SINIFLANDIRICI TOPLULUKLARININ GÜRÜLTÜLÜ VERİLERE KARŞI GÜRBÜZLÜĞÜNÜN DEĞERLENDİRİLMESİ

EVALUATION OF ROBUSTNESS OF ENSEMBLE LEARNERS TO NOISY DATA

M.Özgür Cingiz, Abdülkadir Albayrak, M.Fatih Amasyalı
Bilgisayar Mühendisliği Bölümü
Yıldız Teknik Üniversitesi
İstanbul, Türkiye
mozgur@ce.yildiz.edu.tr, albayrak@yildiz.edu.tr, mfatih@ce.yildiz.edu.tr,

***Özetçe*— Sınıflandırma ile ilgili çalışmalarda veri kümelerinde gürültü verilerin belirlenmesi ve bu veri kümeleri üzerinde doğru sınıflandırma sonucu veren sınıflandırıcıların keşfi problematiktir. Çalışmamızda 36 UCI veri kümesine çeşitli oranlarda gürültü eklendikten sonra gürültülü veri kümeleri üzerinde beş farklı sınıflandırıcı topluluğu yaklaşımı ve iki temel sınıflandırıcı yaklaşımlarının sınıflandırma başarıları birbirleriyle karşılaştırılmıştır. Buna göre gürültüye en dayanaklı sınıflandırıcı Rastgele Uzaylar yaklaşımı ve Bagging yaklaşımı olarak belirlenmiştir.**

Anahtar Kelimeler — Gürültülü Veriler;Sınıflandırıcı Toplulukları; Sınıflandırma

***Abstract*— Discovering noisy data and classification of noisy data sets are problematic issues associated with noisy data sets. In our work, we used 36 UCI data sets that consist of differeent rates of noisy data to measure robustness of five ensemble learners and two basic classifiers to noisy data. According to classification success ratesof our study, Random Subspace and Bagging are more robust to noisy data than other ensemble learners and simple classifiers.**

Keywords — Noisy Data; Ensemble Methods; Classification

I. GİRİŞ

Makine öğrenmesinde en iyi sınıflandırma sonucunu veren sınıflandırıcıların keşfi önemli bir aşamadır. Bu nedenle ilgili veri kümesi üzerinde eğitim ve test aşamasında çok sayıda sınıflandırıcı kullanılarak en iyi sınıflandırma sonucunu veren sınıflandırıcı bulunmaya çalışılır. Bu problemi aşmak ve test verisine doğru sınıfı atama işlemi için birden fazla sınıflandırıcının ortak belirlediği sınıf değerini test verisinin sınıf etiketi olarak belirleyen sınıflandırıcı topluluklarının kullanımı yaygın olarak kullanılan bir yaklaşımdır. Buradaki temel yaklaşım çok sayıda sınıflandırıcının demokrasi usulüyle belirlediği sınıf etiketinin test verisinin sınıf etiketi olarak belirlenmesine dayanır. Sınıflandırıcı topluluklarını oluşturan temel sınıflandırıcıların test verisi için belirledikleri sınıf değerlerinin birbirinden farklı olması ve her bir temel sınıflandırıcıların yüksek sınıflandırma başarısı elde etmesi sınıflandırıcı topluluklarından beklenmektedir. Test verisi için farklı sınıf değeri belirleyen sınıflandırıcılarla oluşturulmuş sınıflandırıcı topluluğunun amacı sınıflandırıcılardan kaynaklanan aşırı uyumu (overfitting) önlemektir. Böylece yeni gelen test verisi için daha genel bir sınıflandırma gerçekleştirilir.

Sınıflandırıcı topluluklarını oluşturan temel öğrenicilerin birbirlerinden farklı sonuçlar üretmesi için aynı veri üzerinde kullanılan farklı sınıflandırıcı seçimi aynı sınıflandırıcıların farklı parametreleri kullanılarak veya farklı yaklaşımlarla oluşturulmuş sınıflandırıcılarla elde edilir.

Günümüzde gelişen ve değişen teknolojiye, artan veri alma kaynaklarının kalitesine rağmen verilerin doğruluğu ve bütünlüğü tam olarak sağlanamamaktadır. Bundan dolayı uygulanan metotların gürültülere karşı duyarlılıkları ve dayanıklılıkları önem kazanmaktadır. Çalışmamızda çeşitli sınıflandırıcı toplulukları yaklaşımlarının gürültülü veriye karşı gürbüzlüğünüz kontrolü belirlenmeye çalışılmıştır. Gürültülü verilerin oranlarına bakılarak farklı sınıflandırıcı toplulukları yaklaşımlarının sınıflandırma başarıları değerlendirilmiştir.

Çeşitli sınıflandırıcıların farklı oranlardaki gürültülü veriler üzerindeki başarılarını gösteren literatürde pek çok çalışma yer almaktadır. Folleco'nun yaptığı çalışmada 11 farklı sınıflandırıcının gürültülü veriler kullanarak elde ettikleri sınıflandırma başarıları değerlendirilmiştir [1]. Folleco gürültüleri iki farklı şekilde ekleyerek sonuçlarını elde etmiştir. Çalışmasında kullandığı ilk gürültü ekleme yöntemi sadece verilerin özellik değerlerinin değiştirilmesi ve bu değerler yerine gürültülü veriler eklenmesine dayanmaktadır. Sadece verilerin özellik değerleri ile oynandığında en iyi sınıflandırma sonuçlarını rastgele orman sınıflandırıcı topluluğu ve kural tabanlı yaklaşımlardan elde edildiği gözlemlenmiştir. Folleco aynı zamanda sadece eğitim verilerinin sınıflandırıcı etiketlerini değiştirip, test verilerinin sınıf etiketlerini aynı bırakarak sınıflandırıcıların sınıflandırma performansını değerlendirmiştir. Buradan elde edilen sonuçlara göre en

başarılı sınıflandırıcı kural tabanlı sınıflandırıcıyken, kural tabanlı sınıflandırıcıyı rastgele orman sınıflandırıcı topluluğu takip etmiştir. Opitz çalışmasında Yapay Sinir Ağları (YSA) ve Karar Ağaçları (KA) temel sınıflandırıcı olarak alarak Bagging ve Adaboost sınıflandırıcı topluluklarının gürültülü verilerle elde ettikleri sınıflandırma başarılarını değerlendirmiştir [2]. Bu çalışma sonucunda temel sınıflandırıcı olarak YSA kullanıldığında gürültülü veriye karşı elde edilen sınıflandırma başarı değerlerinin KA'dan daha yüksek çıktığı gözlemlenmiştir. Sınıflandırıcı toplulukları yaklaşımları açısından ise Bagging yaklaşımının Adaboost yaklaşımından daha iyi sonuç verdiği sonucu elde edilmiştir. Adabooost yaklaşımı Bagging yaklaşımına göre gürültülü veriye daha hassas olduğu sonucu çıkarılmıştır. Melville ise gürültülü verilerde sınıflandırıcıları karşılaştırmak içinAdaboost, Decorate ve Bagging yaklaşımlarını kullanmıştır [xx]. Elde edilen sonuçlardan elde edilen sonuçlar bir önceki çalışmalarda elde edilen sonuçlarla örtüşmektedir. Buna göre gürültü veri oranı arttıkça Adaboost sınıflandırıcı topluluğundan elde edilen sınıflandırma başarı değerleri düşmektedir. Gürültülü verilerde en yüksek başarı yine Baggingyaklaşımıyla elde edilmiştir. Gürültü oranı arttıkça Adaboost sınıflandırıcı topluluğundan elde edilen sınıflandırma performansı tek bir karar ağacından elde edilen sınıflandırma başarı değerlerinin gerisinde kalmıştır.

Çalışmamızda ikinci bölümde kullandığımız temel sınıflandırıcılara yer verilmiştir. Üçüncü bölümde yöntem ve sonuçlardan bahsedilerek sonuçlar değerlendirilmiştir. Dördüncü bölümde ise çalışma sonunda elde edilen çıkarımlar ve gelecekte yapılması düşünülen çalışmalara yer verilmiştir.

II. SINIFLANDIRICILAR

A. Rastgele Orman(RandomForest)

Bir sınıflandırıcı topluluğu olan rastgele orman ilk defa LeoBreiman tarafından önerilmiştir. Temel öğrenicileri (baselearner) karar ağaçlarıdır [4]. Eğitim aşamasında örneklerin %63 ünü kullanan Bagging algoritmasından faydalanmaktadır. Ancak öznitelik seçimi rastgele gerçekleştirilmektedir. Test aşamasında karar verme işlemi demokrasi usulüne dayanır. Sonuç her bir temel öğrenicinin kararlarının birleştirilmesiyle elde edilir. Çalışmamızda Rastsal Orman'da temel sınıflandırıcı olarak C4.5 algoritmasını kullanmaktadır. RandomForest'ta kullanılan ağaç sayısı varsayılan değer olarak belirtilen 10 ağaçtır.

B. Rotasyon Ormanı (RotationForest)

Rotasyon Ormanı, Sınıflandırılması istenen bir veri kümesinden X adet alt küme oluşturulur. Bu X adet alt kümeye Temel Bileşen Analizi (TBA) uygulanarak boyut indirgeme işlemi gerçekleştirilir [5]. Böylece ayırt ediciliği daha fazla olan özellikler seçilmiş olur. Rotasyon ormanlarında, Temel öğrenici olarak bootstrap algoritması kullanılır. Çalışmamızda Rotasyon Ormanları'nda temel sınıflayıcı olarak C4.5 algoritması kullanılmaktadır. Verilerin farklı uzaya taşımak için temel bileşenler analizi uygulanmaktadır. Verilerin sınıflandırılması için doğrusal dönüşüm gerçeklenmektedir.

C. Bagging

Bagging (Bootstrap aggregating) algoritmasında, tüm örneklerin %63'ü olan n tane örnekle temel öğreniciler eğitilip her bir temel öğrenici için bir K eşik değerine göre sınıflandırma işlemi gerçekleştirilir [6]. Örnekler veri kümesindeki örneklerin birebir aynısı olmadığından gürültülü verilere karşı gürbüz (Robust) olması beklenir. Bagging'te temel öğreniciler eğitim verisine duyarlı ise başarılı sonuç vermesi beklenir. Çalışmamızda Bagging için temel sınıflayıcı olarak C4.5 algoritması kullanılmıştır. Bagging'te yeniden örnekleme yapılarak eğitilmiştir. İterasyon sayısı da 10 olarak alınmıştır.

D. Rastgele Altuzaylar (RandomSubSpace)

Rastgele altuzaylar, rastgele ormanın genelleştirilmiş halidir. Rastgele ormanda temel sınıflayıcı olarak karar ağaçları kullanılırken rastgele altuzaylar temel sınıflayıcı olarak farklı sınıflayıcılar kullanabilmektedir. Daha DVM, nearestneigbourhood gibi sınıflayıcıların kullanıldığı doğrusal sınıflandırma işlemlerinde kullanılmaktadır. Rastgele Altuzaylar yönteminde temel sınıflayıcı olarak Reptree(varsayılan parametre) kullanılmıştır. Reptree hızlı karar verme özelliği bulunan karar ağacıdır. Alt uzay büyüklüğü 0.5 ve iterasyon sayısı 10 olarak alınmıştır.

E. AdaBoost

AdaBoost, Temel sınıflayıcılar tarafından seçilecek örneklerin sınıflandırılmasında bagging'den farklı olarak çalışmaktadır. Bagging algoritmasında bütün örneklerin ağırlıklandırması eşit olmaktadır. Buna karşın Adaboost'ta sınıflandırma işlemine müdahale etmeyip yanlış sınıflandırılan örneklerin ağırlıklarını arttırır. Böylelikle doğru olarak snıflandırılan verilerle tekrar işlemek yerine yanlış sınıflandırılan örneklere yoğunlaşıp sınıflandırma başarımını arttırmayı hedeflemektedir. Çalışmamızda, Adaboost için temel sınıflayıcı olarak DecisionStump kullanılmıştır. İterasyon sayısı 10 alınmıştır. Yeniden örnekleme yerine ağırlıklandırma seçilmiştir.

F. NaiveBayes

NaiveBayes yaklaşımı özellikler birbirinden bağımsız olduğu varsayımına dayanır. NaiveBayes özelliklerin her bir sınıfta yer alma olasılıklarının çarpımlarını göz önüne alarak, özellik değerlerinin çarpımlarının maksimum veren sınıf değeri test verisinin sınıf etiketi olarak belirlenir.

G. Karar Ağacı (C4.5)

Karar ağaçları test verilerini bir ağaç yapısın kullanarak sınıflandırmaya çalışır. Ağaçtaki düğümler özellik, dallar özellik değer bilgisini ve yaprak düğümler ise sınıf etiketini belirtmektedir. Karar ağaçlarının yaygın kullanımının nedeni ağaç yapılarının kurallarla ve sade bir şekilde tanımlanabilmesidir. Bu şekilde öğrenilen kurallar kolay bir şekilde aktarılmış olur.

III. SONUÇLAR VE DEĞERLENDİRME

Çalışmamızda sınıflandırıcıların başarı değerleri Tablo 1'de yer alan 36 UCI veri kümesi kullanılarak elde edilmiştir. Tablo 1'deki her bir veri kümesine sırasıyla %5, %10, %25 ve %50 oranında gürültülü veriler eklenerek sınıflandırma gerçeklenmiş ve sınıflandırıcıların gürültülü verilere karşı gürbüzlüğü gözlemlenmiştir. Çalışmada veri kümelerine eklenen gürültüler veri kümelerindeki örneklerin sınıf değerleri değiştirilerek elde edilmiştir. Örneğin %25 gürültü eklenmesi

veri kümesinde yer alan tüm örneklerin %25'inin sınıf değerlerinin değiştirilmesi anlamına gelmektedir.

Tablo.1 UCI Veri Kümeleri

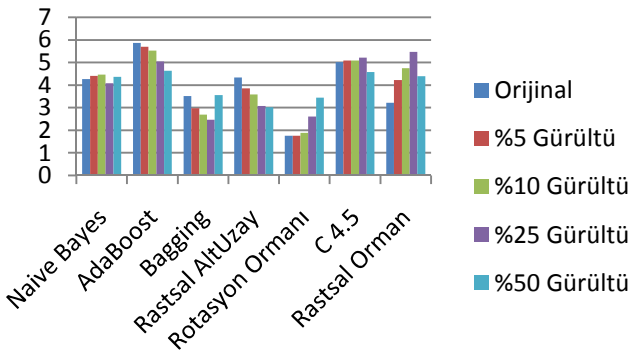
Veri Kümeleri	Öznitelik Sayısı	Sınıf Sayısı	Örnek Sayısı
Abalone	11	19	4153
Anneal	63	4	890
Audiology	70	5	169
Autos	72	5	202
balance-scale	5	3	625
breast-cancer	39	2	286
breast-w	10	2	699
col10	8	10	2019
Colic	61	2	368
credit-a	43	2	690
credit-g	60	2	1000
d159	33	2	7182
Diabetes	9	2	768
Glass	10	5	205
heart-statlog	14	2	270
hepatitis	20	2	155
hypothyroid	32	3	3770
ionosphere	34	2	351
Iris	5	3	150
kr-vs-kp	40	2	3196
Labor	27	2	57
Letter	17	26	20000
lymph	38	2	142
mushroom	113	2	8124
primary-tumor	24	11	302
ringnorm	21	2	7400
segment	19	7	2310
sick	32	2	3772
Sonar	61	2	208
Soybean	84	18	675
Splice	288	3	3190
vehicle	19	4	846
Vote	17	2	435
vowel	12	11	990
waveform	41	3	5000
Zoo	17	4	84

Yedi farklı sınıflandırıcı ve dört farklı gürültü oranlarıyla oluşturulmuş veri kümeleri kullanılarak yapılan sınıflandırma işlemlerinin sonuçları iki farklı şekilde değerlendirilmiştir. İlk olarak yedi farklı sınıflandırıcının her bir veri kümesi üzerindeki sınıflandırma başarıları elde edilmiştir. Bunun için Weka aracı kullanılarak 10 çapraz doğrulama ile yedi farklı aynı veri kümesi için yedi farklı sınıflandırıcı kullanılarak farklı yedi sınıflandırma başarı sonucu çıkarılmıştır. Bir sınıflandırıcı için orijinal ve dört farklı gürültü oranı eklenmiş veri kümeleri üzerinde elde etiketler sınıflandırma başarıları diğer tüm sınıflandırıcıların başarılarıyla karşılaştırılarak bir sıralama gerçekleştirilmiştir. Bir sınıflandırıcı diğer altı sınıflandırıcıdan daha başarılı sınıflandırma sonucu elde etmişse birinci olarak belirtilmiştir. Aynı sınıflandırıcı diğer sınıflandırıcıların başarılı sınıflandırdığı verilerin oranlarına göre bir ile yedi arasında tam sayı değer alabilir. Bu değerler Tablo 2'de gösterilmektedir.

Tablo.2 UCI Veri Kümelerinde Sınıflandırıcıların Ortalama Başarı Değerleri

	Orijinal	%5 Gürültü	%10 Gürültü	%25 Gürültü	%50 Gürültü
Naive Bayes	4,27	4,41	4,47	4,08	4,36
AdaBoost	5,86	5,69	5,52	5,05	4,63
Bagging	3,52	2,97	2,69	2,47	3,55
Rastsal Altuzaylar	4,33	3,86	3,58	3,08	3,02
Rotasyon Ormanı	1,75	1,75	1,89	2,61	3,44
C4.5	5,02	5,08	5,08	5,22	4,58
Rastsak Orman	3,22	4,22	4,75	5,47	4,39

Yedi farklı sınıflandırıcının her bir veri kümesi üzerinde elde ettiği başarı sırasına belirlenerek sınıflandırıcının 36 veri kümesi üzerindeki ortalama başarı değeri çıkartılmıştır. Sınıflandırıcıların 36 UCI veri kümesi üzerinde başarılı sınıflandırma oranlarına göre ortalama sıralaması Şekil 1'de gösterilmiştir.



Şekil.1Farklı Gürültü Oranlara Sahip36 UCI Veri Kümesi Üzerindeki Sınıflandırıcıların Ortalama Başarı Sıralaması

Yukarıda verilmiş şekilde y-ekseni sınıflandırıcıların (sınıflandırıcı toplulukları ve klasik sınıflandırıcılar) ortalama sınıflandırma başarı değerlerini göstermektedir. Bu değerlere göre orijinal verilerde ve gürültü oranı %5 ve %10 gibi az gürültülü veri kümeleri üzerinde en başarılı sınıflandırma sonuçlarının uzak ara Rotasyon Ormanı yaklaşımıyla elde edildiği gözlemlenmiştir. Rotasyon Ormanları diğer sınıflandırıcılara göre ilgili gürültülü veri oranlarında birinciliğe daha yakın çıktığı gözlemlenmiştir. Orijinal ve az gürültülü verilerde Rotasyon Ormanı yaklaşımını Bagging sınıflandırıcı topluluğu takip ettiği gözlemlenmektedir. Gürültü oranının %50 olmasıyla birlikte en iyi ortalama sınıflandırma sonuçlarının Rastgele Altuzaylar sınıflandırıcı topluluğundan elde edilmektedir. Çalışmamızın sonuçları ilgili çalışmalardan

elde edilen sonuçlarla paralellik göstermektedir. Gürültülü veri kümelerin üzerinde Bagging ile elde edilen ortalama başarı değerlerinin AdaBoost ile elde edilen sınıflandırma başarı değerlerinden yüksek olduğu sonucu çıkmaktadır. Aynı zamanda Bagging yaklaşımı genel anlamda temel sınıflandırıcılar olan Naive Bayes ve C4.5'ten daha başarılı sonuçlar vermiştir. Aynı temel sınıflandırıcıların başarı oranları gürültülü veriler üzerinde AdaBoost yaklaşımında daha yüksek çıkmıştır.

Tablo.3 Sınıflandırıcıların Diğer Sınıflandırıcılara İstatistiksel Olarak Üstünlük Sayıları

	Orijinal	%5 Gürültü	%10 Gürültü	%25 Gürültü	%50 Gürültü
Rotasyan Ormanı	69	71	68	53	2
Rastgele Orman	33	18	-1	-49	-21
Bagging	31	34	44	52	25
Rastgele Altuzaylar	29	30	35	47	37
C4.5	3	4	0	-14	-19
Naive Bayes	-53	-52	-45	-17	8
AdaBoost	-112	-105	-101	-72	-32

Tüm sınıflandırıcının UCI veri kümeleri kullanarak elde ettiği sınıflandırma sonuçları birbirleriyle karşılaştırılmalı olarak Tablo 3'te gösterilmiştir. Örneğin Rastgele Ormanlar orijinal UCI veri kümeleri üzerinde tüm sınıflandırıcılara karşı 33 defa istatistiksel olarak üstünlük sağlarken, UCI veri kümelerinde gürültü oranı %50 olduğunda Rastgele Orman diğer sınıflandırıcılara karşı 21 defa kaybetmiştir.

Tablo 3'teki verilere göre gürültü eklenmemiş veya %5, %10 gibi az gürültü veri kümeleri üzerinde Rotasyon Ormanı yaklaşımının diğer sınıflandırıcılara karşı üstünlük sağladığı gözükmektedir. En düşük sınıflandırma başarısı gürültülü verilere karşı hassas olan AdaBoost sınıflandırıcısından elde edilmiştir. Gürültü oranı %25 ve altında Bagging yaklaşımı en iyi sonucu vermekte ama gürültü oranı yüksek olduğunda Bagging yaklaşımında diğer yaklaşımlara üstünlüğü azalmaktadır. Rastgele Altuzaylar sınıflandırıcı topluluğu UCI veri kümelerinde gürültü oranı arttıkça diğer sınıflandırıcılara karşı üstünlüğünün arttığı gözlemlenmektedir. Rastgele Altuzaylar sınıflandırıcısı dışındaki diğer tüm sınıflandırıcılarda gürültü oranları arttıkça sınıflandırıcıların birbirlerine olan üstünlüklerinin azalmıştır. Tablo 3'e göre Naive Bayes sınıflandırıcısının gürültülü veri oranı arttıkça başarı oranının da yükseldiği gözlemlenmiştir.

Naive Bayes ve C4.5 gibi temel sınıflandırıcılarının genel sınıflandırma başarı değerleri ile geri kalan beş farklı sınıflandırıcı topluluğunun sınıflandırma başarı değerleri karşılaştırıldığında AdaBoost dışındaki tüm sınıflandırıcı topluluklarından elde edilen sınıflandırma başarı değerleri temel sınıflandırıcılar olan Naive Bayes ve C4.5 sınıflandırıcılarının başarı değerlerinden yüksek çıkmıştır.

IV. ÇIKARIMLAR VE GELECEK ÇALIŞMALAR

Çalışmamızda çeşitli sınıflandırıcı topluluklarının ve sınıflandırıcıların gürültü verilere karşı sınıflandırma başarıları 36 UCI veri kümesi üzerinde test edilmiştir. Sınıf değerleri değiştirilerek gürültü eklenen veri kümeleri üzerinde genel Adaboost sınıflandırıcı topluluğu dışındaki diğer sınıflandırıcı topluluklarından elde edilen sınıflandırma başarı değerleri temel sınıflandırıcılardan yüksek çıkmıştır. Tablo 2 ve Tablo 3'ten elde edilen değerler birbirleriyle paralellik göstermektedir. İlgili iki tabloya göre gürültüsüz ve az gürültülü veri kümelerinde Rotasyon Ormanı yaklaşımının en iyi sonuç verdiği gözlemlenmiştir. Gürültü oranı arttıkça ise Rastgele Altuzaylar yaklaşımından daha başarılı sonuçlar alınmıştır. Genel bir değerlendirmeye göre gürültülü verilere sırasıyla en gürbüz olan yaklaşımlar Rastgele Altuzaylar ve Bagging olduğu gözlemlenmiştir.

Gelecek çalışmalarda gürültü ekleme işlemi sadece verilerin sınıf etiketlerini değiştirerek değil aynı zamanda özellik değerleri değiştirerek oluşturup sınıflandırıcıların gürültülü verilerde üzerindeki gürbüzlüğü hakkında daha da genel yorumlar yapılacaktır.

KAYNAKÇA

[1] A.A. Felleco, T.M. Khoshgoftaar, J.V. Hulse ve A. Napolitano, "Identifying Learners Robust to Low Quality Data," Information Reuse and Integration (IRI) , pp. 190-195, 2008.

[2] D. Opitz ve R. Maclin, " Popular Ensemble Methods: An Emprical Study", Journal of Artificial Intelligence Research, vol 11,169-198, 1999

[3] P. Melville, N. Shah, L. Mihalkova ve R.J. Mooney, "Experiments on Ensembles with Missing and Noisy Data", Lecture Notes in Computer Science, vol. 3077, 293-302, 2004

[4] L. Breiman, "Random forests", Machine Learning, vol. 45, no. 1, pp.5 -32, 2001.

[5] Juan J. Rodriguez , Ludmila I. Kuncheva , Carlos J. Alonso, "Rotation Forest: A New Classifier Ensemble Method.", IEEE Transactions on Pattern Analysis and Machine Intelligence, v.28 n.10, p.1619-1630, October 2006.

[6] L. Breiman, "Bagging predictors.", Machine Learning,vol. 24, pp. 123–140, 1999.