# Topluluk Algoritması Destekli Yarı-eğiticili Öğrenme

# Semi-supervised Learning Based on Ensemble Algorithm

Abdulkadir Şeker[1], Mehmet Fatih Amasyalı[2]
[1]Bilgisayar Mühendisliği Bölümü, Cumhuriyet Üniversitesi, Sivas, Türkiye
aseker@cumhuriyet.edu.tr
[2]Bilgisayar Mühendisliği Bölümü, Yıldız Teknik Üniversitesi, İstanbul, Türkiye
mfatih@ce.yildiz.edu.tr

***Özetçe*—Eğiticili öğrenmede eldeki etiketli veri miktarının fazla olması tahmin gücü daha yüksek modeller oluşturmamızı sağlar. Etiketli verilerin elde edilmesi etiketsiz verilere göre genelde daha masraflıdır. Eğer etiketsiz veriler elde edilebiliyorsa, yarı eğiticili öğrenme, bu etiketsiz verileri de kullanarak modelin başarısını arttırmayı hedefler. Bu yöntemde, etiketi olmayan veriler, etiketleri tahmin edilerek eğitim kümesine eklenir. Bu çalışmada etiketi olmayan verilerin etiketlerinin tahmininde tekil bir model yerine kolektif öğrenme kullanımının etkisi araştırılmıştır. Denemelerimizde kolektif öğrenme kullanımının başarıya olumlu katkısı açıkça görülmüştür. Bazı veri kümelerinde, kolektif öğrenme destekli yarı-eğiticili öğrenme metodu, tekil bir karar ağacının kullandığı etiketli örnek ile elde ettiği başarıyı yaklaşık 1/3 oranındaki etiketli örnek ile yakalamıştır. Veri kümelerinden etiketi tahmin edilecek elemanlar seçilirken, birbirine yakın elemanlar yerine, daha dağıtık bir şekilde rastgele elemanlar seçmek de başarıyı artıran bir unsur olmuştur.**

***Anahtar Kelimeler — yarı-eğiticili öğrenme; topluluk algoritmaları; kolektif öğrenme; karar ağaçları.***

***Abstract*—In supervised learning, having more amount of labelled data is provided to creating more powerful models for prediction. Obtaining labelled data is more expensive than unlabeled data. If unlabeled data can be accessed, semi-supervised learning aimed to increase success of model as using labelled data. In this method, the unlabeled data add to training set after their labels are predicted. When estimation of unlabeled data, the effect of ensemble learning usage instead of singular model is researched in this study. The positive contribution of ensemble learning usage is obviously seen in our tests. Semi-supervised method based on ensemble algorithm used approximately 1/3 of the labelled samples which a single decision tree used. The two methods obtained close success rate on some datasets. The other impact of increasing success is selecting samples randomly (it means samples are come from more distributed region of dataset.) instead of samples that are closer each other.**

***Keywords — semi-supervised learning; ensemble algorithms; ensemble learning; decision trees.***

## I. GİRİŞ

Eğiticili öğrenme (supervised learning) eğitim yapılan veri setinin çıktılarının bilindiği durumlarda kullanılır. Model eğitim setinin çıktıları doğrultusunda eğitilir ve tahmin yapılır. Sınıflandırma problemleri için bu teknik denenecek olsa, bir grup örneğin hangi sınıftan olduğu ve sınıf sayısı bilinmesi gerekmektedir. Eğiticisiz öğrenmede (unsupervised learning) ise eğitim verisinin çıktıları belli değildir. Sınıf sayısının ve örneklerin hangi sınıfa ait olduğu bilgisi olmadan ilerleyen kümeleme problemleri bu teknik ile çözülmektedir. Bu iki tekniğin hibritleşmesinden ise yarı-eğiticili öğrenme (semi-supervised learning) tekniği ortaya atılmıştır. Bu teknikte, modelin oluşturulmasında etiketli verilerin yanında etiketsiz veriler de kullanılmaktadır. Burada etiketli veri miktarının, etiketsiz veriye göre çok az olduğu düşünülmektedir. Az sayıda etiketlenmiş veriden bir model oluşturulur, bu model ile etiketsiz verilerin etiketleri tahmin edilir. Gerçek ve tahmin edilmiş etiketlere sahip verilerle sonuç model oluşturulur test kümesi üzerindeki performansı ölçülür [1].

Kolektif veya topluluk öğrenme (ensemble learning) algoritmalarında, tek bir öğrenici yerine birden fazla öğrenici eğitilir ve kararları birleştirilir. Bu yöntemlerde kabaca bir problem için uzman birden fazla kişinin çözümlerinin birleştirilmesinin tek kişinin çözümüne göre daha güçlü olduğu düşünülmektedir. Kararların birleştirilmesi işleminde oy birliği, demokrasi, ağırlıklı karar verme, vb. yöntemler kullanılmaktadır [2].

Bu çalışmada yarı-eğiticili öğrenmenin temel yapısı ile birkaç farklı kolektif öğrenme algoritması karşılaştırılmış,

kolektif öğrenmenin yarı eğiticili öğrenme üzerindeki etkisi incelenmiştir.

## II. KULLANILAN YÖNTEM VE VERİ KÜMELERİ

### A. Yarı-eğiticili Öğrenme

Yarı-eğiticili öğrenme, etiketli verinin yanında etiketsiz olanların da kullanıldığı bir öğrenme tekniğidir. Genellikle veri setinde etiketli örneklerin sayısının, etiketsizlere göre çok az miktarda olduğu durumlarda kullanılır. Etiketli veri bulmak veya veriyi etiketlemek maliyetli bir iş olduğundan çoğu çalışma için fazla miktarda etiketsiz veri bulunmaktadır. Bu etiketsiz verilerin etiketleri az miktardaki etiketli veriyle oluşturulan model ile tahmin edilip eğitim kümesine katıldığında test kümesi üzerindeki başarıda kayda değer bir ilerleme görülebilmektedir [3].

Etiketi olmayan verilerin etiketleri tek bir seferde tahmin edilebildiği gibi, iteratif bir süreçle de tahmin edilebilir.

Tek adımlı tahminde, etiketli verilerden bir model oluşturulup tüm etiketsiz veriler bu modele göre etiketlenir. İteratif yöntemde ise, her bir adımda tüm etiketli veri değil bir kısmının etiketi tahmin edilir ve etiketi tahmin edilenler eğitim kümesine dahil edilir, bu yeni eğitim kümesiyle yeni bir model oluşturulur. Ardından bu modelle geriye kalan etiketsiz verilerin yine bir kısmının etiketleri tahmin edilir. Bu süreç etiketsiz veri kalmayana kadar ya da etiketi güvenle tahmin edilebilecek etiketsiz veri kalmayana kadar devam eder [4].

### B. Kolektif Öğrenme (Ensemble Learning)

Makine öğrenmesinde, kolektif metotlar daha iyi tahmin performansı için tek bir öğrenme algoritması yerine birden çok öğrenme algoritması kullanırlar [5].

Kolektif öğrenme, aynı veri kümesi üzerinde farklı algoritmaların çalıştırılmasıyla gerçekleştirilebildiği gibi aynı algoritmayı farklı örneklerle çalıştırmakla da gerçekleştirilir. Bir karar ağacını ele alacak olursak, farklı örnekler ile oluşturulan her ağaç birbirinden farklı olacaktır. Bu sayede hepsinin kendine has kararları çıkacak, bu kararlar bir şekilde birleştirilirse tekil ağaçtan daha güçlü (doğruluğu yüksek) bir model ortaya çıkacaktır [6].

### C. Veri Seti

Bu çalışmada algoritmaların kıyaslanması için farklı alanlardan toplanan 36 veri seti üzerinde çalışılmıştır. Veri kümeleri UCI (University of California, Irvine) veri havuzundan alınmıştır [7]. UCI veri kümeleri, makine öğrenmesi çalışmalarında algoritmaların analizini yapabilmek ve yeni geliştirilenler ile karşılaştırabilmek için akademik çalışmalarda çoğunlukla tercih edilmektedir [8].

## III. DENEYSEL SONUÇLAR

Kullanılan veri kümelerindeki tüm örneklerin etiketleri mevcuttur. Yarı-eğiticili öğrenmeyi simule edebilmek için veri kümesi eğitim ve test olarak ikiye ayrılmış, daha sonra eğitim kümesindeki elemanların bir kısmının etiketleri silinmiştir. Ardından yarı-eğiticili öğrenme ile bu etiketler tahmin edilmiştir. Gerçek etiketli ve tahmini etiketli verilerle oluşturulan sonuç modelin başarısı test kümesi üzerinde ölçülmüştür. Bu işlem 5*2 çapraz geçerleme ile yapılmış ve yapılan 10 deneyin başarı yüzdelerinin ortalaması, genel başarı olarak kabul edilmiştir.

Sonuç ve ara modellerin oluşturulmasında tekil algoritmalar yerine kolektif öğrenme algoritmalarının kullanımının etkisi nasıldır? Çok adımlı (iteratif) yöntemlerde etiketi tahmin edilecek örnekler nasıl seçilmelidir? Bu sorulara cevap vermek için çeşitli deneyler tasarlanmıştır. Bu bölümün devamında bu deneyler ve sonuçları anlatılmıştır.

### A. Etiketi Tahmin Edilecek Verinin Seçiminin Etkisi

Yarı-eğiticili öğrenme az miktarda etiketli veri olan veri kümeleri için önerilen bir tekniktir. Bu sebeple çalışmada elimizdeki veri kümelerinden rastgele %10'luk bir kısmını alınarak modeller oluşturulmuştur. İlk olarak çalışılan tekil yarı-eğiticili öğrenme tekniğinde (SRS: Semi-Supervised-Randomly Singular), %10'luk etiketli veri ile eğitim verisinin kalan %90'lık[1] bölümü tahmin edilmiştir.

Diğer yöntemlerde ise adım adım gidilerek her iterasyonda %10'luk veri için etiketleme yapılmış, %90'a ulaşana kadar veri kümesi genişletilerek öğrenme gerçekleştirilmiştir.

Etiketi tahmin edilecek veri seçiminde başlangıç noktası için 2 farklı yol denenmiştir. Birincisinde rastgele seçilen bir örnek ile (birinci yol), ikincisinde veri setinin merkezine en yakın olan örnekle (ikinci yol) başlanmıştır.

Rastgele seçilen örnek ile (birinci yol) başlanan teknik için 4 farklı algoritma çalıştırılmıştır:

i. Rastgele seçilen örnekler ile eğitim %90'lık bölüme gelene kadar eğitim yapılmış, sonra teste geçilmiştir. (SR: Semi-Supervised-Randomly Selected Elements)

ii. Rastgele seçilen örnekler ile eğitim başlatılmış, her iterasyonda, genişletilmiş veri ile bir model oluşturulmuş, yapılan tahminler kaydedilmiş ve demokrasi (oy çokluğu) yöntemiyle birleştirilmiştir. (SRES: Semi-Supervised-Randomly Selected Elements Ensemble)

iii. Seçilen ilk örneğe en yakın elemanlar seçilerek 'i' algoritmasındaki gibi çalışılmıştır. (SREU: Semi-Supervised Randomly Selected Elements Sorted with Euclidean Distance)

[1] Verinin tamamı ile %90'lık kısmı kullanılarak yapılan denemelerde birbirine yakın sonuçlar çıktığı görülmüş, hız kazanmak için bu miktarın yeterli olduğu düşünülmüştür.

iv. Seçilen ilk örneğe en yakın elemanlar ile eğitim başlatılmış, 'ii' algoritması gibi topluluk öğrenmesi gerçekleştirilmiştir. (SREUES: Semi-Supervised Randomly Selected Elements Sorted Euclidean Distance Ensemble)

Etiketlenecek verilerin seçiminde başlangıç noktası olarak veri kümesinin merkez noktasına en yakın eleman bulunarak (ikinci yol) 2 farklı algoritma denenmiştir:

v. Seçilen ilk örneğe en yakın elemanlar ile model oluşturulmuştur. (SCEU: Semi-Supervised Centroid Selected Elements Sorted Euclidean Distance)

vi. Seçilen ilk örneğe en yakın elemanlar ile kolektif öğrenme yapılmış, model oluşturulmuştur. (SCEUES: Semi-Supervised Centroid Selected Elements Sorted Euclidean Distance Ensemble)

Tablo 1'de 36 UCI veri kümesi üzerinde 6 algoritmanın ortalama başarı yüzdeleri (10 deneme ortalaması) verilmiştir. En son satırda ise algoritmaların topluluk öğrenmesi versiyonları ile kendi içlerindeki sonuçları karşılaştırılmış kaç kez galip geldikleri verilmiştir. Örneğin, SR algoritması ve onun topluluk öğrenmesi versiyonu SRES incelendiğinde; 36 veri kümesinden 15'inde SR, 21'inde ise SRES algoritmasının kazandığı görülmektedir.

| | SR | SRES | SREU | SREUES | SCEU | SCEUES |
|---|---|---|---|---|---|---|
| *iris* | 46.40 | 41.47 | 33.33 | 29.86 | 33.33 | 33.33 |
| *anneal* | 86.99 | 87.24 | 75.64 | 75.07 | 77.10 | 76.94 |
| *audiology* | 38.81 | 32.26 | 27.85 | 29.40 | 29.77 | 33.69 |
| *autos* | 39.60 | 41.78 | 27.42 | 34.65 | 35.64 | 33.26 |
| *balance-scale* | 66.22 | 69.20 | 54.45 | 59.55 | 67.62 | 66.92 |
| *breast-cancer* | 65.10 | 59.16 | 58.61 | 58.46 | 62.02 | 58.53 |
| *breast-w* | 90.74 | 91.46 | 61.17 | 64.49 | 79.05 | 72.26 |
| *col10* | 66.00 | 65.39 | 28.27 | 23.44 | 34.60 | 32.15 |
| *colic* | 62.34 | 72.72 | 75.32 | 62.11 | 75.92 | 74.34 |
| *credit-a* | 78.55 | 79.10 | 73.65 | 72.49 | 80.92 | 79.44 |
| *credit-g* | 63.54 | 67.20 | 59.64 | 66.82 | 61.50 | 62.74 |
| *d159* | 89.34 | 90.56 | 87.62 | 83.93 | 87.14 | 85.93 |
| *diabetes* | 66.88 | 65.62 | 54.03 | 59.16 | 59.47 | 60.33 |
| *glass* | 43.92 | 40.78 | 22.64 | 28.92 | 39.80 | 45.09 |
| *heart-statlog* | 70.15 | 67.48 | 62.51 | 57.85 | 56.29 | 56.96 |
| *hepatitis* | 77.27 | 79.61 | 74.67 | 76.88 | 79.22 | 79.87 |
| *hypothyroid* | 97.82 | 97.33 | 96.14 | 95.78 | 96.53 | 96.31 |
| *ionosphere* | 74.40 | 76.91 | 62.91 | 69.65 | 66.16 | 69.48 |
| *abalone* | 20.70 | 20.79 | 13.42 | 13.98 | 12.79 | 13.38 |
| *kr-vs-kp* | 93.25 | 93.19 | 83.39 | 84.08 | 88.11 | 87.30 |
| *Labor* | 56.79 | 55.36 | 51.07 | 56.78 | 54.28 | 57.50 |
| *letter* | 60.89 | 62.31 | 27.95 | 27.31 | 27.43 | 28.50 |
| *lymph* | 56.62 | 57.46 | 52.11 | 58.59 | 54.36 | 55.63 |
| *mushroom* | 99.34 | 99.29 | 51.47 | 57.04 | 68.99 | 81.88 |
| *primary-tumor* | 22.72 | 26.16 | 23.24 | 17.28 | 23.84 | 24.30 |
| *ringnorm* | 80.40 | 82.35 | 73.11 | 64.41 | 53.47 | 52.88 |
| *segment* | 86.88 | 85.95 | 24.38 | 24.95 | 36.67 | 39.47 |
| *sick* | 96.65 | 96.62 | 95.14 | 95.69 | 94.37 | 94.97 |
| *sonar* | 63.46 | 60.38 | 63.36 | 63.94 | 69.13 | 66.15 |
| *soybean* | 41.10 | 45.70 | 27.01 | 23.70 | 20.65 | 23.47 |
| *splice* | 82.59 | 84.92 | 80.27 | 79.03 | 76.62 | 72.49 |
| *vehicle* | 53.17 | 48.89 | 34.34 | 30.00 | 35.41 | 42.57 |
| *vote* | 92.26 | 92.26 | 63.45 | 63.91 | 80.51 | 82.71 |
| *vowel* | 34.93 | 37.03 | 26.96 | 22.80 | 28.36 | 29.11 |
| *waveform* | 67.74 | 67.94 | 54.65 | 57.63 | 64.78 | 66.06 |
| *zoo* | 54.52 | 49.52 | 32.61 | 21.19 | 27.85 | 32.38 |
| | 15 | 21 | 17 | 19 | 15 | 21 |

**Tablo 1**. Algoritmaların 36 veri kümesi üzerinde başarı (%) sonuçları

Tablo 1 incelendiğinde, modeller kolektif öğrenme durumlarına göre kendi aralarında gruplandırılmıştır. En başarılı[2] algoritmanın rastgele elemanların kolektif öğrenme ile oluşturulduğu model (SRES) olduğu görülmüştür. Diğer algoritmalar kendi aralarında kıyaslandığında ise beklendiği gibi topluluk (ensemble) olanlar, tekil olanlara oranla daha başarılı çıkmışlardır. Ayrıca, rastgele örnek seçmek daha başarılı modeller üretmiştir. Bunun sebebi, rastgele örneklerin en yakın örneklere göre veri setini daha kapsayıcı bir şekilde temsil etmiş olması olabilir.

İlk aşamadan sonra ise yarı eğiticili öğrenmenin etkisini görmek için testler yapılmıştır. Tablo 2'de alınan sonuçlar gösterilmiştir. Bu tablodaki ilk algoritma yarı-eğiticili öğrenme ile verinin sadece %10 kısmı ile oluşturulan tekil bir ağaç (SRS), ikinci rastgele elemanlar ile adım adım ilerleyen yarı-eğiticili öğrenme tekniği olan (SR) ve bu tekniğin topluluk öğrenmesi destekli versiyonu (SRES), sonuncu algoritma ise bütün eğitim verisini (hepsi etiketli) kullanan klasik bir karar ağacı (CDT: Classic Decision Tree) olacaktır. Burada beklenen sonuç, topluluk destekli yarı-eğiticili bir algoritmanın kabaca klasik yarı-eğiticili algoritma ile klasik karar ağacı algoritmasının arasında bir başarı göstermesidir. Ayrıca, SRES algoritmasının, bütün eğitim verisini kullanan karar ağacına (CDT) yakın sonuçlar üretebilen bir yarı-eğiticili algoritma olup olmadığını da incelenmiştir.

Tablo 2'nin en son satırına bakıldığında ilk üç sütunda CDT hariç diğer algoritmaların kendi aralarındaki sonuçlarına ait olan, son sütununda ise CDT algoritmasının diğer üç algoritmaya karşı olan skorları verilmiştir.

Tablo 2'den çıkan sonuçlara göre etiketli veri kümesinin tamamının kullanıldığı algoritma iki veri kümesi hariç (hepatitis ve abalone) hepsinde diğerlerini yenmiştir. Diğer 3 algoritma kendi arasında kıyaslandığında ise SRES daha başarılı çıkmıştır. Topluluk destekli algoritma olan SRES diğerlerine üstünlük sağlamıştır. SRES ile CDT algoritmaları kanser (breast-w), kredi notu (credit-a, credit-g) ve hepatit (hepatitis) veri kümelerinde çok yakın sonuçlar vermiştir. Bu da kolektif öğrenmenin gücünü göstermiştir.

[2] SCEUES ile SRES algoritmalarının ikisi de 21 kez kazanmış fakat kendi aralarında kıyaslama yapıldığında SRES daha başarılı olduğu görülmüştür.

| | SRS | SR | SRES | | CDT |
|---|---|---|---|---|---|
| iris | 32.27 | 46.40 | 41.47 | | 94.67 |
| anneal | 88.29 | 86.99 | 87.24 | | 98.52 |
| audiology | 30.36 | 38.81 | 32.26 | | 84.17 |
| autos | 34.65 | 39.60 | 41.78 | | 63.76 |
| balance-scale | 66.89 | 66.22 | 69.20 | | 78.21 |
| breast-cancer | 66.29 | 65.10 | 59.16 | | 66.50 |
| breast-w | 90.66 | 90.74 | 91.46 | | 93.35 |
| col10 | 65.28 | 66.00 | 65.39 | | 75.88 |
| colic | 67.39 | 62.34 | 72.72 | | 80.33 |
| credit-a | 78.14 | 78.55 | 79.10 | | 81.80 |
| credit-g | 65.86 | 63.54 | 67.20 | | 68.98 |
| d159 | 90.27 | 89.34 | 90.56 | | 97.31 |
| diabetes | 67.71 | 66.88 | 65.62 | | 69.95 |
| glass | 42.06 | 43.92 | 40.78 | | 66.08 |
| heart-statlog | 67.26 | 70.15 | 67.48 | | 75.19 |
| hepatitis | 79.87 | 77.27 | 79.61 | | 77.53 |
| hypothyroid | 96.90 | 97.82 | 97.33 | | 99.47 |
| ionosphere | 78.80 | 74.40 | 76.91 | | 87.54 |
| abalone | 21.60 | 20.70 | 20.79 | | 21.39 |
| kr-vs-kp | 93.25 | 93.25 | 93.19 | | 99.06 |
| Labor | 61.43 | 56.79 | 55.36 | | 82.14 |
| letter | 61.01 | 60.89 | 62.31 | | 82.20 |
| lymph | 50.56 | 56.62 | 57.46 | | 77.89 |
| mushroom | 99.31 | 99.34 | 99.29 | | 99.97 |
| primary-tumor | 26.56 | 22.72 | 26.16 | | 43.44 |
| ringnorm | 81.55 | 80.40 | 82.35 | | 88.47 |
| segment | 86.88 | 86.88 | 85.95 | | 94.15 |
| sick | 96.47 | 96.65 | 96.62 | | 98.28 |
| sonar | 56.06 | 63.46 | 60.38 | | 68.08 |
| soybean | 38.66 | 41.10 | 45.70 | | 89.02 |
| splice | 84.34 | 82.59 | 84.92 | | 92.17 |
| vehicle | 51.89 | 53.17 | 48.89 | | 66.97 |
| vote | 92.30 | 92.26 | 92.26 | | 95.25 |
| vowel | 34.10 | 34.93 | 37.03 | | 67.86 |
| waveform | 69.12 | 67.74 | 67.94 | | 74.00 |
| zoo | 45.24 | 54.52 | 49.52 | | 95.95 |
| | 9 | 12 | 14 | | 34 |

**Tablo 2**. En iyi algoritma ve türevleri ile tekil karar ağacı algoritmasının başarı (%) sonuçları

### *B. Etiketli Veri Miktarının Etkisi*

Son olarak yapılan test ile SRES algoritmasın için etiketli veri miktarının etkisi gözlemlenecektir. Etiketli veri miktarı bütün verinin %10'u ile %50'si arasında olmasının etkisine bakılmıştır. SRES algoritması üzerinde yapılan testlerin sonucu aşağıda verilmiştir. Tablo 3 incelendiğinde beklenildiği gibi veri setindeki etiketli veri miktarı arttıkça başarı yüzdeleri de artmaktadır.

| | %10 | %20 | %30 | %40 | %50 |
|---|---|---|---|---|---|
| iris | 52.53 | 69.07 | 88.13 | 91.87 | 93.07 |
| anneal | 86.79 | 94.94 | 96.92 | 98.02 | 97.44 |
| audiology | 25.60 | 65.60 | 73.45 | 81.19 | 79.40 |
| autos | 34.85 | 35.05 | 47.82 | 51.29 | 51.19 |
| balance-scale | 64.04 | 71.19 | 74.10 | 75.13 | 74.94 |
| breast-cancer | 63.50 | 66.57 | 67.06 | 64.41 | 66.99 |
| breast-w | 92.44 | 92.38 | 92.18 | 93.87 | 94.07 |
| col10 | 64.84 | 69.71 | 70.61 | 72.29 | 73.76 |
| colic | 64.40 | 71.25 | 77.55 | 77.39 | 79.57 |
| credit-a | 79.91 | 82.03 | 80.32 | 82.29 | 84.17 |
| credit-g | 65.92 | 66.68 | 68.36 | 66.76 | 66.48 |
| d159 | 90.06 | 92.36 | 93.96 | 94.95 | 95.49 |
| diabetes | 68.65 | 68.07 | 68.05 | 69.97 | 69.71 |
| glass | 41.86 | 46.76 | 55.88 | 55.20 | 57.75 |
| heart-statlog | 67.48 | 70.30 | 66.30 | 73.04 | 72.22 |
| hepatitis | 78.96 | 75.45 | 78.05 | 76.75 | 75.45 |
| hypothyroid | 96.83 | 98.74 | 98.81 | 99.09 | 99.14 |
| ionosphere | 77.09 | 84.40 | 85.26 | 87.20 | 87.20 |
| abalone | 20.59 | 20.72 | 21.76 | 21.38 | 21.28 |
| kr-vs-kp | 93.13 | 94.71 | 96.14 | 96.63 | 97.44 |
| Labor | 56.07 | 66.43 | 62.14 | 70.36 | 35.71 |
| letter | 61.82 | 69.05 | 72.55 | 75.34 | 76.63 |
| lymph | 60.14 | 67.75 | 73.24 | 71.27 | 75.21 |
| mushroom | 99.19 | 99.52 | 99.74 | 99.85 | 99.85 |
| primary-tumor | 24.64 | 25.96 | 31.26 | 32.78 | 39.80 |
| ringnorm | 82.17 | 83.77 | 86.64 | 86.27 | 86.47 |
| segment | 86.07 | 89.71 | 90.59 | 91.53 | 92.60 |
| sick | 96.30 | 97.31 | 97.65 | 97.72 | 97.83 |
| sonar | 55.10 | 60.77 | 65.10 | 63.27 | 67.31 |
| soybean | 43.00 | 58.43 | 66.05 | 74.96 | 77.45 |
| splice | 83.30 | 88.63 | 90.77 | 90.50 | 91.54 |
| vehicle | 48.87 | 58.23 | 59.65 | 64.68 | 63.48 |
| vote | 91.98 | 94.61 | 93.23 | 95.58 | 94.65 |
| vowel | 35.37 | 42.95 | 51.35 | 55.45 | 58.04 |
| waveform | 70.12 | 71.35 | 71.38 | 72.92 | 73.07 |
| zoo | 32.38 | 57.62 | 72.14 | 67.38 | 48.81 |
| | 1 | 0 | 4 | 8 | 21 |

**Tablo 3**. Etiketli veri miktarının başarıya (%) etkisi

## IV. SONUÇ

Yapılan denemelerde, yarı-eğiticili öğrenmede, etiketi olmayan verilerin etiketlerinin tahmininde tekil bir model kullanmak yerine kolektif öğrenme kullanıldığında daha başarılı sonuçlar alındığı görülmüştür. Veri kümelerinin çoğunda %10'luk veri ile başlayan bir topluluk öğrenme destekli yarı-eğiticili öğrenme algoritmasının (SRES), verinin %30-%40 arasındaki miktarının kullanan bir tekil karar ağacına yakın değerler ürettiği görülmüştür.

Etiketi tahmin edilecek örnekleri seçerken, rastgele seçmenin, kümenin daha geniş bir dağılımdan oluşmasını sağlayarak, daha başarılı sonuçlar verdiği görülmüştür.

Bu çalışmanın verdiği sonuçlara dayanarak, topluluk öğrenmesinin yarı-eğiticili öğrenmedeki başarısına etkisini artıracak yeni deneyler planlanmaktadır. Bu etkinin ne tür veri kümelerinde fazla/az olduğu ayrıca araştırılacak bir konu olabilir. Ayrıca veri setinden rastgele elemanlar

seçmenin daha başarılı olduğu görülmesine karşın, mevcut modele daha yakın örneklerin daha başarılı tahmin edilebileceği fikri üzerine de gidilebilir.

## KAYNAKÇA


[1] Altun, Y., McAllester, D., & Belkin, M. ``Maximum margin semi-supervised learning for structured variables." *Advances in Neural Information Processing Systems (NIPS) 18*, 2005.

[2] Dietterich T. G., *Multiple classifier systems,* Springer, 2000.

[3] Zhu, X., "Semi-Supervised Learning Literature Survey" , Computer Sciences Technical Report 1530, University of Wisconsin–Madison, 2005.

[4] Hanneke, S., Roth, D., "Iterative Labeling for Semi-Supervised Learning", 2004.

[5] Opitz, D., Maclin, R., "Popular ensemble methods: An empirical study". Journal of Artificial Intelligence Research. Vol. 11, 1999, p 169–198.

[6] Banfield, R. E., Hall, L. O., Bowyer, K. W., & Kegelmeyer, W. P., ``A comparison of decision tree ensemble creation techniques", *IEEE transactions on pattern analysis and machine intelligence,* 29(1):173-180, 2007.

[7] Blake, C. L., Merz, C. J., UCI repository of machine learning databases, 1998.

[8] UCI Machine Learning Repository: About. (n.d.). Retrieved September 09, 2016, from https://archive.ics.uci.edu/ml/about.html