Sinop Üniversitesi Fen Bilimleri Dergisi *Sinop Uni J Nat Sci 2(2): 62-78 (2017)*
ISSN: 2536-4383

Kategorik Verilerde Kümeleme İçin Farklı Algoritmaların Karşılaştırılması

Ferhan BAŞ KAMAN[a,*], Semra ERBAŞ[b], Hülya OLMUŞ[b]

[a]*Aksaray Üniversitesi, Şereflikoçhisar Uygulamalı Teknoloji ve İşletmecilik Yüksekokulu, Bankacılık ve Finans Bölümü, Şereflikoçhisar-ANKARA*

[b]*Gazi Üniversitesi, Fen Fakültesi, İstatistik Bölümü, Beşevler-Teknikokullar, ANKARA*

Öz

Kümeleme analizi nesnelerin doğal gruplarını bulmak için kullanılan bir yöntemdir. Kümeleme yapılırken küme içi homojenlik ile kümeler arası heterojenliğin yüksek olması istenir. Literatürde, kategorik verileri kümelemek için çok fazla yöntem yoktur ve var olanların hangisinin en iyi olduğu ile ilgili kesin bir bilgi bulunmamaktadır. Veri sayısına ve veri yapısına göre her bir yöntemin birbirine üstünlükleri ve eksiklikleri vardır. Ayrıca iyi bir kümeleme yapmak için kullanılacak değişken sayısı büyük önem taşımaktadır. Bu çalışmada kategorik verilerin kümelenmesi ile ilgilenildi. Hiyerarşik kümeleme tekniklerinden tek bağlantı tekniği, tam bağlantı tekniği, ortalama bağlantı tekniği ve bölmeli kümeleme tekniklerinden K-modes algoritması kullanılarak kümeleme analizi yapıldı ve sonuçlar karşılaştırıldı. Nitelikli bir karşılaştırma yapmak için literatürde bu tür karşılaştırmaların yapılmasında yaygın olarak kullanılan gerçek veri setlerinden yararlanıldı. Bu analizler MATLAB R2009'da yapılmıştır. Analiz sonuçlarına göre veri sayısı büyüdükçe kümeleme performansı hiyerarşik tekniklerde azalırken K-modes algoritmasında arttığı tespit edildi.

Anahtar Kelimeler: Kümeleme analizi, tek bağlantı tekniği, tam bağlantı tekniği, ortalama bağlantı tekniği ve K-modes algoritma

Comparing Different Algorithms for Clustering in Categorical Data

Abstract

Cluster analysis is a method used to find natural groups of objects. Given a data set the main goal is to produce a partition with high internal intra-cluster similarity and high inter-cluster dissimilaity. In literature, there is not many methods for clustering of categorical data and there is no certain information about which one is best. According to the number of data and data sutructure each has advantages and limitations. Also variable number is important for good clustering results. In this study dealt with clustering of categorical data. Hierarchical clustering techniques which are single linkage, complete linkage, average linkage and partitional clustering technique which is K-modes algorithm were compared. Well known real data sets were used for quality comparison. This analysis was done at MATLAB R2009. According to the analysis results when the number of data set grows clustering performances are decreasing in single linkage, complete linkage, average linkage while K-modes algoritm's is increasing.

Key words: Clustering analysis, single linkage, complete linkage, average linkage and K-modes algorithm

* Corresponding author
E-mail: ferhanbas@aksaray.edu.tr

Received: 15.03.2017
Accepted: 21.11.2017

Giriş

Kümeleme analizi, büyük boyuttaki verileri, aynı gruptaki veriler birbirine en çok benzerlikte, farklı gruplardakiler en az benzerlikte olacak şekilde gruplara bölen bir yöntemdir. Kümeleme analizinin temel amacı nesnelerin doğal gruplarını bulmaktır [1]. Geleneksel kümeleme algoritmaları çoğunlukla sayısal verilerle ilgilenmektedir. Çünkü hesaplamaların yapılması ve benzerliklerin bulunması sayısal verilerle daha kolaydır. Fakat gerçek hayatta birçok veri kategoriktir ve kategorik veriler ile yapılan işlemler daha karmaşıktır [2]. Kategorik verilerde değişkenler birden fazla değere sahip olduğu için benzerlik, ortak nesneler ve ortak değerlerin ilişkisi olarak tanımlanmaktadır. Kategorik verileri kümelemek için birçok algoritma önerilmiştir fakat hangisinin daha iyi sonuçlar verdiği tam olarak belli değildir. Farklı koşullara göre hepsinin birbirine üstünlükleri ve eksiklikleri bulunmaktadır [2]. Kategorik veriler için önerilen bazı kümeleme algoritmaları ROCK, CACTUS, Squeezer, K-modes ve STIRR'dır. K-modes algoritması Huang tarafından 1998 yılında kategorik verileri kümelemek için önerilmiş ve K-ortalamalar algoritmasının genişletilmiş bir versiyonudur. K-ortalamalar kümeleme algoritması kullandığı benzerlik ölçüsünden dolayı kategorik verileri kümeleyemez. K-modes kümeleme algoritması K-ortalamalar algoritması temeline dayanır fakat K-ortalamalar algoritması tarafından getirilen kısıtlamalarda değişiklikler yapılarak kategorik verileri kümelemeye uygun hale getirilmiştir. Bu değişiklikler, kategorik veriler için basit eşleşme katsayısı (Hamming uzaklığı) kullanmak, kümelerin ortalamaları yerine modlarını kullanarak yerlerini değiştirmektir [3]. Gibson ve arkadaşları 1998 yılında STIRR adlı bir algoritma önermiştir. Bu algoritma kategorik veriler için genelleştirilmiş bir spektral yani görsel grafik bölümleme metodudur. STIRR doğrusal olmayan dinamik sistemler için kategorik veri haritası olan tekrarlı bir metottur. Yani değişkene ait her bir nitelik grafik üzerinde ağırlıklandırılmış bir tepe noktasını temsil eder ve bu tepe noktalarından biri asıl küme diğerleri de asıl olmayan kümeler olarak belirlenir. Kümelerdeki yer değişimi bitip bütün kümeler sabitleninceye kadar da kümeleme işlemi devam eder. Kümelerin doğal kombinasyonunun elde edilmesi oldukça kolaydır, fakat STIRR değişken değerleriyle ilişkili yakınlığı tanımlamak için bir ön hazırlık adımı gerektirir. Ek olarak STIRR tarafından kümelerin kesin sınıfları keşfedilemez [4]. Ganti ve arkadaşları tarafından 1999 yılında CACTUS adı verilen verileri özetlemeye dayanan bir algoritma önerilmiştir. CACTUS algoritması bütün değişkenlerin alt kümelerini bulur ve böylece

verinin alt uzay kümelemesini gerçekleştirir [5]. Guha ve arkadaşları 1999 yılında bir hiyerarşik kümeleme metodu olan ROCK algoritmasını sunmuşlardır. ROCK algoritması iki nesne arasındaki ortak komşulukların sayısını hesaplarken bağlantıları kullanmıştır. Bu algoritmada ilk olarak her bir nesne farklı bir küme olarak atanır. Daha sonra kümeler, kümeler arası yakınlığa göre birleştirilir bu yakınlıkta bütün nesne çiftleri arasındaki bağlantıların sayısının toplamı olarak ifade edilir [6]. He ve arkadaşları 2002 yılında Squeezer algoritmasını önermişlerdir. Squeezer algoritması, ilk demeti küme içine koyar ve daha sonraki demetler benzerlik fonksiyonuna dayalı yeni oluşturulmuş bir kümenin benzerliğine göre o kümeye konulur veya reddedilir [7].

Bütün algoritmaların ortak varsayımı her bir nesne sadece bir küme içinde sınıflanabilir ve bütün nesneler bir küme içinde gruplandığında aynı derecede öneme sahiptir. Fakat gerçek hayattaki uygulamalarda kümeler arasında kesin sınırlar çizmek zordur. Bu nedenle, bu çalışmada, iyi bir karşılaştırma yapmak için kategorik verileri kümelemede, literatürde en çok kullanılan gerçek veri setleri kullanılarak, tek bağlantı tekniği, tam bağlantı tekniği, ortalama bağlantı tekniği ve K-modes algoritmalarının kümeleme performansları değerlendirilmiş ve hangi yöntemin performansının daha iyi olduğunu gözlemlemek için doğru kümeleme yüzdeleri kullanılmıştır. Kategorik verileri kümelemek için geliştirilen algoritmalar üzerinden en uygun kümeleme yöntemi belirlenmeye çalışılmıştır.

Kümeleme Analizi

Günümüzde artan bilgi ve buna bağlı olarak artan verilerle sağlıklı bilgiler elde edebilmek için bu verilerden özet bilgi edinme ve sayısını indirgeme ihtiyacı duyulur. Veri setleri büyüdükçe verilerin boyutları, analizleri zorlaştırmakta ve sonuçların doğruluğunu engellemektedir. Böyle büyük veri setlerini kümelemek ise yapacağımız istatistiksel analizlerde bize kolaylık sağlamaktadır [8]. Kümeleme analizinin amacı gruplanmamış verileri benzerliklerine göre sınıflayarak araştırmacıya özet bilgi sağlamak ve çok fazla olan veri sayısını gruplayarak daha az sayıya indirgemektir.

Kümeleme analizi birbirine benzer nesneleri aynı grupta toplarken benzemeyenleri farklı gruplarda toplayan çok değişkenli analiz tekniğidir. Kümeleme analizinde küme içi benzerliğin çok yüksekken kümeler arası benzerliğin düşük olması istenir. Böylece gruplar anlamlı bir şekilde sınıflanmış ve küme içindeki nesneler kümeyi iyi temsil etmiş olur [6].

Değişkenler sayısal (nicel) ve kategorik (nitel) olmak üzere ikiye ayrılır. Sayısal değişkenlerle kümeleme yapabilmek için

ISSN: 2536-4383

genellikle uzaklık ölçümleri kullanılır. Bunlardan en sık kullanılan Öklid uzaklık ölçüsü ve Manhattan City Block uzaklık ölçüsüdür. Kategorik verilerde kümeleme yapabilmek ise sayısal verilerde olduğu kadar kolay değildir. Sayısal veriler üzerinde yapılabilen toplama, bölme, ortalama alma gibi işlemler kategorik veriler üzerinde yapılamaz. Dolayısıyla kategorik verilerde kümeleme yapabilmek için uzaklık ölçümlerini kullanmak uygun değildir. Bunun yerine, daha farklı benzerlik kriterleri kullanmak gereklidir. Kategorik verileri kümelemek için birçok algoritma parametre değerlerinin dikkatli bir şekilde seçimini gerektirir [9]. Bu çalışmada kategorik verilerin kümelenmesi ile ilgilenilmiştir.

Kategorik Verilerde Kümeleme

Kategorik verilerde nesne çiftleri arasında doğal bir uzaklık ölçüsü olmadığı için kümeleme yapmak sayısal verilere kıyasla çok daha zordur. Dolayısıyla kategorik verilerde uzaklık ölçüsü kullanmak uygun değildir. Bunun yerine benzerlik ölçüleri kullanmak gerekir.

Varsayalım ki D p-boyutlu bir veri seti ve V de bütün değişkenlerin birleşimi olsun yani, $V = D_1 \cup D_2 \cup ... \cup D_p$. $x, y \in D$ nesnelerinin herhangi bir çifti için eşleşen ve eşleşmeyen değişken değerlerinin sayısı bir ilişki tablosu kullanılarak gösterilebilir.

Tablo 1. İkiden fazla sonuçlu p sayıda değişken içeren bir nesne çiftinin çapraz tablosu

	$\alpha \in x$	$\alpha \notin x$
$\alpha \in y$	n_{11}	n_{01}
$\alpha \notin y$	n_{10}	n_{00}

Bu tabloda, n_{11} , x ve y nesnelerindeki değişken değerlerinin sayısı; n_{10}, sadece x nesnesindeki değişken değerlerinin sayısı; n_{01}, sadece y nesnesindeki değişken değerlerinin sayısı; n_{00} , ne x ne de y nesnelerindeki değişken değerlerinin sayısıdır.

Eğer x ve y'yi değişken değerlerinin bir dizisi olarak düşünülürse;

$$n_{11} = |x \cap y|$$
$$n_{01} = |x - y|$$
$$n_{10} = |y - x|$$
$$n_{00} = |V - (x \cup y)|$$

olur. İlişki tablosu kullanılarak kategorik değişken değerli nesneler için birçok benzerlik ölçüsü tanımlanabilir. Bu benzerlik ölçüleri aşağıda verilmiştir.

Jaccard Katsayısı :

ISSN: 2536-4383

$$Benzerlik(x,y) = \frac{n_{11}}{(n_{11}+n_{01}+n_{10})} = \frac{|x \cap y|}{|x \cup y|}$$

(1)

Basit Eşleşme Katsayısı :

$$Benzerlik(x,y) = \frac{n_{11}+n_{00}}{(n_{11}+n_{01}+n_{10}+n_{00})} = \frac{|x \cap y| + |V - (x \cup y)|}{|V|}$$

(2)

Kosinüs Katsayısı :

$$Benzerlik(x,y) = \frac{n_{11}}{\sqrt{(n_{11}+n_{01}) \times (n_{11}+n_{10})}} = \frac{|x \cap y|}{\sqrt{|x| \times |y|}}$$

(3)

Örtüşme Katsayısı :

$$Benzerlik(x,y) = \frac{|n_{11}|}{\min(n_{01}, n_{10})} = \frac{|x \cap y|}{\min(|x|, |y|)}$$

(4)

Kategorik veriler üzerine çalışan birçok algoritma benzerlik ölçüsü olarak Eşitlik (1)-(4)'deki formülleri kullanır [10, 11].

Kümeleme Analizi Teknikleri

Bu çalışmada kullanılan kümeleme teknikleri aşağıda verilmiştir.

En yakın komşuluk tekniği (Tek bağlantı) : Tek bağlantı tekniğinde iki küme arasındaki uzaklık mümkün tüm kümeler arasındaki en küçük uzaklık ile başlar. Yani ilk önce en yakın iki küme birleştirilir. D(*r*,*s*), *r* ve *s* kümeleri arası uzaklık, en yakın iki nesne çifti arasındaki uzaklık olarak tanımlanır.

En uzak komşuluk tekniği (Tam Bağlantı): Tam bağlantı tekniğinde iki küme arasındaki uzaklık mümkün tüm kümeler arasındaki en büyük uzaklık ile başlar. Yani ilk önce en uzak iki küme birleştirilir. D(*r*,*s*), *r* ve *s* kümeleri arası uzaklık, en uzak iki nesne çifti arasındaki uzaklık olarak tanımlanır.

Ortalama bağlantı tekniği: Uzaklık *r* ve *s* nesnelerinin bütün çiftleri arasındaki uzaklığın ortalaması olarak tanımlanır. Burada *r* ve *s* farklı kümelere aittir.

Bölmeli kümeleme tekniği: Bölmeli temelli algoritmalar veri setini kullanıcı tarafından belirlenmiş küme sayısına böler. Nesnel bir kriter fonksiyonunun optimize edilmesiyle oluşan kümelerin algoritması olan bölmeli temelli algoritma küme içindeki nesneler arası benzerliği maksimize eder veya uzaklığı minimize eder [12].

K-Modes Algoritması

K-modes algoritması Huang tarafından kategorik verileri kümelemek için önerilmiş bölmeli kümeleme tekniğine sahip K-means algoritmasının genişletilmiş bir versiyonudur [3]. K-ortalamalar kümeleme algoritması kullandığı benzerlik ölçüsünden dolayı kategorik verileri kümeleyemez. K-modes

ISSN: 2536-4383

kümeleme algoritması K- ortalamalar algoritması temeline dayanır fakat K-ortalamalar algoritması tarafından getirilen kısıtlamalarda değişiklikler yapılarak kategorik verileri kümelemeye uygun hale getirilmiştir. Bu değişiklikler, kategorik veriler için basit eşleşme katsayısı veya Hamming uzaklığı kullanmak, kümelerin ortalamaları yerine modlarını kullanarak yerlerini değiştirmektir [13].

Basit eşleşme benzemezlik ölçüsü aşağıdaki gibi tanımlanabilir. *x* ve *y*, *F* kategorik değişken tarafından tanımlanan veri nesneleri olsun. x ve y arasındaki benzemezlik ölçüsü *d(x,y)*, iki nesnenin karşılık gelen kategorik değişkenlerinin eşleşmeyenlerinin toplam sayısıdır. Eşleşmeme ne kadar küçükse iki nesne arasındaki benzerlik o kadar fazladır. Matematiksel olarak aşağıdaki gibi ifade edilebilir.

$$d(x,y)=\sum_{j=1}^{F}\delta(x_j,y_j)$$

(5)

Burada $\delta(x_j,y_j)$, $(x_j=y_j)$ise 0, $(x_j\neq y_j)$ ise 1 değerini alır.

Z kategorik değişkenlerden oluşan bir veri nesne seti olsun, $Z=\{Z_{1,}Z_{2,},...,Z_n\}$ nesnesinin modu $A_1,A_2,...,A_F$, $Q=[q_1,q_2,...,q_F]$ olan bir vektör,

$$D(Z,Q)=\sum_{i=1}^{n}d(Z_i,Q)$$

(6)

eşitliğini minimize eder. Kategorik verilere sahip nesneler için yukarıdaki benzemezlik ölçüsü kullanıldığında amaç fonksiyonu;

$$C(Q)=\sum_{l=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{F}\delta(z_{ij},q_{ij})$$

(7)

şeklindedir.

Burada $Q_l=[q_{l1},q_{l2},...,q_{lm}]\in Q$'dır. K-modes algoritması Eşitlik 7'de tanımlanan amaç fonksiyonunu minimize eder.

K-modes algoritması aşağıdaki adımları içerir:

1. "k" başlangıç modu seçilir.
2. Eşitlik 5'e göre kimin modu kümeye en yakınsa küme içindeki nesneler yer değiştirir.
3. Her yer değiştirmeden sonra küme modları güncellenir.
4. Kümedeki nesnelerde yer değişimi bitinceye kadar 2. ve 3. adımlar tekrarlanır [13].

K-modes algoritması sadece "k" küme sayısının kullanıcı tarafından verilmesini gerektirir. Zaman karmaşıklığı O(nkL) dır. Burada "n" veri setinin büyüklüğü, "k" küme sayısı ve "L" iterasyon sayısıdır [3].

Uygulama

Bu çalışmada dört farklı kümeleme algoritması ele alındı ve dört farklı veri grubuna uygulandı. Hiyerarşik kümeleme tekniklerinden tek bağlantı tekniği, tam bağlantı tekniği, ortalama bağlantı tekniği ve kategorik veriler için geliştirilmiş bölmeli bir kümeleme algoritması olan K-modes algoritması kullanılarak gerçek veri setleri üzerinden kümeleme performansları değerlendirildi. Kategorik verileri kümelemede, kullanılan yöntemlerin nitelikli bir karşılaştırmasını yapmak için literatürde de en çok kullanılan gerçek veri setleri kullanıldı. Bunlar; kongre oylaması, soya fasulyesi, hayvanlar ve arabalara ait veri setleridir[14]. Bu analizler MATLAB R2009'da yapılmıştır. Hiyerarşik kümeleme algoritmalarından tek bağlantı tekniği, tam bağlantı tekniği ve ortalama bağlantı tekniği için MATLAB R2009'da ki hazır fonksiyonlar kullanılmıştır ve benzerlik ölçüsü olarak Jaccard katsayısı kullanılmıştır. K-modes algoritması için de Chaturvedi ve arkadaşlarının 2001 yılında önerdiği MATLAB kodu kullanılmıştır [15].

Hastalıklı Soya Fasulyesine Ait Veri Seti

Bu veri seti soya fasulyesi rahatsızlıkları ile ilgili bilgi içermektedir. 47 tane nesneden oluşmaktadır. Bu nesneler 4 gruba ayrılmıştır: Diaporthe kök pamukçuğu (D1), kömür çürüklüğü (D2), Rhizoctonia kök çürüklüğü (D3) ve Phytopthora çürümesi (D4). Bu grupların dağılımı; D1 için 10, D2 için 10, D3 için 10 ve D4 için 17 tanedir. Her rahatsızlık 36 değişken tarafından tanımlanmaktadır ve hiç kayıp gözlem yoktur. Tablo 2 hastalıklı soya fasulyesi veri setini tanımlamaktadır. Sınıf adlarını içeren son değişken analize dahil edilmeyecektir.

Tablo 2. Soya fasulyesi veri setine ait değişkenler

Değişken Numarası	Değişken İsmi	Değişken Değeri
1	Tarih	Nisan, Mayıs, Haziran, Temmuz, Ağustos, Eylül, Ekim
2	Bitki standı	Normal, Normalden az
3	Yağış	Normalden az, Normal, Normalden fazla
4	Sıcaklık	Normalden az, Normal, Normalden fazla
5	Dolu yağışı	Evet, Hayır

6	Mahsul geçmişi	Geçen yıldan farklı, Geçen yılla aynı, Geçen 2 yılla aynı, Geçen birkaç yılla aynı
7	Hasar görmüş bölüm	Seyrek, Alçak alanlar, Yüksek alanlar, Bütün alan
8	Şiddeti	Küçük, Büyük, Çok büyük
9	Tohum olguları	Hiçbiri, Mantar ilacı, Diğer
10	Çimlenme	%90-100, %80-89, %80'den az
11	Bitki büyümesi	Normal, Anormal
12	Yapraklar	Normal, Anormal
13	Halka yaprak lekeleri	Yok, Sarı halkalar, Sarı olmayan halkalar
14	Marg yaprak lekeleri	w-s marg, w-s marg yok, dna
15	Yaprak lekesi büyüklüğü	1/8'den az, 1/8'den fazla, dna
16	Parça yaprak	Var, Yok
17	Malf yaprak	Var, Yok
18	Yumuşak yaprak	Yok, Üst surf, Alt surf
19	Kök	Normal, Anormal
20	Konaklama	Evet, Hayır
21	Kök pamukçuğu	Yok, Toprağın altı, Toprağın üstü, Yukarıda hasarsız

22	Pamukçuk lezyonu	Dna, Kahverengi, Kahverengi bilinmiyor, Taba rengi
23	Meyve organları	Yok, Var
24	Dış çürüme	Yok, Sağlam ve kuru, Islak
25	Miselyum	Yok, Var
26	Renk solması	Yok, Kahverengi, Siyah
27	Sclerotia	Yok, Var
28	Meyve bölmeleri	Normal, Hastalıklı, Çok az, dna
29	Meyve lekeleri	Yok, Renkli, Kahverengi lekeli, Çarpık, dna
30	Tohum	Normal, Anormal
31	Küf gelişimi	Yok, Var
32	Rengi değişik tohum	Yok, Var
33	Tohum büyüklüğü	Normal, Normalin altı
34	Buruşma	Yok, Var
35	Kökler	Normal, Çürümüş, Tümörlü
36	Sınıf adları	D1, D2, D3, D4

Hayvanlara Ait Veri Seti

Bu veri seti 101 hayvana ait bilgi içermektedir. Veri setinde 18 değişken vardır ve bunlardan birisi hayvanların isimleri, 15 tanesi iki sonuçlu (binary) değerler ve iki tanesi de sayısal değerler içermektedir. Sayısal değişkenlerden birisi hayvanların bacak sayısını vermiştir fakat bunlar kategorikmiş gibi alınıp her birine aynı önem verilecek şekilde analize dahil edilecektir.Yani 2 bacağa sahip olanlar 4 bacağa sahip olanlara daha benzer olmayacak

ISSN: 2536-4383

şekilde 0,2,4,5,6,8 bacağa sahip hayvanların hepsi aynı önemde olacaktır. Diğer sayısal değere sahip olan değişken ise hayvanların ait oldukları sınıfları verdiği için analize dahil edilmeyecektir. 7 tane sınıf var ve hiç kayıp gözlem yoktur. Tablo 3 değişkenleri ve değişkenlerin değerlerini tanımlamakta, Tablo 4 ise hangi hayvanın hangi sınıfa ait olduğunu vermektedir.

Tablo 3. Hayvanlar veri setine ait değişkenler

Değişken Numarası	**Değişken İsmi**	**Değişken Değeri**
1	Hayvan isimleri	Her biri için ayrı ayrı 101 tane
2	Saç	İki sonuçlu
3	Tüyleri	İki sonuçlu
4	Yumurta	İki sonuçlu
5	Süt	İki sonuçlu
6	Uçabilme	İki sonuçlu
7	Suda yaşama	İki sonuçlu
8	Yırtıcılık	İki sonuçlu
9	Dişli	İki sonuçlu
10	Omurgalı	İki sonuçlu
11	Soluma	İki sonuçlu
12	Zehirli	İki sonuçlu
13	Yüzgeçli	İki sonuçlu
14	Bacak	Sayısal (0,2,4,5,6,8)
15	Kuyruk	İki sonuçlu
16	Evcil	İki sonuçlu
17	Kedi boyutlu	İki sonuçlu
18	Sınıfları	Sayısal (1,2,3,4,5,6,7)

Tablo 4. Hayvanlara ait veri seti için sınıfların dağılımı

Sınıf Numarası	Veri sayısı	Sınıf Dağılımı
1	41	Yaban domuzu, antilop, ayı, domuz, bufalo, dana, Güney Amerika'ya özgü kobay, çita, geyik, yunus, fil, yarasa, zürafa, kız, keçi, goril, hamster, tavşan, leopar, aslan, vaşak, vizon, köstebek, firavun faresi, keseli sıçan, Afrika antilopu, ornitorenk, kokarca, midilli, domuz balığı, puma, kedi, rakın, ren geyiği, ayı balığı, deniz aslanı, sincap, vampir, tarla faresi, küçük kanguru, kurt
2	20	Tavuk, karga, güvercin, ördek, flamingo, martı, şahin, kivi kuşu, tarla kuşu, deve kuşu, muhabbet kuşu, penguen, sülün, Amerika deve kuşu, sıyırıcı, yırtıcı martı, serçe, kuğu, akbaba, çalıkuşu
3	5	Çukur engerek, deniz yılanı, kör yılan, kaplumbağa, tuatara
4	13	Levrek, sazan, yayın balığı, tatlı su kefali, kedi balığı, mezgit balığı, ringa balığı, pirana, turna balığı, denizatı, dil balığı, vatoz, ton balığı
5	4	Kurbağa, kurbağa, semender, kara kurbağası
6	8	Pire, sivrisinek, bal arısı, karasinek, uğur böceği, güve, beyaz karınca, arı
7	10	İstiridye, yengeç, kerevit, ıstakoz, ahtapot, akrep, deniz arısı, sümüklü böcek, denizyıldızı, solucan

Kongre Oylarına Ait Veri Seti

Bu veri seti 1984 yılında Amerika Birleşik Devletlerinde yapılan kongre oylarını içermektedir. Her bir gözlem kongredeki bir kişinin oylarını göstermektedir. Oylar vergisiz ihracat, göç etmek gibi 16 durumu

içermektedir. Her bir oy için cevap evet ya da hayırdır. Ne evet ne de hayır demeyenler kayıp gözlem olarak ifade edilmiştir fakat bunlar belirsiz olarak alınacaktır. Bu yüzden her bir kayıp gözlem üçüncü bir durumu ifade eden yani belirsiz anlamına gelen oyları temsil edecektir. 168 tane Cumhuriyetçi ve 267 tane Demokrata ait toplam 435 gözlem vardır. Tablo 5'de değişkenler tanımlanmaktadır. İlk değişken sınıf adlarını içerdiği için analize dahil edilmeyecektir. Veri setindeki kayıp gözlem sayısı 288'dir.

Tablo 5. Kongre oyları veri setine ait değişkenler

Değişken Numarası	**Değişken İsmi**
1	Sınıf adları
2	Özürlü bebekler
3	Su projesi maliyet paylaşımı
4	Bütçe devriminin benimsenmesi
5	Doktor ücretini dondurmak
6	El-Salvador yardımı
7	Okuldaki dini gruplar
8	Önleyici uydu test yasağı
9	Nikaragua'lılar için yardım
10	Füze karışımı
11	Göç etme
12	Şirket azaltma
13	Eğitim harcamaları
14	Süper fon dava hakkı
15	Suç
16	Vergisiz ihracat
17	Afrika dışında ihracat yönetim kanunu

Arabaların Değerlendirilmesine Ait Veri Seti

Bu veri seti arabaların modelleriyle ilgili bilgi içermektedir. 4 farklı model içeren veri setinde 1728 gözlem vardır ve araba modellerinin dağılımı birinci model için 1210, ikinci model için 384, üçüncü model için 69 ve dördüncü model için 65 tanedir.

Arabaların değerlendirilmesi için 6 değişken vardır ve bunlardan 2 değişken sayısaldır. Kapıların sayısını ve arabanın aldığı kişi sayısını içeren sayısal değişkenlerde her birine aynı önem verilecek şekilde kategorikmiş gibi davranılacaktır. Veri setinde kayıp gözlem yoktur. Tablo 6'da arabalar veri setine ait değişkenler yer almaktadır.

Tablo 6. Arabalar veri setine ait değişkenler

Değişken Numarası	**Değişken İsmi**	**Değişken Değeri**
1	Satın alımı	Çok yüksek, yüksek, orta, düşük
2	Bakımı	Çok yüksek, yüksek, orta, düşük
3	Kapıları	2, 3, 4, 5 ve daha fazla
4	Kişi sayısı	2, 4, 4 den fazla
5	Taşıma kapasitesi	Küçük, orta, büyük
6	Güvenlik	Düşük, orta, yüksek

Hastalıklı soya fasulyesi veri seti, hayvanlara ait veri seti, kongre oyları veri seti ve arabalara ait veri seti kullanılarak yapılan analizlerde hiyerarşik kümeleme tekniklerinden tek bağlantı tekniği, tam bağlantı tekniği ve ortalama bağlantı tekniği ile Jaccard benzerlik ölçüsü kullanılarak kümeleme yapılmıştır. Hiyerarşik kümeleme tekniklerinde 'k' küme sayısı kullanıcı tarafından belirlenmektedir. Kümeleme sonuçlarını düzgün bir şekilde değerlendirebilmek için bütün tekniklerde küme sayısı aynı alınmıştır ve küme sayısı belirlenirken verilerin uzmanlar tarafından belirlenmiş sınıf sayıları dikkate alınmıştır. Soya fasulyesi verisi için dört sınıf, hayvanlara ait veri seti için 7 sınıf, kongre oylarına ait veri seti için 2 sınıf ve arabalara ait veri seti için 4 sınıf vardır. K-modes algoritmasında da benzerlik ölçüsü olarak basit eşleşme katsayısı kullanılmakta ve yine 'k' küme sayısı önceden kullanıcı tarafından belirlenebilmektedir. Hiyerarşik kümeleme tekniklerinde de küme sayısı uzmanlar tarafından belirlenmiş sınıf sayısı ile eşit olarak alınmış ve sonuçlar buna göre değerlendirilmiştir.

Tablo 7' de algoritmaların doğru kümeleme yüzdeleri verilmiştir.

Tablo 7. Algoritmaların doğru kümeleme yüzdesi

Veri setleri ve Veri sayısı	Tek Bağlantı Tekniği	Tam Bağlantı Tekniği	Ortalama Bağlantı Tekniği	K-modes Algoritması
Soya fasulyesi veri seti	%100	%100	%100	%75
Hayvanlar veri seti	%86	%85	%88	%79
Kongre oyları veri seti	%61	%83	%61	%82
Arabalar veri seti	%4	%25	%22	%33

Kümeleme sonuçlarına göre 47 nesne ve 35 değişkene sahip soya fasulyesi verisinde hiyerarşik tekniklerin hepsi kümeleri net bir şekilde oluşturabilmiş ve hiç yanlış kümeleme yapılmamıştır. Hiyerarşik tekniklerden her üçü de %100'lük bir doğru kümeleme performansı sergilerken K-modes algoritması %75 oranında doğru kümeleme yapmıştır. K-modes algoritması hiyerarşik kümeleme tekniklerine göre daha düşük bir kümeleme performansı sergilemiş ve kümeleri hatasız olarak ayıramamıştır.

101 tane nesne ve 16 değişken içeren hayvanların sınıflandığı veri setinde bütün yöntemlerde kümeler net bir şekilde ayrılamamıştır fakat ortalama bağlantı tekniği %88, tek bağlantı tekniği %86, tam bağlantı tekniği %85 ve K-modes algoritması %79'luk bir kümeleme performansı sergileyerek iyi sonuçlar vermişlerdir.

435 nesne ve 16 değişken içeren kongre oylarına ait veri setinde ise soya fasulyesi ve hayvanlara ait veri setinde en iyi sonuçları veren tek bağlantı tekniği ve ortalama bağlantı tekniği daha düşük performans göstererek kümeleri ayıramamıştır. Buna karşılık K-modes algoritması ve tam bağlantı tekniği birbirine yakın sonuçlar vererek daha iyi kümeleme performansı sergilemişlerdir. Tek bağlantı tekniği ve ortalama bağlantı tekniği %61, tam bağlantı tekniği %83 ve K-modes algoritması ise %82 oranında doğru kümeleme yapmıştır.

Son olarak 1728 nesne ve 6 değişken içeren arabalara ait veri setinde bütün kümeleme yöntemleri çok düşük kümeleme performansı sergilemiştir. Tek bağlantı tekniği %4,

ortalama bağlantı tekniği %22, tam bağlantı tekniği %25 ve K-modes algoritması %33 oranında doğru kümeleme yapmıştır. Arabalara ait veri setinde kümeleme sonuçlarının kötü olmasının en önemli sebebi nesne sayısı çok fazla iken bunları açıklayan değişken sayısının çok az olmasıdır. Kümeleme analizinde iyi sonuçlar elde edebilmek için her zaman değişken sayısını yüksek tutmak gerekmektedir. Görüldüğü gibi kümelerin net bir şekilde ayrımı 47 gözlem ve 35 değişken içeren soya fasulyesi verisinde sağlanmıştır. Farklı veri ve değişken sayılarına göre yapılan analizlerde değişken sayısının fazla olması küme içi homojenlik kümeler arası heterojenlik kriterinin sağlanması için önemli bir unsur olduğu görülmektedir.

Tablo 7 incelendiğinde değişken sayısının fazla olduğu soya fasulyesi, hayvanlar ve kongre oylarına ait veri setlerinde veri sayısı arttıkça kümeleme performansının hiyerarşik tekniklerde azaldığı, bölmeli bir kümeleme tekniği olan K-modes algoritmasında ise arttığı gözlemlenmektedir. Arabalara ait veri setinde ise veri sayısı fazla iken değişken sayısının çok az olmasından dolayı düşük kümeleme performansları elde edilmiştir fakat en iyi kümeleme performansını K-modes algoritması sağlamıştır.

Sonuç ve Öneriler

İyi bir karşılaştırma yapmak için kategorik verileri kümelemede, literatürde en çok kullanılan gerçek veri setleri kullanılarak, tek bağlantı tekniği, tam bağlantı tekniği, ortalama bağlantı tekniği ve K-modes algoritmalarının kümeleme performansları değerlendirildi.

Kümeleme sonuçlarına göre 47 nesne ve 35 değişkene sahip soya fasulyesi verisinde hiyerarşik tekniklerden her üçü de %100'lük bir doğru kümeleme performansı sergilerken K-modes algoritması %75 oranında doğru kümeleme yapmıştır. 101 tane nesne ve 16 değişken içeren hayvanların sınıflandığı veri setinde ortalama bağlantı tekniği %88, tek bağlantı tekniği %86, tam bağlantı tekniği %85ve K-modes algoritması %79'luk bir kümeleme performansı sergileyerek iyi sonuçlar vermişlerdir. 435 nesne ve 16 değişken içeren kongre oylarına ait veri setinde ise tek bağlantı tekniği ve ortalama bağlantı tekniği %61, tam bağlantı tekniği %83 ve K-modes algoritması ise %82 oranında doğru kümeleme yapmıştır. Son olarak 1728 nesne ve 6 değişken içeren arabalara ait veri setinde tek bağlantı tekniği %4, ortalama bağlantı tekniği %22, tam bağlantı tekniği %25 ve K-modes algoritması %33 oranında doğru kümeleme yapmıştır.

Arabalara ait veri setinde kümeleme sonuçlarının kötü olmasının en önemli sebebi nesne sayısı çok fazla iken bunları açıklayan değişken sayısının çok az olmasıdır. Kümeleme analizinde iyi sonuçlar elde edebilmek için her zaman değişken sayısını

yüksek tutmak gerekmektedir. Görüldüğü gibi kümelerin net bir şekilde ayrımı 47 gözlem ve 35 değişken içeren soya fasulyesi verisinde sağlanmıştır. Farklı veri ve değişken sayılarına göre yapılan analizlerde değişken sayısının fazla olması küme içi homojenlik kümeler arası heterojenlik kriterinin sağlanması için önemli bir unsur olduğu görülmektedir.

Analiz sonuçlarına göre, değişken sayısının yüksek olduğu ilk 3 veri setinde hiyerarşik tekniklerden tek bağlantı, tam bağlantı ve ortalama bağlantı tekniklerinin kümeleme performansı azalırken, değişken sayısının az olduğu 4. veri setinde bölmeli bir kümeleme yöntemi olan K-modes algoritmasının kümeleme performansının arttığı gözlemlenmiştir. İyi bir kümeleme algoritmasının önemli özelliklerinden biri de büyük veri setlerine ve değişken sayısının fazla olduğu durumlara uygulanabilir olmasıdır.

Kaynaklar

[1] Guo L, 2008. Clustering Categorical Response, Master Thesis, Office of Graduate Studies College of Arts and Sciences, Georgia State University, 1-2.

[2] Triphaty B K, Ghosh A, 2011. SSDR: An algorithm for Clustering Categorical Data Using Rough Set Theory, Advances in Applied Science Research, 2(3):314-326.

[3] Huang Z, 1998. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values", Data Mining and Knowledge Discovery, 2(3): 283-304.

[4] Gibson D, Kleinberg J, Raghavan P, 1998. Clustering categorical data: an approach based on dynamical systems, In Proceedings of the 24th VLDB Conference, New York, USA, 311-322.

[5] Ganti V, Gehrke J, Ramakrishan R, 1999. CACTUS: Clustering categorical data using summaries, In Proceedings of ACM SIGKDD, International Conference on Knowledge Discovery&Data Mining, San Diego, CA, USA, 73- 83.

[6] Guha S., Rastogi R., Shim K, 1999. ROCK: A robust clustering algorithm for categorical attributes, Proceedings of the IEEE International Conference on Data Engineering, Sydney, 345-366.

[7] He Z, Xu X, Deng S 2002. Squeezer: An Efficient Algorithm for Clustering Categorical Data, Department of Computer Science and Engineering, Harbin Institue of Technology, 17(5):611-624.

[8] Rezankova H, 2009. Cluster Analysis and Categorical Data, Vysoka Skola Economicka v Praze, Praha, 223-234.

[9] Abdu E, 2009. Clustering Categorical Data Using Summaries and Spectral Techniques, PHD Thesis, Graduate Department of Computer Science, The University of New York, 1-3.

[10] Michel C, 2000. Cardinal Nominal or Similarity Measures in Comparative Evaluation of Information Retrieval Process, The 2st International Conference on Language Resources & Evaluation (LREC), 367.

[11] Barbara D, Couto J, Yi L, 2002. COOLCAT: An Entropy-based Algorithm for Categorical Clustering, In Proceedings of the 11th International Conference on Information and Knowledge Management (CIKM), McLean, VA, USA, 582- 589.

[12] Nemalhabib A, 2006. A Cohesion-based Clustering Technique for Categorical Data, Master Thesis, Graduate Department of Computer Science and Software Engineering, Concordia University, 1-37.

[13] Khan SS, 2007. Computation of Initial Modes for K-modes Clustering Algorithm Using Evidence Accumulation, IJCAI'07 Proocedings of the 20th International Joint Conference on Artifical Intelligence, 2784-2789.

[14] UCI Machine Learning Repository. http://www.ics.uci.edu/~mlear/MLRepository.html, 2013.

[15] Chaturvedi, A., Green, P., Carrol, J, 2001. "K-Modes Clustering", Journal of Classification, 18:35-55.