Aktif Öğrenmede Komitelerin Kullanımı ve Başlangıç Kümesinin Seçimi

Active Learning with Committees and the Selection of Starting Sets

Cem Agan
Bilgisayar Mühendisliği Bölümü
Yıldız Teknik Üniversitesi
İstanbul, Türkiye
cemagan@hotmail.com

M. Fatih Amasyalı
Bilgisayar Mühendisliği Bölümü
Yıldız Teknik Üniversitesi
İstanbul, Türkiye
mfatih@ce.yildiz.edu.tr

***Özetçe*— Etiketli eğitim verisi temin etmek uzun süren maliyetli bir iştir. Aktif öğrenme, makine öğrenmesi algoritmalarının makul başarı oranlarına daha az etiketli eğitim örneği ile ulaşabilmesini amaçlar. Bu amaçla öncelikli olarak hangi örneklerin etiketlerine ihtiyaç duyulduğunun belirlenmesinde kullanılan yöntemlerden biri sınıflandırıcı topluluklarının kararlarından faydalanmaktır. Çalışma kapsamında topluluk tabanlı bir aktif öğrenme uygulaması gerçeklenmiştir ve performansı aktif olmayan öğrenme yöntemleriyle kıyaslanmıştır.**

Anahtar Kelimeler — aktif öğrenme; sınıflandırıcı toplulukları.

***Abstract*— Obtaining tagged training data takes a long time and also is a costly task. Active learning aims machine learning algorithms achieve reasonable accuracies with less tagged training data. To this purpose, one of the methods for determining which samples to be tagged is making use of the decisions of classifier ensembles. Within this work, we implemented a committee-based active learning application and compared it with non-active methods.**

Keywords — active learning; classifier ensembles.

I. GİRİŞ

Eğiticili makine öğrenmesi tekniklerinde yeterli miktarda eğitim örneğinin temin edilmesinin önemi bu konuda çalışanlarca bilinmektedir. Hatta literatürde makine öğrenmesi algoritmalarına ne kadar çok eğitim örneği verilirse o kadar iyi model oluşturabileceğini kanıtlayan araştırmalardan söz edilmektedir [1]. Ancak bilindiği üzere gerçek hayattaki problemlere çözüm geliştirilirken probleme ilişkin oldukça az etiketli eğitim örneği bulunabilmektedir. Bunun nedeni etiketleme işleminin uzmanlar tarafından gerçekleştirilen zaman alıcı ve maliyetli bir iş olmasıdır. Oysa ki bir çok uygulama için etiketsiz veriler oldukça kolay bir şekilde temin edilebilmektedir. Bu durum araştırmacıları daha az etiketli veri ile verimli sınıflandırıcılar üretebilmek için eldeki çok miktardaki etiketsiz örneklerden olabildiğince faydalanmanın yollarını araştırmaya yöneltmiştir.

Araştırmalar çerçevesinde ortaya konan başlıca fikirlerden birisi aktif öğrenmedir [2]. Kabaca ifade edilirse, aktif öğrenmede az sayıda etiketli eğitim örneği ile başlanarak her adımda etiketsiz örnekler mevcut modelle incelenir ve etiketsiz örneklerden mevcut modelin iyileştirilmesinde en fazla katkı sağlayacak olanlar belirlenerek uzmandan bu örneklerin etiketleri sorulur. Aktif öğrenme için bugüne kadar çeşitli yöntemler önerilmiştir. Bunlardan birisi Komite ile Sorgulama (Query by Committee) [3] yöntemidir. Bu yöntemde çoklu sınıflandırıcıların etiketsiz örneklerin sınıflarına dair verdikleri kararlardaki ihtilafa bakılır ve bu örneklerden hangisinin temsil gücü en yüksek (en bilgilendirici) örnek olduğuna karar verilip ona göre sorgulama yapılır. Komitedeki ihtilafın hesaplanmasında çeşitli anlaşmazlık ölçüleri kullanılmaktadır.

Biz bu çalışmamızda aktif öğrenmede toplulukların (komitelerin) kullanımını ve gerçeklemede performansı etkileyen bazı detayları araştırdık. Bu amaçla komite bazlı sorgulama yapan bir aktif öğrenme uygulaması gerçekleyerek performansını bazı standart verisetleri üzerinde pasif öğrenmeyle kıyaslamalı olarak değerlendirdik. Yazının ikinci bölümünde aktif öğrenme ve sorgulama stratejilerine dair literatür bilgisini sunduk. Üçüncü bölümde komite bazlı sorgulama, komitelerin oluşturulması ve kullanılan anlaşmazlık ölçülerine dair temel bilgiler verildi. Bir sonraki bölümde uygulamamızda izlediğimiz yöntem, kullandığımız verisetleri ve test parametrelerini açıkladık ve elde ettiğimiz deneysel sonuçları verdik. Sonuç bölümünde ise araştırmamızın genel sonuçlarını ve ileride yapmayı planladığımız çalışmaları belirterek yazımızı sonlandırdık.

II. AKTİF ÖĞRENME

Genellikle herhangi bir eğiticili öğrenme algoritmasından başarılı sonuçlar alınabilmesi için eğitim işleminin yüzlerce etiketli veri ile gerçekleştirilmesi gerekmektedir. Pek çok eğiticili öğrenme işlerinde etiketli örneklerin temini oldukça zahmetli, vakit alıcı ve maliyetlidir (ör: konuşma tanıma, bilgi çıkarımı, sınıflama vb.) [4].

978-1-4673-5563-6/13/$31.00 ©2013 IEEE

Aktif öğrenmede amaç bir makine öğrenmesi algoritmasının etiketsiz eğitim verisinden seçim yaparak daha az eğitim etiketi ile (daha az veri ile) daha yüksek başarıya ulaşmasını sağlamak ve böylece etiketli veriyi elde etme maliyetini en aza indirmektir. Bu amaçla bir aktif öğrenici, optimal iterasyon sayısında istenen başarıya ulaşabilmek için etiketsiz verilerden etiketlenmesi için bir uzmana (oracle) sorgulama yaparak her adımda etiketlenecek veriyi belirler. Yani etiketleme işlemi rastgele değil aktif öğrenicinin seçimleri doğrultusunda yapılır. Etiketsiz örneklerden öncelikle hangisinin sorgu için seçileceğine karar verilirken bu örneklerin bilgilendiricilikleri (informativeness) hesaplanır. Literatürde bunun için önerilmiş bir çok yöntem bulunmaktadır.

En basit ve yaygın olarak kullanılan sorgu seçim yöntemi belirsizlik örneklemesi (uncertainty sampling) [5] dir. Bu yöntemde aktif öğrenici nasıl etiketleneceği en belirsiz olan örnekleri sorgular. Bu yaklaşım genellikle olasılıksal öğrenme modelleri için doğrudan kullanılabilir. Örneğin iki sınıflı sınıflandırma için bir olasılıksal model kullanılırken posterior olasılığı 0.5'e en yakın örnek sorgulanır. Üç ve daha fazla sınıflı problemlerde ise daha genel bir belirsizlik örneklemesi varyantı kullanılarak en az güvenilir (least confident) tahmine sahip örnek sorgulanır.

Ancak en az güvenilirlik kriteri sadece en olası etiket hakkındaki bilgiyi göz önüne alır ve geriye kalan etiket dağılımından istifade edemez. Bu durumu düzeltmek için bazı araştırmacılar marjin örneklemesi (margin sampling) olarak adlandırdıkları farklı bir çok-sınıflı belirsizlik örneklemesi varyantı kullanmaktadırlar. Buna göre birinci ve ikinci en olası sınıf tahminleri arasındaki olasılık farkı marjini oluşturur ve en küçük marjin değerine sahip örnek en belirsiz örnek olarak belirlenerek sorgulanır. Buna rağmen çok büyük etiket kümesine sahip problemlerde marjin yaklaşımı hala geriye kalan sınıfların çıkış dağılımını göz ardı etmektedir.

Daha genel bir belirsizlik örneklemesi stratejisi ise belirsizlik ölçüsü olarak entropi [6] kullanmaktadır. Bu yaklaşımda en belirsiz örnek belirlenirlenirken olası tüm etiketler için tahminler değerlendirilir.

Literatürde sorgu seçimi için belirsizlik örneklemesi dışında, başka yöntemler de yer almaktadır. Komite ile Sorgulama yönteminde eldeki az miktardaki etiketli verilerle bir sınıflandırıcı topluluğu oluşturulur ve etiketsiz veriler için bu topluluğun üyelerinin sınıf tahminleri alınır. En bilgilendirici örneğin, sınıfı hakkında en fazla ihtilafa düşülen örnek olduğu düşünülür ve sorgulaması yapılır. Beklenen model değişimi (Expected model change) yönteminde, etiketi bilinmesi halinde mevcut modelde en fazla değişime neden olacak örnek sorgulanır. Beklenen hata azaltımı (Expected error reduction), modelin ne kadar değişeceğini değil de genelleştirme hatasının ne kadar düşeceğini ölçer ve hatayı en çok düşüren örnekleri sorgular. Varyans Azaltımı (Variance reduction) ise hesaplama maliyetini düşürmek için çıkış varyansını en aza indiren örnekleri sorgulayarak bu hatayı dolaylı yoldan değerlendirmiş olur. [4]

III. KOMİTE İLE SORGULAMA

Aktif öğrenmede sorgu seçiminde kullanılan yöntemlerden birisi komitelerden yararlanmaktır. Bu stratejide rastgeleleştirilmiş öğrenme algoritmalarından gelen bir dizi hipotez kullanılırak hakkındaki sınflandırma tahminleri etiketlere en çok yayılmış olan (komitenin en çok ihtilafa düştüğü) etiketsiz örnek sorgulama için seçilir. Hipotezler aynı zamanda test örneklerinin sınıflandırılmasında da kullanılırlar.

Komite ile sorgulamada kullanılmak üzere önerilen ilk tekil öğrenme algoritması Gibbs algoritmasıydı [7]. Teorik açıdan iyi incelenmiş olmasına karşın pratikte hesapsal açıdan kontrol edilemez durumdaydı. Bu nedenle daha sonra getirilen yaklaşımlar birden çok sınıflandırıcının eğitilmesini mümkün kılan popüler yöntemleri kullandılar. Abe ve Mamitsuka, Özyüklemeli Kümeleme ile Sorgulama (Query by Bagging) ve İteleme ile Sorgulamanın (Query by Boosting) etiketleme maliyetlerini düşürmede etkili olduğunu göstermiştir [7]. Bu yöntemler aşağıdaki genel plana göre çalışıyordu:

Girişler : Bir öğrenme algoritması A, Bir dizi etiketli eğitim örneği L, Bir dizi etiketsiz eğitim örneği U, Aktif öğrenme iterasyonlarının sayısı k, Her adımda seçilen örnek sayısı m
Aşağıdaki adımları k kere tekrar et :

- Bir sınıflandırıcı topluluğu oluştur
  C* = ToplulukYöntemi(A, L)
- $\forall x_j \in U$ için eldeki topluluk ile
  Bilgi_Değeri(C*, x_j) yi hesapla
- En bilgilendirici m örnekten oluşan S alt kümesini seç
- S için uzmandan etiketleri iste
- S'deki elemanları U'dan sil ve L'ye ekle

Cıkış: ToplulukYöntemi(A, L)

Fig. 1. Komite bazlı sorgulama için genel algoritma planı [7]

Abe ve Mamitsuka'nın sınıflandırıcı komitesini oluşturmakta kullandığı topluluk algoritması bu iki komite ile sorgulama yöntemi arasındaki farkı belirliyordu. İlk yaklaşım olan Özyüklemeli Kümeleme ile Sorgulama eğitim kümesi üzerinde bir kaç kez özyükleme (bootstrap) örneklemesi yaparak çalışıyordu. Öte yandan İteleme ile Sorgulamada, her adımda eğitim örneklerine atanan ağırlık dağılımlarının değiştirilerek bir topluluğun oluşturulduğu daha adaptif bir yaklaşım kullanılmıştır. İki versiyonda da sorgulanacak etiketsiz örnek seçilirken komite üyelerinin etiket tahminindeki ihtilafına bakılmaktaydı.

Melvile ve Monney (2004) ise özel bir meta-learning algoritmasını aktif öğrenme döngüsünün içersine katarak Aktif DECORATE(Active DECORATE)[8] adında başka bir komite ile sorgulama yaklaşımı geliştirmiştir. Bu yaklaşım topluluk içinde çeşitliliği daha yüksek tekil sınıflarıcılar oluşturmak için yapay olarak üretilmiş ek eğitim örnekleri kullanmaktaydı.

Daha önce belirtildiği gibi komite ile sorgulama yöntemlerinde etiketsiz örneklerin komitede ne kadar ihtilafa yol açtığına bakılmaktadır. Hakkında en çok anlaşmazlık olan örneğin en bilgilendirici ve temsil gücü en kuvvetli örnek olduğu düşünülür. Literatürde komitenin ihtilafının belirlenmesinde çeşitli 'anlaşmazlık ölçüleri'nin kullanıldığını görmekteyiz.

Abe ve Mamitsuka [7] marjin kavramını kullanmıştır; buna göre örnek için en çok tahmin edilen sınıf ile en çok tahmin edilen ikinci sınıfın komite tarafından aldığı oyların farkları marjin olarak tanımlanıyordu ve en küçük marjin değerine sahip örneğin en bilgilendirici (en belirsiz) örnek olduğuna karar veriliyordu.

Bir başka anlaşmazlık ölçüsü ise Melvile ile Monney [9] tarafından kullanılan ve olasılık dağılımları arasındaki benzerliğin bir ölçüsü olan Jensen–Shannon uyumsuzluğu (JS-divergence)'dur. Bu ölçüde tekil sınıflandırıcılardan etiketsiz örneklerin sınıflara ait olma olasılıkları alınıp anlaşmazlığın hesaplanmasında kullanılmaktadır.

IV. UYGULAMA VE TEST

Biz de çalışmamızda aktif öğrenmede topluluklardan faydanlanmayı araştırmak ve test etmek için Matlab ortamında bir uygulama gerçekledik. Literatürdeki mevcut komite ile sorgulama yöntemlerinin karşılaştırmalarını içeren yayınlar [10] incelendiğinde standart veri setlerinde şu ana kadar tespit edilen en başarılı yaklaşımın Aktif DECORATE olduğu görülmektedir. Özyüklemeli Kümeleleme ve İteleme ile sorgulamaların ise onu yakından takip ettikleri görülmektedir. Ancak Aktif DECORATE yönteminde aktif öğrenme sürecinin içersinde gerçekleştirilen yapay veri üretimi yöntemin çalışma süresini diğerlerine göre oldukça uzatmaktadır. Biz çalışmamızda hem gerçeklenmesinin kolaylığı, hem çalışma hızı, hem de gürültülü verilere olan dayanıklılığından ötürü komiteleri Özyüklemeli Kümeleme ile oluşturmayı tercih ettik.

Komite anlaşmazlık ölçüleri için gerçekleştirilmiş kıyaslamalı testlere bakıldığında ise kullanılagelen anlaşmazlık ölçülerinin verisetlerinin büyük bir kısmı için anlamlı performans değişimlerine neden olmadığı görülmektedir [10]. Bu durumda JS-uyumsuzluğu gibi hesapsal karmaşıklığı daha yüksek olan ölçülerin kullanılmasının anlamlı kazanç sağlamadığı söylenebilir. Melvile ve Monney yaptıkları araştırmada hem marjin hem de JS- uyumsuzluğu ölçülerini test etmiş ve marjin fikrini baz alan ölçülerin doğrudan karar sınırlarını tanımlamaya yönelimli olduklarını belirtmişlerdir [9]. Bu durumda marjinlerin aktif öğrenme için, sınıf olasılık dağılımlarındaki belirsizliği azaltmaya göre daha uygun olduğu söylenebilir. Bu nedenle biz de çalışmamızda marjin ölçüsünü tercih ettik. Gerçeklemede marjinin genelleştirilmiş versiyonunu kullandık; yani oy sayıları yerine sınıf tahminlerinin olasılık dağılımlarını göz önüne aldık. Topluluk gerçeklemelerinin çoğunda tekil sınıflandırıcılar örneklerin sınıflara aidiyet olasılıklarını üretebilmektedir. $P_{C_{i,y}(x)}$, bir x örneğine C_i tekil sınıflandırıcısı tarafından y sınıfının atanma olasılığını göstermek üzere bu x örneğine C* komitesi tarafından y sınıfının atanmasının olasılığı şöyle tanımlanmaktadır:

$$P_y(x) = \frac{\sum_{C_i \in C^*} P_{C_{i,y}(x)}}{boyut(C^*)} \qquad (1)$$

Gnelleştirilmiş marjin tanımına göre Eşitlik 1 tüm sınıflar için hesaplandığında bunlar içersinden en yüksek olasılık ile en yüksek ikinci olasılık arasındaki fark marjini oluşturmaktadır.

Bölüm III'de komite bazlı sorgulama için, gerçeklememizde de kullandığımız genel bir algoritma planı vermiştik. Bu plana göre algoritmaya giriş olarak hem etiketli hem de etiketsiz örneklerden oluşan iki veri kümesi verilmektedir. Yani ilk modellemenin yapılabilmesi için başlangıçta hali hazirda etiketlenmiş bir miktar veri olmalıdır. Elde bir çok etiketsiz veri varken bunlardan hangilerinin etiketlenerek ilk modellemede kullanılacağı komite bazlı yöntemleri tasarlarken karşımıza çıkan bir sorudur. Bu noktada çok basit bir yaklaşımla bu küme rastgele seçilebileceği gibi bir kümeleme algoritması kulanılarak küme çekirdeklerine en yakın örneklerden seçme yoluna da gidilebilir. Biz gerçeklememizde her iki yolu da deneyerek arada nasıl bir farkın oluştuğunu görmek istedik.

A. Verisetleri

Gerçekliğimiz aktif öğrenme uygulamasının testi için biri Delve [11] dördü UCI [12] veri havuzundan olmak üzere 5 veriseti kullandık. Aktif öğrenmenin etkisini net olarak görebilmek için örnek sayısının mümkün olduğunca büyük olduğu verisetlerini tercih etmeye çalıştık. Aşağıda test için kullandığımız verisetlerinin bazı özellikleri yer almaktadır:

TABLO I. TESTLERDE KULLANILAN VERİSETLERİ

Verisetinin Adı	Örnek Sayısı	Boyut Sayısı	Sınıf Sayısı
d159	7182	33	2
letter	20000	16	26
mushroom	8124	112	2
ringnorm	7400	20	2
spambase	4601	57	2

Testlerde verisetlerinin her birini iki parçaya böldük. Örneklerin yarısını eğitim (ilk modellemede kullanılacak örnekler ile üzerinde aktif öğrenmenin gerçekleştirileceği ve uzmandan etiketlerinin sorgulanacağı etiketsiz örnekler kümesi) diğer yarısın da test için kullandık.

B. Test Koşulları

Test için 'aktif toplu öğren' gerçeklemesinin yanı sıra karşılaştırma yapmak amacıyla 'toplu öğren' (pasif) ve 'tekil öğren' (pasif) gerçeklemeleri de geliştirilmiştir. Tüm testlerde öğrenme iterasyonlarının sayısı eşit ve 50 olarak belirlenmiştir. Aktif toplu öğren ve toplu öğren yöntemleri için topluluk yöntemi olarak Özyüklemeli Kümeleme kullanılmıştır ve topluluk boyutu 5 olarak seçilmiştir. Her yöntem test edilirken ilk modelleme kümesinin seçimi için kmeans kümeleme kullanılmıştır. Bunun nedeni kıyaslamanın sağlıklı yapılabilmesi için yöntemlerin birbirlerine yakın başlangıç kümeleriyle başlamalarını sağlamaktır.

Ayrıca başlangıç kümesinin rastgele ve çeşitli kümeleme yöntemleriyle seçimi arasındaki değişimi görebilmek adına 'aktif toplu öğren' başlangıç kümesi rastgele ve dört farklı kümeleme yöntemiyle seçilecek şekilde verisetleri üzerinde test edilmiştir.

C. Deneysel Sonuçlar

Tablo II'de aktif öğrenme ve aktif olmayan öğrenme gerçeklemelerinin beş standart veri kümesindeki artan eğitim örneği sayıları için başarı oranları verilmiştir.

Tablo III'de ise aktif öğrenme gerçeklemesinin farklı başlangıç kümesi oluşturma koşulları için beş veri kümesinde elde ettiği başarılar gösterilmiştir.

TABLO II. AKTİF-PASİF ÖĞRENME TEST SONUÇLARI

D159				LETTER				MUSHROOM			
eös	**atop**	**top**	**tek**	**eös**	**atop**	**top**	**tek**	**eös**	**atop**	**top**	**tek**
144	88.1	82.8	82.4	**382**	58.0	55.2	47.7	**161**	97.9	97.5	99.5
216	88.7	83.9	86.3	**582**	61.9	61.6	55.7	**242**	99.7	96.3	99.5
288	93.3	86.4	90.2	**782**	63.6	65.3	60.6	**323**	99.6	98.8	99.8
360	91.6	89.2	90.2	**982**	67.9	69.7	63.1	**404**	99.9	98.7	99.5
432	91.3	89.6	91.1	**1182**	70.2	69.7	64.6	**485**	100.0	98.6	99.8
504	93.8	94.2	91.7	**1382**	73.1	73.5	67.9	**566**	100.0	99.3	99.6
576	93.6	92.1	91.1	**1582**	73.4	74.6	68.2	**647**	100.0	99.2	99.6
648	94.5	94.3	93.3	**1782**	76.6	75.3	69.7	**728**	100.0	99.7	99.6
720	96.4	94.9	94.3	**1982**	77.2	75.9	68.8	**809**	100.0	99.6	99.6
792	95.8	94.7	93.3	**2182**	78.7	76.5	71.1	**890**	100.0	99.6	99.7
864	95.9	92.4	93.7	**2382**	78.5	77.7	70.9	**971**	100.0	99.8	99.7
936	96.9	94.4	93.2	**2582**	79.3	77.7	71.0	**1052**	100.0	99.5	99.7
1008	95.6	93.7	93.0	**2782**	80.7	79.1	72.4	**1133**	100.0	99.7	99.7
1080	96.3	92.9	91.4	**2982**	80.7	80.0	72.7	**1214**	100.0	99.7	99.7
1152	96.0	95.1	92.2	**3182**	81.6	80.1	73.8	**1295**	100.0	99.7	99.8
1224	96.8	95.7	92.9	**3382**	83.2	80.7	73.8	**1376**	100.0	99.9	99.8
1296	97.3	95.3	92.8	**3582**	83.3	80.5	74.1	**1457**	100.0	99.9	99.8
1368	95.5	96.9	92.8	**3782**	84.6	81.1	75.7	**1538**	100.0	100.0	99.8
1440	98.1	96.4	93.6	**3982**	84.5	81.9	74.4	**1619**	100.0	99.8	99.9
1512	97.7	96.2	93.1	**4182**	86.0	81.4	75.3	**1700**	100.0	99.7	99.9

RINGNORM				SPAMBASE			
eös	**atop**	**top**	**tek**	**eös**	**atop**	**top**	**tek**
148	84.2	85.5	80.0	**92**	86.1	85.3	78.4
222	85.1	88.0	78.0	**138**	87.2	87.6	84.1
296	88.3	87.7	80.4	**184**	85.0	87.8	75.8
370	89.9	87.7	79.3	**230**	88.7	87.5	81.0
444	88.1	86.8	83.5	**276**	88.5	87.5	86.9
518	90.0	88.6	84.3	**322**	89.1	87.9	87.4
592	89.8	89.4	84.3	**368**	89.9	87.5	88.3
666	90.1	88.8	83.4	**414**	90.5	89.2	88.6
740	91.5	88.3	84.1	**460**	90.8	88.9	87.6
814	90.0	89.0	84.7	**506**	91.7	88.8	88.5
888	90.7	88.7	81.7	**552**	92.2	89.4	87.9
962	91.6	89.6	86.8	**598**	91.2	90.8	86.9
1036	90.5	89.3	85.9	**644**	91.1	90.0	88.5
1110	91.8	90.6	85.5	**690**	91.3	90.7	87.4
1184	91.5	90.4	84.5	**736**	90.3	91.2	88.3
1258	92.0	90.0	87.3	**782**	91.7	90.0	87.9
1332	91.1	90.9	87.3	**828**	90.8	91.6	88.7
1406	91.8	90.8	86.3	**874**	92.3	91.0	88.7
1480	92.0	90.4	85.6	**920**	91.6	90.8	88.6
1554	92.5	90.4	86.5	**966**	92.9	91.9	89.6

[a] Aktif toplu öğrenme (atop), toplu öğrenme (top) ve tekil öğrenmenin (tek) beş standart verisetindeki artan eğitim örneği sayıları (eös) için başarı oranları

TABLO III. AKTİF ÖĞRENMEDE KÜMELEMELİ BAŞLANGIÇ

D159						MUSHROOM					
eös	**ras**	**km**	**mog**	**fcm**	**som**	**eös**	**ras**	**km**	**mog**	**fcm**	**som**
46	74.0	70.1	71.0	75.8	82.0	54	90.9	94.5	96.0	86.6	92.6
140	87.3	84.1	83.5	86.8	82.4	160	99.2	97.9	93.6	97.7	96.1
234	90.4	90.1	92.3	90.8	86.7	266	99.7	98.2	99.4	99.0	98.8
328	89.8	91.5	91.6	84.5	89.5	372	99.9	99.0	99.8	98.4	99.7
422	88.7	91.6	91.3	91.1	91.9	478	100	99.2	100	99.9	100
516	93.0	95.1	94.3	93.9	92.7	584	99.7	99.3	99.9	100	100
610	92.2	94.5	93.7	94.3	91.8	690	100	99.3	100	100	100
704	91.6	93.6	95.4	96.2	94.5	796	100	100	100	100	100
798	95.1	95.1	96.2	93.9	95.7	902	100	100	100	100	100
892	94.7	97.1	96.3	94.8	95.3	1008	100	100	100	100	100
986	94.4	96.4	96.0	96.4	97.1	1114	100	99.9	100	100	100
1080	96.8	96.0	96.6	95.7	97.1	1220	100	100	100	100	100
1174	96.5	95.7	96.4	96.8	96.3	1326	100	100	100	100	100
1268	96.8	96.9	96.3	95.8	96.8	1432	100	100	100	100	100
1362	96.3	96.1	97.3	97.2	97.4	1538	99.9	100	100	100	100

RINGNORM						SPAMBASE					
eös	**ras**	**km**	**mog**	**fcm**	**som**	**eös**	**ras**	**km**	**mog**	**fcm**	**som**
48	68.4	77.1	77.0	71.3	69.7	30	80.1	79.6	77.1	78.3	80.8
144	83.6	82.2	85.5	85.2	80.5	90	82.7	85.1	86.2	82.6	84.2
240	85.7	87.3	86.5	85.1	86.6	150	84.4	85.2	86.4	86.8	87.8
336	85.8	86.8	86.9	87.2	85.8	210	88.0	87.2	86.2	87.6	87.5
432	89.9	90.2	87.8	88.1	89.1	270	89.0	87.8	88.5	88.5	88.0
528	88.6	90.1	86.7	88.7	86.9	330	87.8	88.2	88.7	89.5	89.2
624	90.6	89.1	89.6	89.8	91.5	390	90.1	88.2	92.2	90.1	89.9
720	90.4	89.8	90.1	90.2	90.1	450	90.6	91.5	89.5	91.4	91.1
816	90.8	90.7	90.8	91.1	90.1	510	89.2	90.4	90.1	89.6	91.6
912	91.5	92.0	91.5	91.5	90.9	570	90.5	91.5	92.2	90.9	92.1
1008	91.2	90.7	91.0	91.1	90.8	630	90.7	90.4	91.9	90.5	91.6
1104	92.1	91.3	92.1	90.4	91.3	690	90.8	91.4	91.3	92.4	93.7
1200	92.6	90.3	90.6	91.6	92.5	750	91.5	91.7	92.3	92.5	92.0
1296	91.8	92.5	92.2	92.0	91.6	810	92.1	93.1	92.6	91.2	92.9
1392	91.3	91.8	91.4	92.4	93.1	870	93.4	92.1	91.8	91.6	91.5

[b] Aktif Öğrenmede başlangıç kümesinin rastgele (ras) ve kmeans (km), mixture of Gaussians (mog), fuzzy cmeans (fcm), som kümeleme yöntemleriyle seçildiği durumlar için başarılar

V. SONUÇ

Test sonuçları aktif toplu öğrenmenin 5 veriseti için de toplu öğrenmeye göre bir adım önde olduğunu göstermiştir. Bu durum iyi seçilmiş (aktif öğrenme ile) az sayıdaki örnek ile eğitilen algoritmaların pasif olarak çok sayıda örnekle eğitilmiş algoritmalarla kıyas götürür başarı oranlarını yakalayabileceğini kanıtlamaktadır. İlk modelleme örneklerinin belirlenmesinde kümeleme kullanmanın aktif öğrenmenin erken safhalarında bazı verisetleri için performansta hissedilir artışlar oluşturduğu görülmüştür. Bu safhada hangi kümeleme algoritmasının daha iyi olduğuna dair genel bir çıkarımda bulunulamasa da söz konusu veri kümesindeki dağılım için en iyi çalışan kümeleme algoritmasının makul bir başarı hedefine diğerlerine göre daha çabuk ulaştığı söylenebilir.

Sonuç olarak gerçekleştirilen testler aktif öğrenmede etiketlenecek örneklerin seçimi için Özyüklemeli Kümeleme gibi kolektif öğrenme yöntemleriyle üretilmiş komitelerden faydanılabildiğini ve pasif öğrenmeye kıyasla daha az eğitim örneğiyle makul başarı oranlarının yakalanabildiğini göstermiştir. İleriki araştırmalarımızda toplulukların oluşturulmasında Özyüklemeli Kümelemeden farklı yöntemlerin ve aktif öğrenme sürecinde marjin dışındaki diğer anlaşmazlık ölçülerinin genel başarıyı nasıl etkilediğini inceleyen çalışmalar gerçekleştirmeyi planlıyoruz.

KAYNAKÇA

[1] Davy, M. (2005), A review of active learning and co-training in text classification. Dep. of Computer Science, Trinity College Dublin, Research Report, TCD-CS-2005-64, 39 pp.

[2] Cohn, D., Atlas, L. and Ladner, R. (1994), Improving generalization with active learning. Machine Learning, 15(2): 201–221.

[3] Seung, H.S., Opper, M. and Sompolinsky, H. Query by committee. Proceeding COLT '92, Pages 287-294

[4] Settles, B. *Active Learning Literature Survey.* Computer Sciences Technical Report 1648. University of Wisconsin–Madison.

[5] Lewis, D. and Gale, W. A sequential algorithm for training text classifiers. In Proceedings of the ACM SIGIR, pages 3–12. ACM/Springer, 1994.

[6] Shannon, C.E. A mathematical theory of communication. Bell System Technical Journal, 27:379–423,623–656, 1948.

[7] Abe, N. and Mamitsuka, H., Query learning strategies using boosting and bagging, in Proceedings ICML-98, 1–10.

[8] Melville, P. and Mooney, R. (2003), Constructing diverse classifier ensembles using artificial training examples, in Proceedings of IJCAI, 505–510.

[9] Melville, P. and Mooney, R. (2004), Diverse ensembles for active learning, in Proceedings of the 21st Int. Conference on Machine Learning, 584–591.

[10] Stefanowski, J., Pachocki, M. (2009), Comparing Performance of Committee based Approaches to Active Learning, Recent Advances in Intelligent Information Systems, 457-470

[11] Data for Evaluating Learning in Valid Experiments [http://www.cs.toronto.edu/~delve/]

[12] UC Irvine Machine Learning Repository [http://archive.ics.uci.edu]