küçültülmüş ve dengesizleştirilmiş veri kümelerinde azınlık sınıf için en yüksek kesinliği vermiştir. Küçük ve dengesiz veri kümelerinin uzaklık tabanlı azınlık sınıf çoğaltılmasında ROSE ve SMOTE yöntemleri, diğerler yöntemlere tercih edilebilir. Küçük ve dengesiz veri setlerinde Naïve Bayes sınıflandırılması önerilmektedir.

Sonraki çalışmada, küçük ve dengesiz veri kümeleri bahsi geçen uzaklık tabanlı azınlık sınıf çoğaltma yöntemleri ile kademeli olarak çoğaltılacak ve sınıflama skorları karşılaştırılacaktır. Çoğaltma yönteminden bağımsız olarak, üretilecek olan sentetik veri sayısı araştırma konusudur.

KAYNAKLAR

[1] Kaya E, Bulun M, Arslan A, "Tıpta Veri Ambarları Oluşturma ve Veri Madenciliği Uygulamaları", Çukurova Üniversitesi, Akademik Bilişim 2003.

[2] Düzgün O, "Veri Madenciliği Yöntemleri Ve İşkur İçin Uygulamaya Yönelik Model Önerisi", Uzmanlık Tezi, 2017.

[3] Longadge R, Dongre S, Malik L, "Class Imbalance Problem in Data Mining: Review", International Journal of Computer Science and Network (IJCSN) Volume 2, Issue 1, February 2013, ISSN 2277-5420.

[4] Berner E, "Clinical Desicion Support Systems",Department of Health Services Administration University of Alabama at Birmingham, USA, Springer, ISBN -10: 0-387-33914-0, 2006.

[5] Lateh MA, Muda A , Yusof Z , Muda N and Azmi M, "Handling a Small Dataset Problem in Prediction Model by employ Artificial Data Generation Approach: A Review", IOP Conf. Series: Journal of Physics: Conf. Series 892, 2017.

[6] C.-H. Tsai and D.-C. Li, "Improving Knowledge Acquisition Capability of M5' Model Tree on Small Datasets," 2015 3rd Int. Conf. Appl. Comput. Inf. Technol. Int. Conf. Comput. Sci. Intell., pp. 379–386, 2015.

[7] D. C. Li, C. C. Chang, C. W. Liu, and W. C. Chen, "A new approach for manufacturing forecast problems with insufficient data: The case of TFT-LCDs," J. Intell. Manuf., vol. 24, no. 2, pp. 225–233, Jul. 2013.

[8] D. C. Li, C. Sen Wu, T. I. Tsai, and Y. S. Lina, "Using mega-trend-diffusion and artificial samples in small data set learning for early flexible manufacturing system scheduling knowledge," Comput. Oper. Res., vol. 34, no. 4, pp. 966–982, 2007.

[9] Torgyn Shaikhina Natalia A.Khovanova, "Handling limited datasets with neural networks in medical applications: A small-data approach", Artificial Intelligence in Medicine, Pages 51-63, Volume 75, January 2017.

[10] Justin Collins, BS; Jordan Brown, BS; Christine Schammel, PhD; Kevin Hutson, PhD; and W. Jeffery Edenfield, "Meaningful Analysis of Small Data Sets: A Clinician's Guide", SC (J.B., K.H.); and Pathology Consultants, Greenville, SC (C.S.), GHS Proc. 2 (1): 16-19, June 2017.

[11] https://www.techopedia.com/definition/32064/cross-validation

[12] Japkowicz N, "Learning from Imbalanced Data Sets: A Comparison of Various Strategies " AAAI Technical Report WS-00-05, 2000.

[13] Chawla N, "Data Mining for Imbalanced Datasets: An Overview", Data Mining and Knowledge Discovery Handbook pp 875-886, Springer, US, 2010.

[14] Steven Abney, "Bootstrapping", Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, pp. 360-367, July 2002.

[15] Meng-Fong Tsai, Shyr-Shen Yu, "Distance Metric Based Oversampling Method for Bioinformatics and Performance Evaluation", Journal of Medical Systems, 2016.

[16] Sotiris Kotsiantis, Dimitris Kanellopoulos, "Handling imbalanced datasets: A review",GESTS International Transactions on Computer Science and Engineering, Vol.30, 2006

[17] https://archive.ics.uci.edu/ml/datasets/hepatitis