SELÇUK ÜNİVERSİTESİ
MÜHENDİSLİK FAKÜLTESİ

S.Ü. Müh. Bilim ve Tekn. Derg., c.7, s.1, ss. 154-164, 2019
Selcuk Univ. J. Eng. Sci. Tech., v.7, n.1, pp. 154-164, 2019
ISSN: 2147-9364 (Electronic)
DOI: 10.15317/Scitech.2019.189

FARKLI BAĞLANTI YÖNTEMLERİ İLE HİYERARŞİK KÜMELEME TOPLULUĞU

Derya BİRANT

Dokuz Eylül Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, İZMİR
derya@cs.deu.edu.tr

(Geliş/Received: 13.03.2018; Kabul/Accepted in Revised Form: 04.11.2018)

ÖZ: Kümeleme topluluğu, yüksek kümeleme performansı sağlaması nedeniyle son yıllarda tercih edilen bir teknik haline gelmiştir. Bu çalışmada, Bağlantı-tabanlı Hiyerarşik Kümeleme Topluluğu (BHKT) olarak isimlendirilen yeni bir yaklaşım önerilmektedir. Önerilen yaklaşımda, topluluk elemanları farklı bağlantı yöntemleri kullanarak hiyerarşik kümeleme yapmakta ve sonrasında çoğunluk oylaması ile ortak karar üretmektedir. Çalışmada kullanılan bağlantı yöntemleri: tek bağlantı, tam bağlantı, ortalama bağlantı, merkez bağlantı, Ward yöntemi, komşu birleştirme yöntemi ve ayarlı tam bağlantıdır. Ayrıca çalışmada, farklı boyutlardaki hiyerarşik kümeleme toplulukları incelenmiş ve birbiriyle karşılaştırılmıştır. Deneysel çalışmalarda, hiyerarşik kümeleme toplulukları 8 farklı veri setinde uygulanmış ve tek bir kümeleme algoritmasına göre daha iyi sonuçlar elde edilmiştir.

Anahtar Kelimeler: *Bağlantı yöntemleri, Hiyerarşik kümeleme, Kümeleme topluluğu*

Hierarchical Clustering Ensemble with Different Linkage Methods

ABSTRACT: Clustering ensemble has become a preferred technique in recent years due to the high clustering performance it provides. In this study, a new approach called Link-based Hierarchical Clustering Ensemble (LHCE) is proposed. In the proposed approach, the ensemble members perform hierarchical clustering using different linkage methods and then make joint decisions with majority voting. Linkage methods used in this study are single linkage, complete linkage, average linkage, centroid linkage, Ward method, neighbor joining and adjusted complete linkage. In this study, hierarchical clustering ensembles with different sizes were also investigated and compared with each other. In the experimental studies, hierarchical clustering ensembles were applied on 8 different datasets and better results were obtained rather than a single clustering algorithm.

Key Words: *Clustering ensemble, Hierarchical clustering, Linkage methods*

GİRİŞ (INTRODUCTION)

Kümeleme, veri elemanlarının sahip oldukları özelliklere göre belirli bir yakınlık kriteri ile değerlendirilerek gruplara ayrılmasıdır. Kümeleme, sınıflandırmadan farklı olarak eğitim aşaması içermeyen bir yöntemdir. Etiketi olmayan verileri yakınlığa (benzerliğe) göre bir gruba dahil etmektedir. Başlıca kümeleme yöntemleri: bölümlemeli, hiyerarşik, yoğunluk tabanlı, ızgara tabanlı ve model tabanlı yöntemlerdir. Bu makalede sunulan çalışmada, hiyerarşik kümeleme yöntemi kullanılmaktadır.

Hiyerarşik kümeleme, örnekleri birbirlerine yakınlıklarına göre değişik aşamalarda bir araya getirerek kümeleri ardışık bir biçimde oluşturmaktadır. Temelde iki yaklaşıma sahiptir: birleştirici (agglomerative) ve ayrıştırıcı / bölücü (divisive) (Rafsanjani ve diğ., 2012). *Birleştirici hiyerarşik kümelemede*, her bir örnek başlangıçta ayrı bir küme olarak kabul edilir, her adımda kümelerin birbirlerine uzaklıkları hesaplanır ve en yakın iki küme birleştirilerek bir üst küme oluşturulur. Böylece,

her adımda küme sayısı bir azaltılır ve işlem bütün örneklerin dahil olduğu tek bir büyük küme oluşuncaya kadar devam ettirilir. Bu süreç, dendogram olarak adlandırılan hiyerarşik bir ağaç diyagramı ile görselleştirilebilmektedir. *Ayrıştırıcı hiyerarşik kümelemede* ise, tam tersi bir biçimde, tüm örnekler başlangıçta tek bir küme olarak kabul edilir ve her aşamada benzer olmayan gözlemler belirlenerek daha küçük kümeler oluşturulur. Bu süreç, her gözlem tek başına küme oluşturuncaya kadar sürdürülür. Bu makalede sunulan çalışmada, birleştirici hiyerarşik kümeleme yöntemi kullanılmaktadır.

Son yıllarda, sınıflandırıcı toplulukları (classifier ensembles) ile elde edilen yüksek başarılar kümeleme alanının da ilgisini çekmiş ve tek bir kümeleyici kullanmak yerine bir *kümeleme topluluğu* (veya *kümeleyici topluluğu*) (clustering ensemble) oluşturulması ve topluluk elemanlarının ortak karar vermesinin sağlanması gündeme gelmiştir. Yapılan çalışmalar (Sarumathi ve diğ., 2015; Cornuejols ve diğ., 2018), kümeleme topluluğu içerisindeki farklı kümeleyicilerin fikir birliği sonucunda çıkan ortak çözümün, tek bir kümeleme algoritmasının ürettiği çözümden daha doğru olduğunu göstermiştir.

Bu motivasyondan yola çıkarak, bu makalede Bağlantı-tabanlı Hiyerarşik Kümeleme Topluluğu (BHKT) olarak isimlendirilen yeni bir kümeleme topluluğu yaklaşımı önerilmektedir. Önerilen yaklaşımda, her bir topluluk elemanı aynı hiyerarşik kümeleme algoritmasını farklı bir bağlantı yöntemi ile çalıştırmaktadır. Deneysel çalışmalarda kullanılan bağlantı yöntemleri şunlardır: tek bağlantı (single linkage), tam bağlantı (complete linkage), ortalama bağlantı (average linkage), merkez (centroid) bağlantı, Ward yöntemi, komşu birleştirme (neighbor joining) yöntemi ve ayarlı tam bağlantı (adjusted complete linkage).

Kümeleme topluluğunu oluşturacak kümeleyicilerin türü, sayısı ve kararlarının birleştirilmesinde (consensus) kullanılan yöntem, ortak üretilecek çözümü etkileyen önemli faktörlerdir. Bu makaledeki çalışmanın bir amacı da, hiyerarşik kümeleme için en uygun (optimum) sayıda ve bağlantı türünde en doğru sonuca ulaştıran topluluk elemanlarının belirlenmesidir. Bu amaca yönelik olarak, alternatif topluluk tasarımları, makine öğrenmesi çalışmalarında sıklıkla kullanılan 8 veri seti üzerinde denenmiş ve karşılaştırılmıştır. Deneysel çalışmalarda elde edilen sonuçlara göre, hiyerarşik kümeleme toplulukları, tek bir kümeleyiciye göre doğruluk açısından daha yüksek başarı göstermiştir.

İLGİLİ ÇALIŞMALAR (RELATED WORK)

Literatürdeki *topluluk öğrenmesi* (ensemble learning) çalışmaları ilk olarak *gözetimli* (supervised) öğrenme yaklaşımı olan sınıflandırma üzerinedir, ancak sonrasında aynı başarıyı sağlayabilmek için *gözetimsiz* (unsupervised) öğrenme olan kümeleme konusunda da çeşitli çalışmalar yapılmaya başlanmıştır (Sarumathi ve diğ., 2015; Cornuejols ve diğ., 2018). Ancak, kümeleme algoritmalarının sonuçlarının birleştirilerek ortak karar üretilmesi çok kolay bir işlem değildir. Kümeleme topluluklarının başarısını etkileyen faktörlerin detaylı olarak araştırılması gerekmektedir (Amasyalı ve Ersoy, 2008).

Geçmiş çalışmalar incelendiğinde, kümeleme topluluğu konusu üzerine farklı alanlarda çalışıldığı görülmektedir, örneğin; turizm (D'Urso ve diğ., 2013), kimya (Saeed ve diğ., 2012), biyoloji (Pirim ve Şeker, 2012), yüksek hızlı tren (Xiao ve diğ., 2016) ve elektrik enerjisi (Khan ve diğ., 2016). Kümeleme toplulukları farklı türlerde oluşturulabilmektedir: sıralı kümeleme (sequential clustering), kooperatif kümeleme (cooperative clustering) ve işbirlikçi kümeleme (collaborative clustering) toplulukları bulunmaktadır (Cornuejols ve diğ., 2018). Bazı kümeleme topluluğu çalışmaları *torbalama* (bagging) tabanlı (D'Urso ve diğ., 2013), bazıları ise *hızlandırma* (boosting) tabanlı (Smeraldi ve diğ., 2011) olarak gerçekleştirilmiştir.

Literatürdeki bazı çalışmalar (Smeraldi ve diğ., 2011; D'Urso ve diğ., 2013) topluluğun üretim aşamasına odaklanmıştır, bazıları (Saeed ve diğ., 2012; Xiao ve diğ., 2016) ise sonuçların birleştirilmesi aşamasını iyileştirmeye yönelik yapılmıştır. Son yıllarda, kümeleme topluluğu içerisindeki başarılı kümeleme çözümlerini seçen, *budama* (pruning) yöntemi sonrasında ortak karar üreten çalışmalar (Yu ve diğ., 2014; Zhao ve diğ., 2017; Akyüz ve Otar, 2017) gerçekleştirilmiştir. Ayrıca, *ağırlıklı* (weighted) kümeleme topluluğu yaklaşımlarının da denendiği (Liu ve diğ., 2017) ve başarılı sonuçlar elde edildiği görülmektedir.

K-means algoritması en yaygın kullanılan kümeleme algoritmalarından birisi olduğu için, kümeleme topluluğu çalışmalarında da yine bu algoritmanın sıklıkla kullanıldığı görülmektedir (Liu ve diğ., 2017; Yang ve diğ., 2017; Ren ve diğ., 2017). Ancak, hiyerarşik kümeleme algoritmasını da kullanan çeşitli kümeleme topluluğu çalışmaları (Li ve Chen, 2009; Rashedi ve Mirzaei, 2013; Xiao ve diğ., 2016) mevcuttur. Li ve Chen (2009), hiyerarşik kümeleme topluluğu ve birliktelik kuralları (association rules) konularını birleştiren hibrit bir yaklaşım önermişlerdir. Rashedi ve Mirzaei (2013), hızlandırma (boosting) tabanlı bir hiyerarşik kümeleme topluluğu önermiş, her bir örneğe bir ağırlık değeri atayarak iteratif bir yol izlemişlerdir. Xiao ve arkadaşları (2016) yarı denetimli (semi-supervised) hiyerarşik kümeleme topluluğu önermiş ve çalışmalarında CHAMELEON algoritmasını kullanmışlardır.

Bu makalede sunulan çalışmada, mevcut çalışmalardan farklı olarak, hangi bağlantı yöntemleri ve üye sayısı ile en iyi hiyerarşik kümeleme topluluğunun üretilebileceği araştırılmaktadır. Literatürdeki bazı çalışmalar (Gionis ve diğ., 2007; Yi ve diğ., 2012) birkaç farklı bağlantı türünü içeren hiyerarşik kümeleme algoritmasını topluluk oluşturulmasında kullanmıştır. Ancak, bu çalışmalar sadece hiyerarşik kümeleme algoritmaları kullanmamış, aynı zamanda başka kümeleme algoritmalarını da (k-means, k-medoids, fuzzy c-means gibi) topluluğa dahil etmişlerdir. Ayrıca, kümeleme sonuçlarının birleştirilmesinde de bu çalışmadakinden farklı algoritmalar kullanmışlardır.

Hiyerarşik Kümeleme (Hierarchical Clustering)

Hiyerarşik kümelemenin her aşamasında, Öklid veya başka bir uzaklık ölçütü kullanarak, birleştirilecek olan birbirine en benzer iki kümenin belirlenmesinde farklı yaklaşımlar uygulanabilmektedir. İki küme arasındaki uzaklığın hesaplanmasında kullanılan başlıca yöntemler: tek bağlantı, tam bağlantı, ortalama bağlantı, merkez bağlantı, Ward yöntemi, komşu birleştirme yöntemi ve ayarlı tam bağlantıdır.

Tek bağlantı (*single-link*): İki küme (k_1 ve k_2) arasındaki uzaklık (U), k_1 kümesi ile k_2 kümesinin birbirine en yakın olan iki elemanı (x_1 ve x_2) arasındaki uzaklık olarak kabul edilir. En kısa mesafe esasına dayandığı için *en yakın komşuluk* tekniği olarak da bilinmektedir (Murtagh ve Contreras, 2017).

$$U(k_1, k_2) = \min_{x_1 \in k_1, x_2 \in k_2} U(x_1, x_2) \tag{1}$$

Tam bağlantı (*complete-link*): İki küme (k_1 ve k_2) arasındaki uzaklık (U), k_1 kümesi ile k_2 kümesinin birbirine en uzak olan iki elemanı (x_1 ve x_2) arasındaki uzaklık olarak kabul edilir. Kümeler arası eleman çiftleri arasındaki maksimum uzaklık dikkate alındığı için *en uzak komşuluk* tekniği olarak da bilinmektedir (Murtagh ve Contreras, 2017).

$$U(k_1, k_2) = \max_{x_1 \in k_1, x_2 \in k_2} U(x_1, x_2) \tag{2}$$

Ortalama bağlantı (*average-link*): Birinci küme (k_1) ile ikinci küme (k_2) elemanları arasındaki bütün uzaklıklar hesaplanır ve bunların ortalaması iki küme arasında uzaklık (U) olarak kabul edilir. Başka bir deyişle, karşılıklı iki küme arasındaki tüm mesafelerin ortalamasıdır (Murtagh ve Contreras, 2017).

$$U(k_1, k_2) = \frac{1}{|k_1|} \frac{1}{|k_2|} \sum_{x_1 \in k_1} \sum_{x_2 \in k_2} U(x_1, x_2) \tag{3}$$

Merkez (*centroid*) *bağlantı*: Birinci kümenin (k_1) merkezi (p elemanlı ortalama vektör) ve ikinci kümenin (k_2) merkezi arasındaki uzaklık hesaplanır (Murtagh ve Contreras, 2017).

$$U(k_1, k_2) = U\left(\left(\frac{1}{|k_1|}\sum_{x\in k_1}\vec{x}\right), \left(\frac{1}{|k_2|}\sum_{x\in k_2}\vec{x}\right)\right) \quad (4)$$

Ward yöntemi: Bir kümenin merkezinde bulunan örneğin, kümenin içinde bulunan örneklerden ortalama uzaklığını dikkate alır. Yani, toplam küme içi varyansı minimize etmeyi hedefler. Bu amaçla, küme içi kareli sapmalardan yararlanarak hata kareler toplamını hesaplar (Murtagh ve Contreras, 2017).

$$TU_{k_1\cup k_2} = \sum_{x\in k_1\cup k_2} U(x, \mu_{k_1\cup k_2})^2 \quad (5)$$

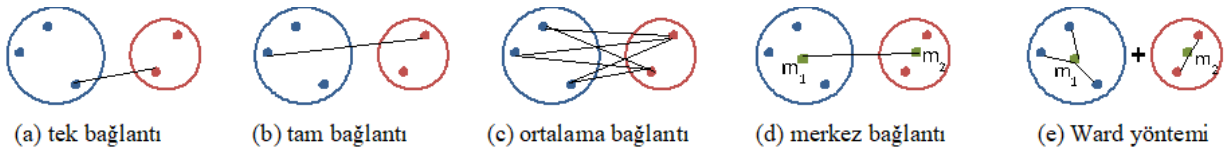
Komşu birleştirme yöntemi (neighbor joining): Diğer yöntemlere göre daha karmaşık olan bu yöntem aşağıdaki adımları içermektedir (Sharma ve diğ., 2018).

- Tüm ikili uzaklıklar (pairwise distances) hesaplanır.
- Göreceli uzaklıkları (relative distances) minimum olan iki eleman *i* ve *j* seçilir.
- Seçilen elemanlara (*i* ve *j*) ilişkin satırlar ve sütunlar uzaklık matrisinden silinir.
- Seçilen elemanları temsil edecek yeni ana elemanın (parent node) (*k*) diğer elemanlara olan uzaklıkları hesaplanır ve uzaklık matrisine eklenir.
- İki eleman kalıncaya kadar aynı işlemler tekrarlanır.

Ayarlı tam bağlantı (*adjusted complete-link*): İki kümenin (k_1 ve k_2) birbirine en uzak olan iki elemanlarının (x_1 ve x_2) arasındaki uzaklıktan, kümelerin *küme içi uzaklık* (KİU) (within cluster distance) değerlerinden büyük olanının çıkartılması ile elde edilir. Başka bir deyişle, k_1 ve k_2 kümelerinin birleşimi ile elde edilen "küme içi uzaklık" değerinden, kümelerin ayrı ayrı "küme içi uzaklık" değerlerinden büyük olanının çıkartılmasıyla hesaplanır (Kamvar ve diğ., 2002).

$$U(k_1, k_2) = \max_{x_1\in k_1, x_2\in k_2} U(x_1, x_2) - \max_{i\in\{1,2\}} K\dot{I}U(k_i) \quad (6)$$

Hiyerarşik kümelemede yaygın olarak kullanılan bağlantı yöntemleri Şekil1'de gösterilmektedir.



(a) tek bağlantı (b) tam bağlantı (c) ortalama bağlantı (d) merkez bağlantı (e) Ward yöntemi

Şekil 1. Hiyerarşik kümelemede kullanılan alternatif bağlantı yöntemleri
Figure 1. *Alternative linkage methods used in hierarchical clustering*

Kümeleme Topluluğu (Clustering Ensemble)

Kümeleme toplulukları, tek bir kümeleyici kullanmak yerine; birden fazla kümeleyici kullanmak ve sonrasında her bir topluluk elemanının ürettiği çıktıları en iyi şekilde birleştirmek ilkesiyle çalışır. Kümeleme topluluklarının oluşturulması iki temel aşama içermektedir: *topluluk üretimi* (ensemble generation) ve *fikir birliği* (consensus) *fonksiyonunun uygulanması*.

Başlıca kümeleme topluluğu üretme yöntemleri şunlardır: (Vega-pons ve Ruiz-Shulcloper, 2011)

- *Farklı kümeleme algoritmaları*
  Topluluk elemanları farklı kümeleme algoritmaları (örneğin; k-means, bulanık c-means, hiyerarşik kümeleme, beklenti maksimizasyonu (EM) gibi) çalıştırmaktadır. Böylece farklı bakış açısına sahip çözümlerin birleştirilmesi ile başarılı sonuçlar elde edilebilmektedir. (Pirim ve Şeker, 2012; Yi ve diğ., 2012)

- *Farklı parametreler veya farklı başlangıçlar*
  Tek bir kümeleme algoritmasının farklı parametreler ile çalıştırılması sonucu oluşan kümeleme çözümleri birleştirilmektedir (Akyüz ve Otar, 2017). Örneğin; k-means algoritması farklı rastgele sayılar ile başlatılabilmekte veya farklı uzaklık metrikleri (Euclidean, Manhattan, Minkowski, vb.) kullanılabilmektedir.
- *Farklı nesneler*
  Veri setinden rastgele nesne seçimleri ile farklı alt gruplar oluşturulabilmekte, böylece kümeleme topluluğunun kendi içinde çeşitlilik (diversity) içermesi sağlanabilmektedir. Örneğin; önyükleme (bootstrap) yöntemi kullanılabilmektedir. (Alqurashi and Wang, 2018)
- *Farklı özellik seçimleri*
  Kümeleyicilerin veri koleksiyonunun rastgele seçilen farklı özellik alt-grupları üzerinde verdikleri kararlar birleştirilmektedir. (Yu ve diğ., 2014)
- *Farklı nesne temsilleri* (object respresentation)
  Veri dönüştürme veya yeni özellik çıkarımı gibi yöntemlerle farklı nesne temsilleri yapılabilmektedir. (Sarumathi ve diğ., 2015)
- *Alt uzaylara izdüşüm* (projection to subspaces)
  Yüksek boyutlu verinin kümelenmesinde izdüşümler yöntemi ile verinin boyutu düşürülebilmektedir. (Akyüz ve Otar, 2017)

Bazı çalışmalar (Akyüz ve Otar, 2017) ise birden fazla yaklaşımı bir arada uygulamaktadır. Bu makaledeki çalışmada, ikinci sıradaki yaklaşım uygulanmakta, yani aynı kümeleme algoritması farklı yöntem parametreleri ile çalıştırılarak topluluk oluşturulmaktadır.

Kümeleme topluluğu oluşturulmasının ikinci aşaması olan "fikir birliği fonksiyonunun uygulanmasında" kullanılan başlıca yöntemler şunlardır: (i) etiketleme / oylama tabanlı yöntemler, (ii) eş-ilişkili matris tabanlı yöntemler, (iii) grafik tabanlı yöntemler, (iv) hiyerarşik yöntemler ve (v) medyan bölme tabanlı yaklaşımlar. Bazı çalışmalarda (Yu ve diğ., 2014; Yi ve diğ., 2014; Zhao ve diğ., 2017; Yang ve diğ., 2017; Ren ve diğ., 2017), kümeleme çözümlerinin birbirleri ile olan yakınlığını ya da çeşitliliklerini ölçen Normalize Karşılıklı Bilginin (Normalized Mutual Information) maksimize edilmesi hedeflenmektedir. Bu makaledeki çalışmada ise oylama tabanlı bir yöntem kullanılmıştır.

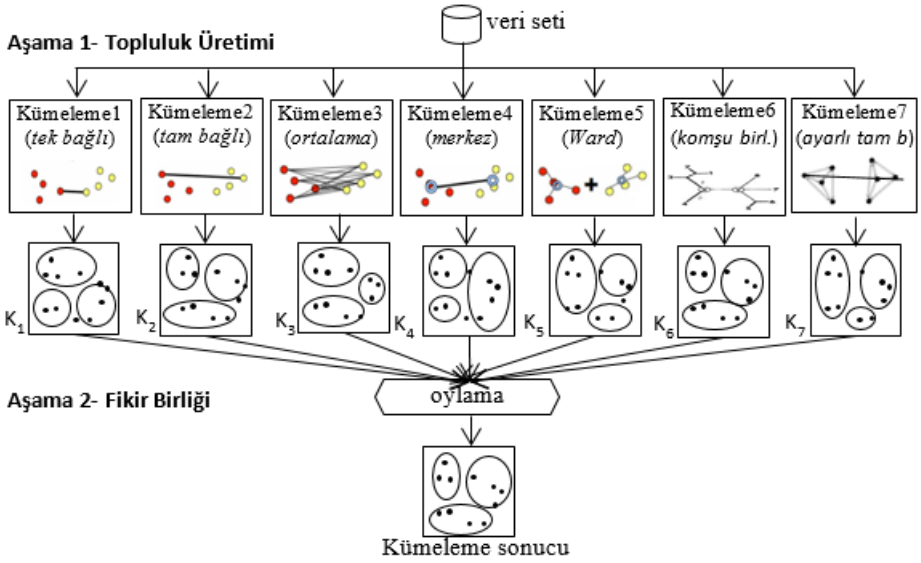
Kümeleme topluluklarının sağladığı başlıca avantajlar aşağıda listelenmektedir:

- *Doğruluk* (*Accuracy*): Toplu karar sayesinde, kümeleme algoritmasının yanlış seçim yapma olasılığı azalmaktadır. Çünkü yanlış karar vermeleri için kümeleyicilerin yarıdan fazlasının yanlış karar vermiş olması gerekir. Azınlıkta bulunan yanlış kümelemeler dikkate alınmamakta, böylece tek bir kümeleme algoritmasına göre ortalamada daha iyi performans elde edilebilmektedir.
- *Sağlamlık* (*Robustness*): Birden fazla modelin iyi sonuçları birleştirilmekte, böylelikle daha sağlam ve güvenilir kümeleme sonuçları elde edilebilmektedir.
- *Kararlılık* (*Stability*): Gürültü ve aykırı değerlere karşı daha düşük hassasiyetli sonuçlar elde edilebilmektedir. Ayrıca, sonuçların algoritma parametrelerine bağımlılığını azaltabilmektedir.
- *Yenilik* (*Novelty*)*:* Kümeleme toplulukları, tek bir kümeleme algoritmalarıyla erişilemeyen çözümlerin bulunmasına izin vermektedir.
- *Model seçimi* (*Model selection*): Kümeleme toplulukları, elde edilecek son kümelerin sayısını belirlemek için temel çözümler arasındaki eşleşmeyi göz önünde bulundurarak model seçim problemine bir yaklaşım sağlamaktadır.
- *Bilginin yeniden kullanımı* (*Knowledge reuse*): Geçmiş zamanlarda elde edilmiş olan kümeleme sonuçlarının birleştirilerek yeniden kullanılması mümkündür.
- *Dağıtık hesaplama ve paralelleştirme* (*Distributed computing and parallelization*): Dağıtık veri kaynaklarında paralel bir şekilde elde edilen kümeleme sonuçlarının bir merkezde birleştirilebilmesine imkan tanır.
- *Çoklu bakış* (*Multiple views*): Farklı kümeleme algoritmalarının aynı problem için kullanılabilmesini ve böylece farklı bakış açısına sahip çözümlerin bileştirilebilmesini sağlar.

Ayrıca, farklı özellik seçimleri veya nesne seçimleri ile farklı açılardan yapılan değerlendirmeler birleştirilebilir.

ÖNERİLEN HİYERARŞİK KÜMELEME TOPLULUĞU (PROPOSED HIERARCHICAL CLUSTERING ENSEMBLE)

Bu makalede, BHKT (Bağlantı-tabanlı Hiyerarşik Kümeleme Topluluğu) adı verilen yeni bir hiyerarşik kümeleme topluluğu yaklaşımı önerilmektedir. Önerilen yaklaşıma göre, birleştirici (agglomerative) hiyerarşik kümeleme yönteminin farklı bağlantı türleri (tek, tam, ortalama, merkez, vb.) ile çalıştırılması sonucu oluşan kümeleme çözümleri oylama yöntemi ile birleştirilmektedir. Şekil 2'de önerilen yaklaşımın genel yapısı gösterilmektedir.



Şekil 2. Önerilen yaklaşım: BHKT (Bağlantı-tabanlı Hiyerarşik Kümeleme Topluluğu)
Figure 2. *Proposed approach: LHCE (Link-based Hierarchical Clustering Ensemble)*

Önerilen yaklaşımda kullanılan çoğunluk oylaması (majority voting) tekniğine ilişkin bir örnek Şekil 3'de gösterilmektedir. Veri örneklerinin ($ö_1$, $ö_2$, $ö_3$, $ö_4$, $ö_5$) her bir kümeleyici (K_1, K_2, K_3, K_4, K_5) sonucuna göre yer aldığı küme numarası bir matris olarak düşünülebilir. Oylama işlemi ile her bir örnek en çok hangi kümeye yerleştirildiyse, nihai sonuçta o kümede yer alır. Mesela, $ö_1$ örneğinin dört kümeleyici tarafından 1 nolu kümeye, bir kümeleyici tarafından da 2 nolu kümeye yerleştirildiği görülmektedir. Bu durumda, çoğunluk kararı ile $ö_1$ örneği 1 nolu kümede yer alacaktır.

	K_1	K_2	K_3	K_4	K_5		K^*
$ö_1$	1	1	1	2	1		1
$ö_2$	1	1	1	1	1		1
$ö_3$	2	2	3	2	2	oylama →	2
$ö_4$	2	2	2	2	2		2
$ö_5$	3	3	1	3	2		3
$ö_6$	3	3	3	3	3		3

Şekil 3. Çoğunluk oylaması tekniği ile kümeleme sonuçlarının birleştirilmesi
Figure 3. *Combining clustering results by majority voting technique*

DENEYSEL ÇALIŞMALAR (EXPERIMENTAL STUDIES)

Deneysel çalışmalarda, önerilen hiyerarşik kümeleme topluluğunun (BHKT) performansını etkileyen kümeleyicilerin türü ve sayısı faktörleri araştırılmıştır.

Veri Setleri (Datasets)

Deneysel çalışmalarda kullanılan, farklı özelliklere (gürültü, eksik veri, farklı veri türleri vb.) sahip, farklı alanlara ilişkin ve kümelemeye uygun 8 veri seti, UCI makine öğrenmesi veri havuzundan (UCI, 2018) alınmıştır. Kullanılan veri setlerine ait örnek sayısı, küme sayısı, özellik sayısı ve veri türleri bilgileri Çizelge 1'de sunulmaktadır.

Çizelge 1. Veri setleri hakkında bilgi

Table 1. *Information about datasets*

No	Veri Seti	Örnek Sayısı	Küme Sayısı	Özellik Sayısı	Özellikler	
					Sürekli	Ayrık
1	blogger	100	2	5	0	5
2	breast-cancer	286	2	9	0	9
3	dermatology	366	6	34	1	33
4	ecoli	336	8	7	7	0
5	iris	150	3	4	4	0
6	seeds	210	3	7	7	0
7	seismic-bumps	2584	2	18	14	4
8	zoo	101	7	16	1	15

Deneysel Sonuçlar (Experimental Results)

Bu çalışmada, hiyerarşik kümeleme topluluklarının performansını etkileyen kümeleyici türü ve sayısı faktörleri 8 veri seti üzerinde araştırılmıştır. Çizelge 2'de sunulan üç deney gerçekleştirilmiş, her bir deneyde farklı özelliklere sahip bir kümeleme topluluğu denenmiş ve elde edilen doğruluk değerleri karşılaştırılmıştır. Kümele toplulukları (3-BHKT, 5-BHKT ve 7-BHKT) oluşturulurken, olası tüm kombinasyonlar denenmemiş, bağlantı türlerinin tek başlarına gösterdikleri başarılara göre seçim yapılmıştır. Çünkü, yüksek doğruluğa sahip kümeleme sonuçlarının oylamaya dahil edilmesi, nihayi sonucun da daha iyi olmasını sağlayacaktır.

Çizelge 2. Hiyerarşik kümeleme toplulukları

Table 2. *Hierarchical clustering ensembles*

Deney No	Topluluk Adı	Eleman Sayısı	Kümeleyici Türleri						
			Tek Bağlı	Tam Bağlı	Ortalama Bağlı	Merkez Bağlı	Ward Yöntemi	Komşu Birleştirme	Ayarlı Tam Bağlı
1	3-BHKT	3	√	√	√				
2	5-BHKT	5	√	√	√	√	√		
3	7-BHKT	7	√	√	√	√	√	√	√

Birinci kümeleme topluluğu (3-BHKT), üç elemandan oluşmaktadır: birinci eleman tek bağlantı, ikinci eleman tam bağlantı, üçüncü eleman ile ortalama bağlantı yöntemi ile hiyerarşik kümeleme yapmaktadır. Veri setlerindeki örneklerin sınıf etiketleri mevcut olduğundan, küme sayısı etiket sayısına eşit olacak şekilde verilmiştir. Çoğunluk kararı ile oluşturulan nihai kümeleme sonuçlarının doğruluk değerleri sınıf etiketleri ile karşılaştırılarak hesaplanmıştır. Uzaklık metriği olarak Öklid uzaklık ölçütü kullanılmıştır. Çizelge 3'ün 1., 2., ve 3. kolonlarında, üç farklı bağlantı türünün ayrı ayrı uygulanması

sonucu elde edilen sonuçlar sunulmaktadır. Çizelge 3'ün 8. kolonunda (3-BHKT) ise üç farklı bağlantı türü ile elde edilen çözümlerin oylama ile birleştirilmesi sonrası elde edilen sonuçlar verilmektedir. Kümeleme doğruluk oranları incelendiğinde, tüm veri setlerinde 3-BHKT yönteminin, tek bir kümeleme algoritmasına göre eşit veya daha doğru sonuçlar ürettiği görülmektedir.

İkinci kümeleme topluluğu (5-BHKT), beş farklı bağlantı yöntemini (tek, tam, ortalama, merkez ve Ward) kullanan beş ayrı kümeleyiciden oluşmaktadır. Çizelge 3'deki 5-BHKT yaklaşımına ilişkin sonuçlar karşılaştırıldığında, çoğu veri setinde (8 tanenin 6 tanesinde) 5-BHKT yönteminin, tek bir kümeleyiciye göre eşit veya daha başarılı performans gösterdiği gözlenmektedir. Ortalamada, tek bir kümeleme algoritması ile en fazla %78,39 başarı elde edilebilirken, bu oran 5-BHKT yönteminde %80,60'dır.

Üçüncü kümeleme topluluğu (7-BHKT), yedi farklı bağlantı yönteminin tümünü oylama işlemi ile birleştirmektedir. Çizelge 3'deki sonuçlar incelendiğinde, bu kümeleme topluluğunun veri setlerin yarısında daha başarılı olduğu, ancak ortalama doğruluk oranı (%80,46) olarak tüm tekli kümeleyicilerden daha iyi performans gösterdiği görülmektedir.

Çizelge 3. Karşılaştırma sonuçları

Table 3. *Comparison results*

Veri seti	Kümeleme Doğruluğu (%)									
	Tek bağlı	**Tam bağlı**	**Ort. bağlı**	**Merkez bağlı**	**Ward**	**Komşu birleşt.**	**Ayarlı tam bağlı**	**3-BHKT**	**5-BHKT**	**7-BHKT**
blogger	70,00	70,00	58,00	67,00	50,00	68,00	63,00	70,00	**72,00**	68,00
breast-cancer	**70,63**	55,59	69,93	**70,63**	66,43	70,28	**70,63**	**70,63**	**70,63**	70,28
dermatology	30,87	65,03	66,12	66,12	63,66	30,60	30,87	**71,31**	69,13	66,94
ecoli	44,94	62,50	74,40	77,38	54,17	42,56	44,05	77,08	**77,68**	77,38
iris	66,00	88,00	**88,67**	66,00	83,33	33,33	66,00	**88,67**	83,33	83,33
seeds	34,76	84,29	89,52	91,90	86,67	33,33	33,81	90,95	89,52	**94,29**
seismic-bumps	93,38	93,00	93,38	93,38	58,86	**93,42**	93,38	93,38	93,38	93,38
zoo	87,13	84,16	87,13	70,30	65,35	40,59	40,59	89,11	89,11	**90,10**
Ortalama	62,21	75,32	78,39	75,34	66,06	51,51	55,29	**81,39**	80,60	80,46

Çizelge 3'ün sol tarafındaki sonuçlar (ilk 7 kolon), her bir kümeleme türünün ayrı ayrı uygulanmasıyla elde edilmiştir. Çizelge 3'ün sağ tarafındaki 3-BHKT, 5-BHKT ve 7-BHKT isimli kolonlardaki sonuçlar ise, sırasıyla ilk 3, 5, ve 7 farklı kümeleme türleri ile elde edilen çözümlerin oylama işlemi ile birleştirilmesi sonrası elde edilmiştir. Çoğunluk oylaması (majority voting) tekniği ile her bir örnek en çok hangi kümeye yerleştirildiyse, nihai sonuçta o kümede yer almaktadır. Mesela, 5-BHKT isimli kümeleme topluluğunda, bir örneği dört kümeleyici 2 nolu kümeye, bir kümeleyici 3 nolu kümeye yerleştiriyorsa, çoğunluk kararı ile söz konusu örnek 2 nolu kümeye yerleştirilir. Kaç tane örneğin doğru kümeye yerleştirildiğine bakılarak Çizelge 3'de verilen "Kümeleme Doğrulukları" hesaplanmıştır.

Farklı eleman sayısına sahip hiyerarşik kümeleme topluluklarının (3-BHKT , 5-BHKT ve 7-BHKT) 8 veri seti üzerinde elde ettikleri doğruluk oranlarının ortalaması sırasıyla %81,39, %80,60 ve %80,46'dır. Dolayısıyla, ortalamada en yüksek başarının (%81,39) üç elemanlı kümeleme topluluğu (3-BHKT) ile elde edildiği görülmektedir.

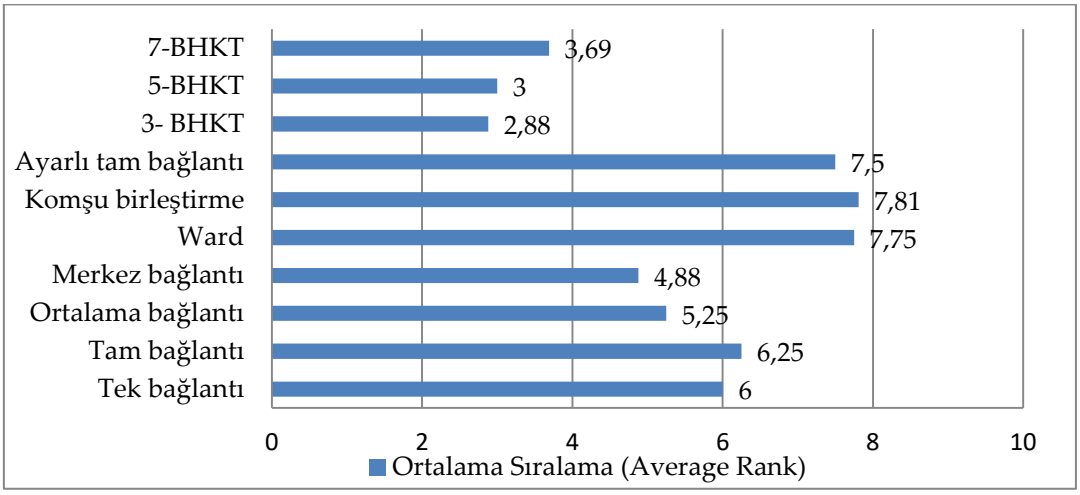
Çizelge 4'de her bir yöntemin 8 veri seti üzerinde elde ettiği başarı sıraları (rank) ve ortalamaları gösterilmektedir. Örneğin; tam bağlı hiyerarşik kümeleme yönteminin diğer 9 yönteme göre veri setlerinde elde ettiği başarı sıralarının ortalaması (3 + 10 + 6 + 6 + 3 + 7 + 9 + 6) / 8 = 6,25'dir. Aynı değere sahip birden fazla yöntem olması durumunda, ortalama başarı sırası atanmıştır. Örneğin; "iris" isimli veri setinde en yüksek değer (%88,67) iki tane olduğu için, 1. ve 2. belirlenememiş, iki yönteme de 1.5 başarı sırası verilmiştir.

Çizelge 4. Yöntemlerin başarı sıralamaları

Table 4. *The ranks of the methods*

Veri seti	Tek bağlı	Tam bağlı	Ort. bağlı	Merkez bağlı	Ward	Komşu birleşt.	Ayarlı tam bağlı	3-BHKT	5-BHKT	7-BHKT
blogger	3	3	9	7	10	5,5	8	3	1	5.5
breast-cancer	3	10	8	3	9	6,5	3	3	3	6,5
dermatology	8,5	6	4,5	4,5	7	10	8,5	1	2	3
ecoli	8	6	5	2.5	7	10	9	4	1	2,5
iris	8	3	1,5	8	5	10	8	1,5	5	5
seeds	8	7	4,5	2	6	10	9	3	4,5	1
seismic-bumps	5	9	5	5	10	1	5	5	5	5
zoo	4,5	6	4,5	7	8	9,5	9,5	2,5	2,5	**1**
Ortalama	6,00	6,25	5,25	4,88	7,75	7,81	7,50	2,88	3,00	3,69

Şekil 4'de gösterilen ortalama başarı sonuçlarına göre, kümeleme topluluklarının (3-BHKT, 5-BHKT ve 7-BHKT) veri setlerinde kümeleme doğruluğu olarak hep ilk sıralarda yer aldığı görülmektedir. Alternatif yedi farklı bağlantı türü karşılaştırıldığında en kötü sonuçların "komşu birleştirme", Ward ve "ayarlı tam bağlantı" yöntemleri ile elde edildiği gözlenmiştir. Kümeleme toplulukları (3-BHKT, 5-BHKT ve 7-BHKT) oluştururulurken bu sonuçlar göz önüne alınmış, olası tüm kombinasyonlar denenmemiştir. Çünkü, doğruluk oranı düşük kümeleme sonuçlarının oylamaya girmesi nihayi sonucu daha iyi hale getirmeyecektir. Tüm olası kombinasyonlar denenmeden, tek başına yüksek kümeleme doğruluğuna sahip olan bağlantı türleri 3-BHKT ve 5-BHKT çözümleri için seçilmiştir. Son olarak da, hepsi birden 7-BHKT çözümünde denenmiştir.



Şekil 4. Alternatif yöntemlerin başarı sıralarının ortalaması

Figure 4. *Average ranks of alternative methods*

Topluluk öğrenmesi yöntemlerinin, tekli yöntemlere göre veri setleri üzerindeki kazanma - kaybetme - berabere kalma sayıları bir matris olarak Çizelge 5'te sunulmaktadır. Örneğin, matristeki 5-2-1 değeri, 8 veri setinin 5 tanesinde 7-BHKT yönteminin başarılı olduğunu, 2 tanesinde tek-bağlı hiyerarşik kümeleme algoritmasının daha iyi performansa sahip olduğunu ve 1 veri setinde de bu iki yöntemin berabere kaldığını ifade etmektedir. Çizelge 5'deki sonuçlar bir bütün olarak değerlendirildiğinde, veri seti sayısı olarak kümeleme topluluğu yöntemlerinin, tek bir kümeleme algoritmasına göre daha başarılı olduklarını söylemek mümkündür.

Çizelge 5. Yöntemler arasındaki kazanma - kaybetme - berabere kalma sayıları
Table 5. *The number of wins - losses - ties among the methods*

	Tek Bağlantı	**Tam Bağlantı**	**Ortalama Bağlantı**	**Merkez Bağlantı**	**Ward Yöntemi**	**Komşu Birleştirme**	**Ayarlı tam Bağlantı**
3-BHKT	5-0-3	7-0-1	6-0-2	4-2-2	8-0-0	7-1-0	6-0-2
5-BHKT	6-0-2	7-1-0	5-1-2	5-1-2	7-0-1	7-1-0	6-0-2
7-BHKT	5-2-1	6-2-0	6-1-1	5-1-2	7-0-1	5-1-2	6-1-1

SONUÇ ve GELECEK ÇALIŞMALAR (CONCLUSION and FUTURE WORK)

Kümeleme toplulukları, farklı kümeleme çözümlerini dikkate alarak ortak bir karar alma gayesindedir. Bu çalışmada, Bağlantı-tabanlı Hiyerarşik Kümeleme Topluluğu (BHKT) olarak isimlendirilen yeni bir yaklaşım önerilmektedir. Önerilen yaklaşımda, her bir topluluk elemanı hiyerarşik kümeleme algoritmasını farklı bir bağlantı yöntemi ile çalıştırmakta ve sonrasında sonuçlar çoğunluk oylaması ile birleştirilmektedir. Çalışmada, hiyerarşik kümeleme topluluklarının performansını etkileyen kümeleyici türü ve sayısı faktörleri incelenmiş ve karşılaştırmalar yapılmıştır. Deneysel çalışmalarda, hiyerarşik kümeleme toplulukları 8 farklı veri setinde uygulanmış ve tek bir kümeleme algoritmasına göre daha başarılı sonuçlar elde edilmiştir. Sonuç olarak; farklı bağlantı yöntemleri kullanan hiyerarşik kümeleyicilerin sonuçlarını birleştirmenin performansa olumlu bir katkı yaptığını söylemek mümkündür.

Gelecek çalışma olarak; hiyerarşik kümeleme dışındaki diğer kümeleme yöntemlerine (bölümlemeli, yoğunluk tabanlı, ızgara tabanlı ve model tabanlı) yönelik kümeleme topluluklarının oluşturulması planlanmaktadır.

KAYNAKLAR (REFERENCES)

Akyüz, S., Otar, B.Ç., 2017, "Doğruluk ve çeşitlilik ödünleşimlerinin eniyilemesi ile kümeleme topluluklarının seçilmesi", *25th IEEE Signal Processing and Communications Applications Conference (SIU)*, 15-18 Mayıs 2017, Antalya, Türkiye.

Alqurashi, T., Wang, W., 2018, "Clustering ensemble method", *International Journal of Machine Learning and Cybernetics*, ss. 1-20.

Amasyalı, M.F., Ersoy, O., 2008, "Kümeleyici topluluklarının başarısını etkileyen faktörler", *IEEE 16th Signal Processing, Communication and Applications Conference (SIU 2008)*, 20-22 Nisan 2008, Aydın, Türkiye.

Cornuejols, A., Wemmert, C., Gançarski, P., Bennani, Y., 2018, "Collaborative clustering: Why, when, what and how", *Information Fusion*, Cilt 39, ss. 81-95.

D'Urso, P., Giovanni, L.D., Disegna, M., Massari, R., 2013, "Bagged clustering and its application to tourism market segmentation", *Expert Systems with Applications*, Cilt 40, ss. 4944-4956.

Gionis, A., Mannila, H., Tsaparas, P., 2007, "Clustering aggregation", *ACM Transactions on Knowledge Discovery from Data*, Cilt 1, Sayı 1, ss. 1-30.

Kamvar, S., Klein, D., Manning, C., 2002, "Interpreting and Extending Classical Agglomerative Clustering Algorithms Using a Model-Based Approach", *19th International Conference on Machine Learning (ICML 2002)*, 8-12 Temmuz 2002, Sydney, Australia, ss. 283-290.

Khan, I., Huang, J. Z., Ivanov, K., 2016, "Incremental density-based ensemble clustering over evolving data streams", *Neurocomputing*, Cilt 191, ss. 36-43.

Li, T., Chen, Y., 2009, "Hierarchical clustering ensemble algorithm based association rules", *International Conference on Wireless Communications, Networking and Mobile Computing*, 24-26 Eylül 2009, Beijing, Çin, ss. 5320-5323.

Liu, H., Wu, J., Liu, Tao, D., Fu, Y., 2017, "Spectral ensemble clustering via weighted k-Means: theoretical and practical evidence", *IEEE Transactions on Knowledge and Data Engineering*, Cilt 29, Sayı 5, ss. 1129-1143.

Murtagh, F., Contreras, P., 2017, "Algorithms for hierarchical clustering: an overview II", *WIREs Data Mining and Knowledge Discovery*, Cilt 7, Sayı 6, ss. 1-16.

UCI Machine Learning Repository. https:// archive.ics.uci.edu/ml/datasets.html, ziyaret tarihi: 12 Mart 2018.

Rafsanjani, M.K., Varzaneh, Z.A., Chukanlo, N.E., 2012, "A survey of hierarchical clustering algorithms", *The Journal of Mathematics and Computer Science*, Cilt 5, Sayı 3, ss. 229-240.

Ren, Y., Domeniconi, C., Zhang, G., Yu, G., 2017, "Weighted-object ensemble clustering: methods and analysis", *Knowledge and Information Systems*, Cilt 51, ss. 661-689.

Sarumathi, S. , Shanthi, N., Ranjetha, P., 2015, "Analysis of diverse cluster ensemble techniques", World Academy of Science, Engineering and Technology, *International Journal of Computer and Information Engineering*, Cilt 9, Sayı 11, ss. 2386-2396.

Saeed, F., Salim, N., Abdo, A., 2012, "Voting-based consensus clustering for combining multiple clusterings of chemical structures", *Journal of Cheminformatics*, Cilt 4, Sayı 37, ss. 1-8.

Sharma, A., Jaloree, S., Thakur, R.S., 2018, "Review of Clustering Methods: Toward Phylogenetic Tree Constructions", *International Conference on Recent Advancement on Computer and Communication, Lecture Notes in Networks and Systems*, Cilt 34, ss. 475-480.

Smeraldi, F., Bicego, M., Cristani, M., Murino, V., 2011, "CLOOSTING: CLustering data with bOOSTING", MCS 2011, *Lecture Notes in Computer Science*, Cilt 6713, ss. 289-298.

Pirim, H., Seker, S.E., 2012, "Ensemble clustering for biological datasets", *Bioinformatics*, InTech publisher, ss. 287-298.

Rashedi, E., Mirzaei, A., 2013, "A hierarchical clusterer ensemble method based on boosting theory", *Knowledge-Based Systems*, Cilt 45, ss. 83-93.

Vega-pons, S., Ruiz-Shulcloper, J., 2011, "A survey of clustering ensemble algorithms", *International Journal of Pattern Recognition and Artificial Intelligence*, Cilt 25, Sayı 3, ss. 337-372.

Yang, F., Li, T., Zhou, Q., Xiao, H., 2017, "Cluster ensemble selection with constraints", *Neurocomputing*, Cilt 235, ss. 59-70.

Yi, J., Yang, T., Jin, R., Jain, A.K., Mahdavi, M., 2012, "Robust ensemble clustering by matrix completion", *IEEE 12th International Conference on Data Mining*, ss. 1176-1181.

Yu, Z., Li, L., Gao, Y., You, J., Liu, J., Wong, H.-S., Han, G., 2014, "Hybrid clustering solution selection strategy", *Pattern Recognition*, Cilt 47, ss. 3362-3375.

Xiao, W., Yang, Y., Wang, H., Li, T., Xing, H., 2016, "Semi-supervised hierarchical clustering ensemble and its application", *Neurocomputing*, Cilt 173, ss. 1362-1376.

Zhao, X., Liang, J., Dang, C., 2017, "Clustering ensemble selection for categorical data based on internal validity indices", *Pattern Recognition,* Cilt 69, ss. 150-168.