KARMA TİPTEKİ VERİLERİ KAMILA, K-ORTALAMALAR, K-ORTAYLAR ve K-PROTOTİPLER ALGORİTMALARIYLA KÜMELEME: KARŞILAŞTIRMALI BİR UYGULAMA

Emrah BİLGİÇ[1]

Özet

Kümeleme Analizi Sosyal Bilimlerden Fen Bilimlerine birçok alanda yaygın olarak kullanılan önemli bir araçtır. Kümeleme Analizini gerçekleştirebilmek için hazırlanmış pek çok algoritma mevcuttur. Günümüzde bu algoritmalar ile ilgili olarak en çok tartışılan hususlardan ilk ikisinin, karma tipteki veri setleri için hangi kümeleme algoritmasının kullanılması gerektiği ve en iyi küme sayısının nasıl belirlenebileceği olduğu söylenebilir. Bu çalışmada, farklı ölçeklerle ölçülmüş karma tipteki değişkenlerin değerlerini içeren bir veri seti, bu tip veriler için yeni ve çok iddialı bir şekilde oluşturulmuş olan KAMILA algoritması ile analiz edilecektir. Daha sonra veri seti bu algoritmadan önce karma tipteki veriler için kullanılagelen k-ortalamalar, k-ortaylar ve k-prototipler gibi algoritmalarla da kümelere ayrılacaktır. Bu doğrultuda, İstanbul'da faaliyet gösteren yerel bir süpermarket zincirinden sağlanan alışveriş işlem verileri, R programlama dili kullanılarak analiz edilmiştir. Mağazaları İstanbul'un farklı semtlerinde bulunan bu firmanın müşterileri farklı demografik özelliklere ve farklı satın alma davranışlarına sahiptir. İşlem kolaylığı açısından 999 müşteri için sağlanmış olan veri kümesi, müşterilerin firmanın kârlılığı açısından önem arz eden ürün kategorilerinden alış veriş yapıp yapmadıklarını ve satın alınan ürünlerin toplam fiyatının ne kadar olduklarını içermektedir. Bu veriler müşteri segmentasyonu amacıyla kümeleme analizine tâbi tutulmuştur. Sonuç olarak, KAMILA algoritmasının altın segment olarak isimlendirebilecek segmentteki müşterileri başarıyla tespit edebildiği gözlenmiştir.

[1] Dr. Öğr. Üyesi, Kayseri Üniversitesi, Develi Sosyal ve Beşeri Bilimler Fakültesi, emrahbilgic@kayseri.edu.tr, ORCID: 0000-0002-9875-2299

A Comparative Application on Clustering of Mixed-type Data Sets with kamila, k-means, k-medoids and k-prototypes Algorithms

Abstract

Cluster Analysis is one of the crucial tools which is being used in many areas of scientific researches. As known, there are many algorithms for performing Cluster Analysis. Nowadays, the main two debates relating to these algorithms are; which one to use for mixed-type data sets and how to decide selecting the best number of clusters. In this study, KAMILA algorithm which is created very ambitiously and other algorithms used before KAMILA such as k-means, k-medoids and k-prototypes algorithms will be performed for clustering the values of different scaled variables. With this aim, a data set of a grocery store in Istanbul will be analyzed. The company has stores in different districts of Istanbul and the customers have different demographic characteristics and different purchasing behaviors. The data set provided for 999 customers includes information such as; whether the customers are purchasing the product categories that are crucial for the company's profitability and how much the total price of the purchased items are. These data were subjected to clustering analysis for customer segmentation. As a result, it is observed that KAMILA algorithm can successfully identify the customers in the segment that can be named the gold segment.

Key Words: Mixed-type data sets, Cluster Analysis, KAMILA Algorithm

GİRİŞ

İçerisinde bulunduğumuz Büyük Veri çağında günlük yaşantımız adeta verilerle çevrelenmiş durumdadır. Kişiler, şirketler ve kurumlar tarafından oluşturulan bu veriler yalnız sayısal değil aynı zamanda yalnız sözel veya karma yapıdadırlar. Dolayısıyla günümüzde Sosyal Bilimleri ilgilendiren verilerde tek tip ölçekle ölçülmüş değişkenlerle karşılaşmak oldukça güçtür. Boyut olarak çok büyük olan günümüz verileri nominal, ordinal ve oranlı ölçeklerle ölçülmüş sürekli ve kategorik değişkenleri aynı anda barındırabilmektedirler.

Kümeleme Analizi, Gözetimsiz Öğrenme tekniklerinden biri olup bir veri kümesini, veri noktalarının birbirleri ile olan benzerliklerini dikkate alarak gruplara ayırmaktadır. Bahsedilen karma tipteki veriler için kümeleme analizi yapabilmek oldukça problemli olduğundan araştırmacılar farklı yaklaşımlarla yeni algoritmalar türetmektedirler. Ayrıca ilgili alan yazında bu tip verilerin kümelenmesinden sonra en iyi küme sayısının kaç olarak alınacağı tartışması da devam etmektedir.

Bu çalışmada kümeleme analizi ve karma tipteki verileri kümeleme problemi ile ilgili literatür incelendikten sonra, bir süpermarketten sağlanan müşterilerin satın alma işlem verilerini içeren karma yapıdaki veri kümesi KAMILA (Foss vd., 2016) algoritması ile ve ilgili literatürdeki diğer algoritmalar olan k-ortalamalar, k-ortaylar ve k-prototipler algoritmaları ile kümelere ayrılacaktır. Bu analizin amacı firmanın sahip olduğu müşterileri daha önceden sayısı bilinmeyen segmentlere ayırmak, aynı zamanda KAMILA algoritmasının performansını kendisinden önce kullanılagelen diğer algoritmalarla karşılaştırmaktır. Elde edilen sonuçlardan en iyisinin hangisi olduğuna *dirsek noktası*, *ps* değeri ve *ortalama silüetler* gibi ölçülerle karar verilecektir. Ayrıca farklı algoritmalarla elde edilen kümelemelerin uyum içinde olup olmadıkları da karşılaştırılacaktır.

Bu çalışmanın amacı, karma tipteki verilerin yeni oluşturulmuş ve ileride birçok çalışmada kullanılabilecek potansiyele sahip KAMILA algoritması ile pazarlama alanında bir kümeleme analizi uygulaması yapmak ve algoritmanın performansını daha önce karma tipteki verileri kümelemek için kullanılagelen diğer algoritmalarla karşılaştırmaktır. Bu çalışma ilgili literatürü detaylı bir şekilde inceleyen Türkçe dilindeki ilk çalışma olup aynı zamanda KAMILA algoritmasının pazarlama alanındaki ilk uygulamasıdır. Ayrıca, ilerleyen bölümlerde inceleneceği üzere karma tipteki verileri kümeleme konusunda yapılan pazarlama uygulamalarının sayısı oldukça az olduğundan ve günümüz gerçek hayat veri kümeleri karma tipte olduğundan böyle bir çalışmanın gerekli olduğu düşünülmektedir. Son yıllarda yapılan çalışmaların birçoğu sadece satış işlem verileri ile yapılmış (tek tip değişken) ve araştırmacılar Birliktelik Kuralları Analizini kullanarak satın alınan ürün gruplarını tespit etmiş ve elde edilen sonuçlar kümeleme analizine tâbi tutularak müşteri segmentasyonu gerçekleştirilmiştir. Bu çalışmada daha önce yapılan çalışmalardan farklı olarak karma yapıdaki bir veri kümesi kullanılıp müşteri segmentasyonu gerçekleştirilecektir.

I. Kümeleme Analizi

Bir veri setini dağılımı ile ilgili herhangi bir bilgiye sahip olmaksızın anlayabilmek ve hakkında yorumlar yapabilmek ancak kümeleme analizi ile mümkün olabilmektedir. Kümeleme analizinin en önemli gözetimsiz öğrenme tekniklerinden biri olmasının sebebi budur. Kümeleme, kısaca gruplanmamış veri noktalarını küme adı verilen doğal gruplara ayırma işidir. Gruplara ayırmada göz önünde bulundurulan kriter ise aynı grup içerisindeki veri noktalarının birbirine çok benzemesi ve farklı kümelerdeki veri noktalarının da birbirine

benzememesini sağlamaktadır. (Zaki ve Meira Jr, 2014, 28; Kalaycı, 2010, s. 349; Alpar, 2011, 309).

Kümeleme analizi sonlu bir grup nesne için yapılan bir sınıflandırma çeşididir. Nesneler arasındaki ilişkiler bir yakınlık matrisinde temsil edilirler. Eğer bahsi geçen nesneler d-boyutlu ölçü uzayındaki örnekler veya veri noktaları ise, yakınlıklar nokta çiftleri arasındaki uzaklıklardır (örneğin Öklid uzaklığı gibi). Eğer, nokta çiftleri arasında bir uzaklık veya benzerlik olmazsa kümeleme analizi yapılamaz. Bunun sebebi benzerlik (yakınlık) matrisinin kümeleme analizini yapabilmek için tek girdi olmasındandır (Jain ve Dubes, 1988:55). Dolayısıyla kümeleme analizi ham veri kümesi ile değil, kümelemesi yapılacak gözlem çiftlerinin benzerlik (yakınlık) veya benzeşmezlik (uzaklık) değerlerinin oluşturduğu matris ile yapılabilmektedir. Ayrıca bu analizde verilerin normalliği varsayımı fazla önemli olmayıp uzaklık değerlerinin normalliği yeterli görülmektedir (Tatlıdil, 1992, s.252).

Kümeleme analizi yapmak için izlenecek adımlar en genel anlamda; değişkenlerin seçimi, uzaklık fonksiyonu ile kümeleme algoritmasının seçimi, kümelemenin geçerliliği ve sonuçların yorumlanması şeklinde sıralanabilir. Bu aşamaların her biri birbiriyle bağlantılı olup birbirlerinden karşılıklı beslenirler ve elde edilen kümeleri yani kümeleme analizinin sonucunu belirleyen etmenler olarak karşımıza çıkarlar. Kümeleme analizinde izlenilen yöntemler en genel anlamda iki alt gruba ayrılabilir. Bunlar hiyerarşik ve hiyerarşik olmayan yöntemlerdir. Erilli (2009), bu iki alt grubu aşağıdaki gibi detaylanmıştır:

i. Hiyerarşik Kümeleme Yöntemleri

a. Toplanmış Hiyerarşik Kümeleme

i. Tek Bağlantı Kümeleme

ii. Ortalama Bağlantı Kümeleme

iii. Tam Bağlantı Kümeleme

iv. Mc Quitty Bağlantı Kümeleme

v. Küresel Ortalama Bağlantı Kümeleme Yöntemi

vi. Ortanca Bağlantı Kümeleme Yöntemi

vii. Ward Bağlantı Kümeleme Yöntemi

b. Bölen Hiyerarşik Kümeleme

ii. Hiyerarşik Olmayan Kümeleme Yöntemleri

a. Merkez Tabanlı Kümeleme

i. K-Ortalamalar Yöntemi

ii. K-Medoid Yöntemi

b. Yoğunluk Tabanlı Kümeleme

i. DBSCAN Yöntemi

ii. DENCLUE Yöntemi

iii. OPTISC Yöntemi

c. Izgara (Grid) Tabanlı Kümeleme

i. STING Yöntemi

ii. WAVECLUSTER Yöntemi

iii. CLIQUE Yöntemi

d. Kategorik Kümeleme

i. Rock Yöntemi

ii. K-Mode yöntemi

iii. K-Mod Yöntemi

e. Bulanık Kümeleme

A. Araştırmaya Konu Olan Kümeleme Yöntemleri

1. Merkez Tabanlı Kümeleme Yöntemleri

Daha önceden belirlenen bir kriter ışığında veri setini direk olarak belirli sayıdaki alt gruplara böler. Bu kriter, veri setinin tamamını ilgilendiren veya oluşacak kümeleri ilgilendiren, ve yinelemeli (iterative) şekilde optimize edilecek bir kriter olabilir. k-ortalamalar, k-ortaylar ve k-prototipler en yaygın bölümleyici algoritmalardır.

k-ortalamalar algoritmasında izlenen adımlar şu şekilde gösterilebilir.

- Atama adımı: Bu adımda her bir veri noktası başlangıçta belirlenen küme sayısı kadar kümelerden kendisine en yakın olanına atanır.

- Merkez güncelleme adımı: Bu aşamada küme ortalamaları hesaplanır. Daha önceden belirlenen kriter sağlanıyorsa algoritma durur, sağlanmıyorsa yeni veri noktaları kümelere atanarak tekrar küme ortalamaları hesaplanır ve algoritma bu şekilde yinelenir.

k-ortaylar algoritmasında izlenen adımlar aşağıdaki şekilde özetlenebilir. K-ortaylar k-ortalamalardan farklı olarak uzaklık karelerini minimize etmek yerine, gözlem değerleri ile ortay olarak seçilen gözlemin arasındaki mutlak uzaklıkları minimize eder.

- Başlangıç adımı: Başlangıçta belirlenen küme sayısı kadar veri noktası küme temsilcileri olarak seçilir.
- Atama adımı: Bu adımda her bir veri noktası kendisine en yakın kümeye (temsilciye) atanır.
- Uzaklık hesaplama adımı: Bu adımda küme içi uzaklıklar toplamı hesaplanır, en küçük uzaklık bulunduysa algoritma durur. Daha küçük bir uzaklık mümkünse küme içi en küçük uzaklık bulununcaya kadar atama adımı tekrarlanır.

k-prototipler algoritmasında izlenecek olan adımlar aşağıdaki şekildedir. Bu algoritma sayısal veriler için k-ortalamalar, sözel veriler için k-modlar tekniğini esas alarak çalışır.

- Başlangıç adımı: Daha önceden belirlenen küme sayısı kadar veri noktası kümeler için prototip teşkil etmesi amacıyla seçilir.
- Atam adımı: Her bir gözlem değeri kendisine en yakın kümeye atanır. Burada uzaklık matrisi esas alınır.
- Bütün veri noktaları kümelere atandıktan sonra prototipler ile veri noktaları karşılaştırılır. Bir veri noktası bulunduğu kümedeki prototip yerine başka bir kümedeki prototipe benziyorsa, prototipler güncellenir.
- Kümelerde değişecek prototip kalmayıncaya kadar son işlem tekrarlanır.

2. Karma Tipteki Verileri Kümeleme Yöntemleri

Son yıllarda kümeleme algoritmalarının sayısında her ne kadar çok hızlı bir artış gözlemlense de, karma tipteki verilerin kümelemesi için oluşturulmuş olan algoritmaların sayısında artış gözlenmemektedir. Günümüz Büyük Veri ve Endüstri 4.0 çağında karma tipteki veri setleri ile karşılaşmak kaçınılmazdır. Çünkü gerek internet ortamında kullanıcıların bıraktıkları sayısal yapıdaki (satın alma verileri, tıklama sayıları, sayfa ziyareti süreleri vb.) ve sözel yapıdaki (yapılan yorumlar, eleştiriler, şikâyetler vb.) veriler, gerek WiFi, sensorlar ve mobil cihazlar aracılığı ile oluşan yapısal olmayan veriler bir araya getirildiklerinde yapısal olmayan karma veri setleri karşımıza çıkmaktadır. Kümeleme analizi her çeşit veri setini

başarıyla anlamlı gruplara ayırabildiğinden sadece sayısal verilerin kullanıldığı araştırmalarda değil, görüntü analizi, web analizi, metin madenciliği gibi alanlarda da yaygın kullanılmaktadır. Karma tipteki veri setleri için yapılan uygulamalı kümeleme analizi çalışmalarında genel olarak aşağıda detayları verilen yollar takip edilmelidir. (Foss vd., 2019):

Araştırmacılar tarafından başvurulan yöntemlerden biri, farklı tipteki değişkenlerin tek tip haline dönüştürülmesidir. Örneğin kategorik veriler, nümerik veri haline dönüştürüldükten sonra Minkowski, Öklid gibi uzaklık ölçüleri hesaplanarak kümeleme analizi yapılmaktadır. Bazı çalışmalarda nümerik veriler kategorik hale dönüştürüldükten sonra işlem yapılmıştır. Bu gibi yöntemlerde kategorik veya nümerik değişkenlere gereğinden fazla ağırlık verilmesi kaçınılmazdır (Hennig ve Liao, 2013).

Kategorik verileri nümerik hale dönüştürmeden, basit eşleştirme gibi yöntemlerle de kümeleme analizi yapmak da mümkündür. Eğer bir veri setindeki iki veri noktası herhangi bir kategorik değişken için aynı değeri almışlarsa 1, farklı değer almışlarsa 0 puan atayarak elde edilecek matris ile kümeleme analizi yapılabilir. Örneğin k-prototipler algoritması, kategorik verileri basit eşleştirme tekniği ile ele alırken nümerik verileri de Öklid uzaklığı ile analize tâbi tutar. k-prototipler algoritması daha önce vurgulandığı gibi k-ortalamalar ve k-modlar tekniklerinin karışımıdır. Öklid uzaklık ölçüsünün kullanıldığı k-ortalamalar metodundan yola çıkılarak bu algoritmaya kategorik değişkenler için de bir uzaklık ölçüsü eklenmiş, dolayısıyla k-prototipler nümerik ve kategorik değişkenli karma verileri de kümeleyebilir hale gelmiştir (Huang, 1998). Ji vd. (2013), veri noktaları ve kümelerin prototipleri arasında yeni bir benzeşmezlik ölçüsü hesabı geliştirmiş k-prototipler algoritmasını genişletmişlerdir.

Farklı tipteki değişkenler için hazırlanmış olan uzaklık ölçüleri de mevcuttur(örneğin: Gower, 1971). Bu tip ölçülerle elde edilen uzaklık matrisleriyle kümeleme analizi gerçekleştirilebilir. Gower'in ölçüsü (d_G) ile Öklid'in uzaklık ölçüsü (d_E) arasındaki ilişki ve farklılıklar bir örnekle şu şekilde açıklanabilir: d_E, bilindiği gibi, farkların (uzaklıkların) karesini almaktadır. Bu durum büyük farklara sahip değişkenlere daha çok ağırlık vermektedir. Dolayısıyla eğer bir değişken için iki gözlem arasındaki fark çok büyükse ve diğer değişkenler için bu fark küçükse bu gözlemler birbirlerine, bütün değişkenler için farkların orta seviyede olduğu durumundakinden daha az benzerdir diyebiliriz. Pazarlama açısından düşünüldüğünde aynı segmentte yer alabilecek müşterilerin, örneğin satın alma tutarları ve eğitim seviyeleri değişkenleri için farklılıkların çok büyük olması istenmeyen bir durumdur. Bu durum d_E

uzaklığını üstün kılmaktadır. Çünkü vurgulandığı gibi d_E farkların karesini alarak birbirine benzemeyen gözlemleri ayrı kümelere atmaktadır (Hennig, 2013).

Karma tipteki verileri kümelemek için başvurulabilecek son yöntem modele dayalı yöntemlerdir. Bu yöntemlerde temel varsayım gözlemlerin normal-multinomial sonlu karma modellerini izlediğidir. Fakat bu varsayım sağlanmadığı takdirde kümeleme analizi sağlıklı olarak yapılamamaktadır. Örneğin Gauss Karma modelinde X_i vektörlerinin Gauss dağılımlarından birine uyduğu varsayılmaktadır. Karma modellerin parametreleri Beklenti Maksimizasyonu yöntemi ile kestirilmektedir. Sürekli değişkenler için Kernel Yoğunluk Fonksiyonu kullanılarak varsayımlar biraz yumuşatabilmekle beraber bu işlemi uygulamak oldukça fazla zaman ve işlem gerektirmektedir (Foss vd., 2019:13).

Karma tipteki verileri kümeleme çalışmalarına pazarlama uygulamalarında sık rastlanmamaktadır. Yu vd. (2006) kernel k-yığışım (k-aggregate) algoritması ile karma tipteki müşteri verilerini kümelere ayırmış ve uzmanların da onayıyla işe yarar müşteri segmentleri elde etmiştir. Buradaki temel fikir kümeleme analizini Kernel metodu ile harmanlamaktır. Hsu ve Chen (2007), CAVE isimli algoritmayla katalog pazarlama amaçları doğrultusunda bir müşteri segmentasyonu gerçekleştirmiştir. Algoritma varyans ve entropiye dayalı bir mantıkla çalışmaktadır. Varyans sürekli değişkenler arasındaki benzerlikler için kullanılmakta, uzaklık hiyerarşisi de kategorik değişkenler arasındaki uzaklığı belirlemektedir. Kategorik değişkenlerin birbirleri ile olan uzaklıklarının hesaplanmasında ayrıca entropi de kullanılmaktadır. Morlini ve Zani (2010), karma tipteki değişkenleri içeren bir veri kümesinin farklı yöntemlerle analiz edilmesi ve sonuçların karşılaştırılması ile ilgili uygulamalı çalışmasında, Gower'in uzaklık ölçüsünü kullanmıştır. Bir ürüne ait farklı markaların piyasaya sunmuş oldukları modelleri inceleyen çalışmada, kategorik ve nümerik değişkenlerden elde edilmiş gözlem değerleri ile çalışılmış, Gower'in formülünün oluşturduğu uzaklık matrisi k-ortalamalar tekniği ile kümelere ayrılmıştır. Ji vd. (2013) k-prototipler algoritmasına bazı eklemeler yapıp geliştirmiş ve bir pazarlama uygulamasında kullanmışlardır. Swenson vd. (2016) sağlık pazarını segmente etmek amacıyla, ABD'deki bir hastaneden hizmet almış 700 hastaya ait demografik veriler ve uygulanan anket yoluyla elde edilen verileri kullanarak kümeleme analizi gerçekleştirmişlerdir. Çalışmada uzaklık ölçüsü olarak Gower'in formülü ve kümeleme analizi tekniği olarak da k-ortalamalar kullanılmıştır.

a. KAMILA Algoritması

KAMILA algoritması (Foss vd., 2016) karma tipteki verileri kümeleme için oluşturulmuş diğer algoritmalardan farklı olarak, çok güçlü parametrik varsayımlar yapmaksızın sürekli ve kategorik değişkenlere dengeli bir şekilde ağırlık atayarak kümeleme analizini gerçekleştirmektedir. Literatürdeki önceki çalışmalarda, değişkenlerin kümelemedeki ağırlıklarını belirleme işini ya kullanıcıya bırakmakta ya da ağırlıklar sürekli değişkenlerden veya kategorik değişkenlerden yana baskın olarak atamaktaydılar. k-ortalamalar tekniğinin yarı-parametrik genelleştirilmesi olan KAMILA algoritması ise sürekli ve kategorik değişkenlerin ağırlıklarını (katkılarını) otomatik olarak dengelemektedir.

KAMILA algoritmasının özellikleri kısaca şu şekilde özetlenebilir: İlk olarak algoritma, kullanılacak olan değişkenlerin tiplerini değiştirmeden yani bütün değişkenleri kategorik yapıya veya sürekli yapıya dönüştürmeden de işlem yapabilmektedir. İkincisi, algoritma sürekli ve kategorik değişkenlere eşit etki (eşit ağırlık) şansı vermektedir. Üçüncü olarak algoritma sınırlayıcı etkisi olan parametrik varsayımları kullanmamakta ve son olarak bu algoritmada kümelemeye girdi olacak değişkenler için tek tek ağırlık atanmasına da gerek duyulmamaktadır. Daha önce vurgulandığı üzere, sürekli değişkenler için yapılabilecek varsayımlar Kernel Yoğunluk Fonksiyonu ile biraz yumuşatabilmektedir. Fakat bu işlemi uygulamak oldukça fazla zaman ve işlem gerektirdiğinden KAMILA algoritması bunun için alternatif bir yöntem olarak geliştirilmiştir.

μ merkezi etrafında küresel simetrik olan X sürekli rassal değişkenler vektörü için Foss vd.'nin (2016) ikinci önermesinde belirttiği gibi, X'in olasılık yoğunluk fonksiyonu Eşitlik 1'de verildiği gibi tanımlanabilir.

$$f_x(x) = \frac{f_R(r)\,\Gamma\left(\frac{p}{2}+1\right)}{p r^{p-1}\,\pi^{\frac{p}{2}}} \qquad (1)$$

Buradaki $R = \sqrt{(X-\mu)^T(X-\mu)}$ ve f_R, R'nin olasılık yoğunluk fonksiyonudur. Her bir kümenin küresel olduğu varsayımı altında yukarıdaki eşitlik her bir kümeyi modellemek için kullanılabilir. Ayrıca tek değişkenli Kernel yoğunluk tahmini R'nin yoğunluğunun bir tahmini ($\hat{f}_R$) için kullanılır. Bu sebeple f_R yerine $\hat{f}_R$ konulduğunda, $\hat{f}_x$ küme yoğunluk fonksiyonu elde edilir. Her bir küme için aynı veya farklı olabilen $\Sigma_g$ölçekleme matrisi yoluyla bu süreç eliptik (elliptical) kümelere genişletilebilir. Burada nominal değişkenler her bir küme içerisinde multinom olarak modellenmiştir. Bu arada, KAMILA her ne kadar bu durumla ilgili

olarak agnostik ve her türlü olasılık dağılımına uygun olsa da, vurgulandığı gibi nominal değişkenler her bir küme içerisinde multinom olarak modellenmiştir. Aralıklı ve nominal ölçekli değişkenlerin kümeye özel parametreleri Beklenti Maksimizasyonu algoritmasında olduğu gibi iteratif olarak tahmin edilir. Ayrıca *t.* iterasyonda f_R'nin Kernel yoğunluk tahmini şu şekilde kurgulanır.

$$\hat{f}_R^{(t)}(r) = \frac{1}{Nh^{(t)}} \sum_{i=1}^{N} k\left(\frac{r - r_i^{(t)}}{h^{(t)}}\right) \qquad (2)$$

Burada *N* örneklem büyüklüğü iken *k(.)* ise *t* iterasyonundaki ilgili bant-genişliği (bandwidth) parametreli ($h^{(t)}$) kernel fonksiyonudur. Ayrıca $r_i^{(t)}$ ise *i.* gözlem ile onun bir önceki iterasyondaki en yakın merkezle ($\mu^{(t-1)}$) arasındaki uzaklığıdır. Tahmin sürecinin detayları Foss vd.'nin (2016) çalışmasında verilmiştir. KAMILA daha önce belirtildiği üzere aralıklı ölçekle ölçülmüş değişkenleri kesikli hale veya nominal değişkenleri nümerik hale dönüştürmeye gerek duymamaktadır. Ayrıca aralıklı ve nominal ölçekle ölçülmüş değişkenler arasında geleneksel sonlu karma modellerine nazaran daha uygun bir denge kurmaktadır. KAMILA algoritması hem çarpık hem de kalın kuyruklu (heavy-tailed) aralıklı ölçekli küme dağılımlarında, normal-multinom karma modellerine üstün gelmektedir. Eğer kümelerin gerçek dağılımları biliniyorsa bu dağılımı kullanmak tercih sebebidir. Son olarak, diğer karma modellerde olduğu gibi nominal değişkenlerdeki kategorilerin sayısı arttıkça buna paralel olarak gözlem sayıları da arttırılmalıdır.

B. Kümelemenin Geçerliliği

Aynı veri seti için farklı kümeleme algoritmalarının sonuçlarını karşılaştırabilmek amacıyla veya hakkında daha önceden herhangi bir bilgi sahibi olmadığımız küme sayısını en doğru şekilde belirleyebilmek amacıyla bazı ölçülere ihtiyaç duyulmaktadır (Gan vd,, 2007). Kümeleme analizi sonucu elde edilebilecek küme sayıları gözlem çiftleri arasındaki benzerlikleri (benzeşmezlikleri) hesaplayan yönteme ve bölünmeleri sağlayacak olan algoritmaya doğrudan bağlıdır (Kassambara, 2017:128). Dolayısıyla en iyi küme sayısının ne olacağı hususunda kesin bir cevap bulunmamakla beraber yapılan birçok çalışmada kümelemenin geçerliliği meselesi sadece en iyi küme sayısını belirleme problemi ile eş görülmektedir.

Kümelemenin geçerliliği hakkında kapsamlı bir çalışmayı Halkidi vd. 2001 yılında gerçekleştirmiştir. Kümelemenin geçerliliği meselesi bu çalışmada üç bölümde incelenmiş olup konuyu bu şekilde ele almak en sağlıklı yol olarak görülmektedir. Daha önce yapılan kümeleme analizi çalışmalarında da geçerlilik aynı şekilde üç başlıkta incelenmiştir (Jain ve Dubes, 1988; Theodoridis ve Koutroubas, 1999). Buna göre geçerlilik üç ölçütle incelenebilir ki bunlar, Dışsal Ölçütler, İçsel ölçütler ve Göreceli Ölçütlerdir. Bu arada, ölçüt kavramı bir küme yapısının geçerliliğini belirlemek için izlenecek stratejiyi ifade ederken, ileride bahsedilecek olan geçerlilik ölçüleri de geçerliliği test etmeye yarayan istatistiklerdir (Jain ve Dubes, 1988). Bazı çalışmalarda (Maulik ve Bandyopadhyay, 2002) kümelemenin geçerliliği tek başlık altında incelenmiş ve içsel dışsal ayırımı yapılmadan, aşağıda bahsedilecek İçsel Ölçütler başlığındaki ölçüler tartışılmıştır. Bazı çalışmalarda ise kümelenin geçerliliği başlığı altında Dışsal Ölçütler konusu, İçsel ve Göreceli Ölçütlerden bahsedilmeksizin tek başına ele alınmıştır (Wu vd., 2009).

Dışsal ölçütlerde, herhangi bir kümeleme algoritması ile ortaya çıkan kümeler, aynı veri seti için daha önceden belirli olan kümelerle (bir uzmanın belirlediği kümeler de olabilir) veya başka bir algoritma tarafından elde edilen kümelerle karşılaştırılır. Burada kullanılabilecek ölçülerden bazıları, entropi ölçüsü, F ölçüsü, bilginin değişimi (variation of information) ölçüsü, ortak bilgi (mutual information) ölçüsü, Rand istatistiği, Düzeltilmiş Rand İstatistiği, Jaccard'ın katsayısı, Fowlkes ve Mallows ölçüsü, Hubert'in istatistiği gibi ölçülerdir (Gan vd., 2007:302; Jain ve Dubes, 1988).

İçsel Ölçütlerde, bir algoritma tarafından üretilen kümelemenin yapısı sadece veri kümesinden intikal eden bazı özellikler ve nicelikler kullanılarak değerlendirilir. Hiyerarşik bir kümelemede oluşan hiyerarşiler bu duruma örnek olarak verilebilir (Gan vd., 2007:303). Bu ölçütte konu ile alakalı literatürde en çok karşılaşılan ölçüler Kophenetik Korelasyon Katsayısı (Sokal ve Rohlf, 1962) ve Kaufmann ile Roussew'in (1990) önerdiği Birleştirici Katsayısıdır (Agglomerative Coefficient). Diğer taraftan uygulamalı çalışmalarda küme içi sapmalar ile kümeler arası sapmalar ve bunların oranları ile ilgilenen ölçüler de karşımıza çıkmaktadır. Pratik olmaları sebebiyle uygulamada en çok kırılma (dirsek) noktası tekniği, Silhouette İndeksi ve Gap İstatistiği kullanılmaktadır. Bunun dışında daha birçok ölçü de mevcuttur ve ölçü geliştirme çalışmaları devam etmektedir (Cui vd., 2017; Starczewski, 2017).

İçsel Ölçütler Liu vd.'nin (2010) çalışmasında sıkışıklık (bir kümede yer alan gözlemlerin ne kadar birbirine yakın olduğu) ve ayrışma (bir kümenin diğer kümelerden ne kadar uzak, farklı olduğu) isimli iki ana başlık altında detaylı olarak şu ölçülerle incelenmiştir:

Hubert'in değiştirilmiş Γ istatistiği, Davies-Bouldin indeksi, Dunn'ın indeksi, I indeksi, SD geçerlilik indeksi, RMSSTD indeksi, RS indeksi, Calinski Harabasz ve S_Dbw indeksi. Bu ölçüler bazı çalışmalarda kümeleme sonuçlarının kalitesini ölçen ölçüler olarak da karşımıza çıkmaktadır (Saitta vd., 2008; Tomašev ve Radovanović, 2016).

Üçüncü ve son olarak kümeleme sonuçlarının değerlendirilmesi konusunda Göreceli Ölçütler karşımıza çıkmaktadır. Bu ölçütteki temel fikir bir kümeleme yapısının aynı algoritmayla fakat başka parametrelerle üretilmiş olan kümeleme yapısıyla karşılaştırılmasıdır (Halkidi vd., 2001:123). Bu noktada akla ilk gelen örnek, hiyerarşik kümelemede karşımıza çıkan Tek, Tam ve Ortalama gibi Bağlantı yöntemleridir (Jain ve Dubes, 1988:161).

Bir veri setinde net olarak "doğru" kümelerden veya net olarak "en iyi" kümelemelerden söz etmek pek mümkün değildir. Bu yüzden araştırmacıların aradığı küme yapılarını daha önceden belirlemesi ortaya çıkabilecek küme yapılarını daha kolay değerlendirmesine olanak sağlar. Hennig ve Liao (2013) karma tipteki verileri kümeleme ile ilgili çalışmasında; iki kümeleme yaklaşımını; modele dayalı kümeleme yöntemleri (gizli sınıflar veya karma dağılımlar) ve temelinde olasılık olmayan uzaklığa dayalı (k-ortaylar yöntemi gibi) kümeleme yöntemlerini karşılaştırmaktadır. Çalışmada özellikle vurgulanan nokta, küme sayısı hususunun kesinlikle bir uzman tarafından ele alınması gerektiğidir. Çalışmada kullanılan veri kümesi farklı sosyo-ekonomik değişkenleri içeren veri kümesidir. Çalışmayı diğerlerinden ayıran özellik ise, çalışmanın sonuna ekledikleri, kümeleme alanında uzman 40'tan fazla araştırmacıdan aldıkları görüşlerdir. Bu görüşlerde de karma tipteki verilerin kümelenmesi probleminin araştırılan ve geliştirilen bir konu olduğu ayrıca Hennig ve Liao'nun küme sayısını bir uzmanın belirlemesi gerektiği görüşünün isabetli olduğu vurgulanmıştır.

II. Uygulama

Perakende müşterilerine ait veriler genellikle karma yapıdadırlar. Çünkü bu veriler ürün satın alma verilerini, TL veya başka bir para cinsinden tutarları hatta müşteriye mağaza kartı verilirken bir form doldurulmuşsa sosyo-ekonomik değişkenleri de içerebilmektedir. Dolayısıyla verilerde bir tarafta müşterilerle alakalı demografik bilgiler yer alabilirken diğer tarafta satın aldıkları ürünler ve satın alma tutar ve miktarları da yer alabilir. Bu çalışmanın amacı karma yapıdaki, perakende ile alakalı bir veri setini yeni oluşturulmuş ve ileride birçok

çalışmada kullanılabilecek potansiyele sahip KAMILA algoritması ile, ayrıca bu algoritmadan önce karma yapıdaki verileri kümelemek için kullanılagelen diğer algoritmalarla kümeleme analizine tabi tutmaktır. Bu kümeleme analizi uygulamasıyla analize konu olan firmanın müşterileri segmentlere (pazarlama literatüründeki pazar/müşteri segmentasyonu) ayrılacaktır.

Çalışmada uygulamaya konu olan veri setinde, firmanın pazarlama faaliyetleri için önemli gördüğü 8 ana ürün grubuna ait, 999 müşteri tarafından oluşturulmuş satın alma işlem verileri yer almaktadır. Müşterilerin bu ürünleri satın alıp almadıklarının yanı sıra bu ana ürün grubundan kaç adet satın aldığı ve alışveriş sepetinin TL cinsinden toplam tutarı da yer almaktadır. Şu durumda eldeki verilerde müşterilerin belirli ürünü satın alıp almadıkları (nominal) ve iki adet de nümerik değişken vardır. Firmanın amacı, bu müşterileri sahip olunan veriler ışığında daha önceden sayısı bilinmeyen segmentlere ayırmaktır. Firmalar müşterilerini farklı değişkenlere göre (demografik veriler, satın alma verileri) farklı sayıdaki segmentlere (genç-orta yaş-yaşlı, pırlanta-altın-gümüş-bronz…gibi) ayırmaktadır.

Tablo1: Uygulamada Kullanılan Veri Kümesi: İlk Beş Müşteriye Ait ve 999 Müşterinin Toplamına Ait Veriler

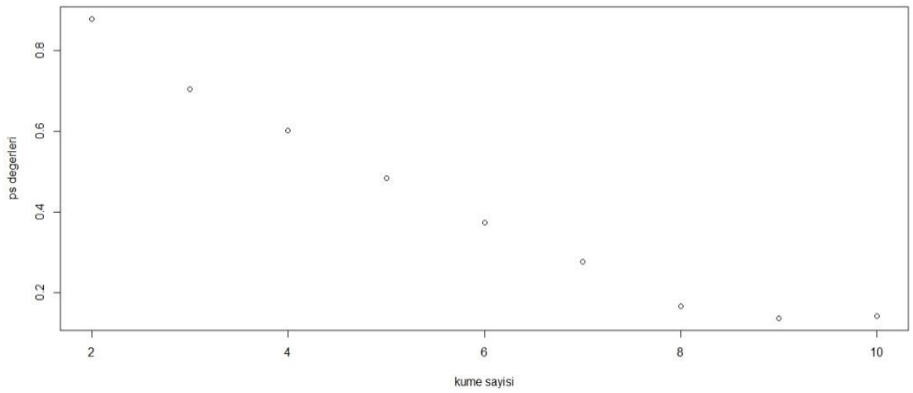
No	Kahvaltılık	Temizlik ürünü	Elbise	Kuru gıda	İçecek	Meyve	Ev gereçleri	Atıştırmalık	Toplam ödeme (TL)	Toplam ürün adedi
1	0	1	0	1	0	0	0	0	5.47	2
2	1	0	0	1	1	0	0	0	23.72	9
3	0	0	0	1	0	0	0	0	14.47	4
4	1	0	0	1	0	0	0	1	30.84	9
5	0	0	0	1	0	0	0	0	26.66	2
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
Toplam	592	416	14	714	322	243	113	425	34061.8	8226

Firma tarafından sağlanan ham veri seti, kahvaltılık reyonu, temizlik ürünleri, elbise, kuru gıda reyonu, içecek reyonu, meyve reyonu, plastik vb. ev gereçleri, atıştırmalık ürünler gibi, 8 ana ürün grubundan satın alınan ürünleri içeren ve ayrıca, toplam satın alınan ürün miktarı ve toplam ödenen tutarı da içeren bir veri setine dönüştürüldükten sonra Tablo 1'deki gibi bir veri seti elde edilmiştir. Buna göre, toplam satırında görüldüğü üzere 999 müşterinin 592'si kahvaltılık ürünler almış, 999 müşteri toplamda 8226 ürün satın almış ve bunlara 34 bin 61 TL ödemiştir.

Veri seti ilk olarak R programlama dilinde yer alan *kamila* paketindeki işlevlerle KAMILA algoritması kullanılarak kümeleme analizine tabi tutulmuştur. Kullanılan kodlar Ek 1'de sunulmuştur.

KAMILA algoritması ile yinelemeli şekilde kümeleme yapılmış ve gözlem noktalarının bulunduğu küme değişiklik göstermeyinceye kadar yineleme devam etmiş ve sonlanmıştır. En iyi küme sayısının belirlenmesi için Tibshirani ve Walther'in (2005) *ps* (prediction strength: kestirimin gücü) değeri kullanılmıştır. Daha önce vurgulandığı üzere, kümeleme analizinin en zor aşamalarından birisi, küme sayısına karar verebilmektir. Literatürde birçok ölçüt olmasına karşın, hangisinin en iyi küme sayısını verdiği hususunda bir uzlaşma bulunmamaktadır.

Kestirimin Gücü değeri kısaca şu şekilde açıklanabilir: Bu ölçüdeki anahtar fikir, kümeleme analizinin sınıflandırma analizi gibi görülmesi ve problemin doğru sınıf değerlerini tahmin etme şeklinde varsayılmasıdır. Bu ölçü ile veri setinden kaç küme oluşturulabileceği ve bu kümelemenin ne kadar iyi olduğu anlaşılabilmektedir. Bilindiği üzere bir sınıflandırma analizi probleminin temel hedefi kurulan bir model yardımıyla ve en az hata ile veri noktalarını daha önceden bilinen sınıflara atamaktır. Kestirimin Gücü ölçüsünde ilk önce veri setinden alınan bir eğitim veri seti (örneğin k-ortalamalar tekniği ile) *k* kadar kümeye ayrılır. Daha sonra eğitim veri setiyle *k* kadar elde edilen kümelerin merkezleri kullanılarak veri setinden alınan test veri seti de *k* kadar kümeye ayrılır. Son olarak eğitim veri seti kullanılarak yapılmış kümeleme analizi ve test veri setiyle yapılmış kümeleme analizinde aynı kümelerde yer alan veri noktalarının sayısının kaç olduğu sorusunun cevabı aranır. KAMILA algoritması yapıldıktan sonra Şekil 1'de görüldüğü üzere, kestirimin gücü değerinin en yüksek olduğu 0,87'nin karşılığı küme sayısı ikidir. Küme sayısının üç olduğu durumda *ps* değeri 0,70'e düşmektedir. Küme sayısı iki alındığında ve küme sayısı üç alındığında elde edilen kümelerin karşılaştırılması ile (Tablo 2), her iki kümelemede de müşterilerin çoğunun (708 müşteri) ilk kümede yer aldığı, 127'sinin ise her iki kümelemede de ikinci kümede yer aldığı gözlemlenmiştir.



Şekil 1: KAMILA algoritması ile elde edilen farklı küme sayılarına karşılık gelen kestirimin gücü (*ps*) değerleri

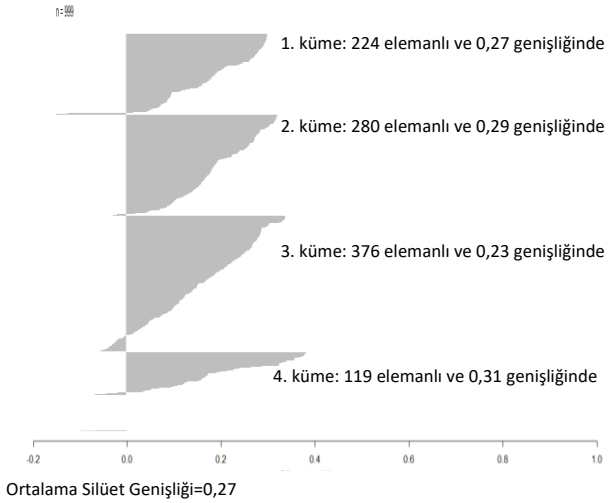
Dolayısıyla küme sayısını iki veya üç aldığımızda müşterilerin %83'ü aynı kümelerde yer almaktadır. Fakat Tibshirani ve Walther (2005) ps değeri 0,8 değerinden yüksek olan küme sayısının tercih edilmesini önermektedir. Dolayısıyla KAMILA algoritmasının ortaya çıkardığı kümeler için küme sayısını iki almak daha uygundur. KAMILA algoritması ile elde edilen iki kümede yer alan gözlemler (müşteriler) incelendiğinde Tablo 3'deki özet bilgilerle karşılaşılmaktadır.

Tablo2: KAMILA algoritması ile elde edilen iki ve üç kümenin uyumu

	1	**2**	**3**
1	708	105	0
2	0	127	59

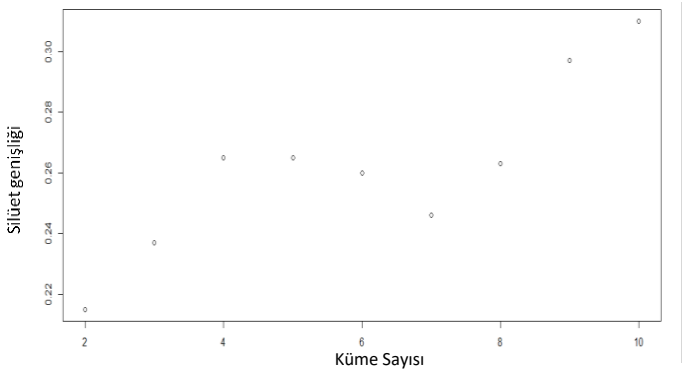
Veri setine k-ortaylar tekniğini, *pam* (partitioning around medoids) algoritmasını (Kaufmann ve Roussew, 1990) ve Gower'in uzaklık ölçüsü ile bulduğumuz uzaklık matrisi ile uyguladığımızda en iyi sonucun silüetler ortalaması hesabına göre dört küme olduğu tespit edilmiştir (Şekil 2). Sırasıyla iki küme için, üç küme için, dört küme için, beş küme için ve altı küme için silüetler ortalaması 0.22, 0.24, 0.2656, 0.2653, 0.2608 hesaplanmıştır (Şekil 3).

Ortalama silüet genişliği, aynı kümedeki veri noktalarının birbirleri ile olan uzaklıklarını en yakın başka bir kümede yer alan veri noktalarının birbirlerine olan uzaklıkları ile karşılaştırılmaktadır. Bu sayede bir küme ile komşu kümenin farklılıkları ortaya çıkarılmaktadır. Küme sayısı 4'ten sonra bu değer Şekil 3'de görüldüğü üzere düşmeye başlamıştır. K-ortaylar algoritması ile elde edilen dört kümeye ait bilgiler Tablo 4'te verilmiştir.

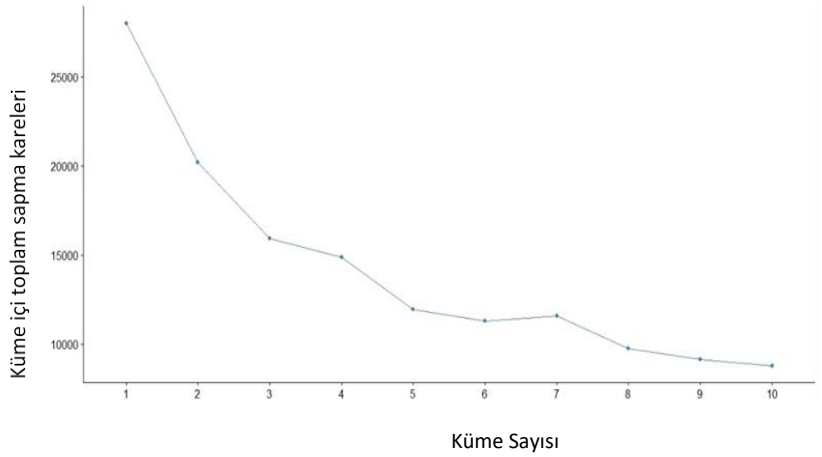


Şekil 2: K-ortaylar algoritması ile elde edilen dört kümeye ait silüet grafiği

Veri setine k-ortaylar algoritmasından sonra k-ortalamalar algoritması Gower'in uzaklık matrisi kullanılarak uygulandığında veri setinin üç kümeye ayrılması uygun görülmektedir. Şekil 4'ten de görüldüğü üzere, küme sayısı üçten sonra küme içi kareler toplamının düşüşü azalmaktadır. K-ortalamalar algoritmasının ortaya çıkardığı segmentlere ait özellikler ise Tablo 5'te verilmiştir.

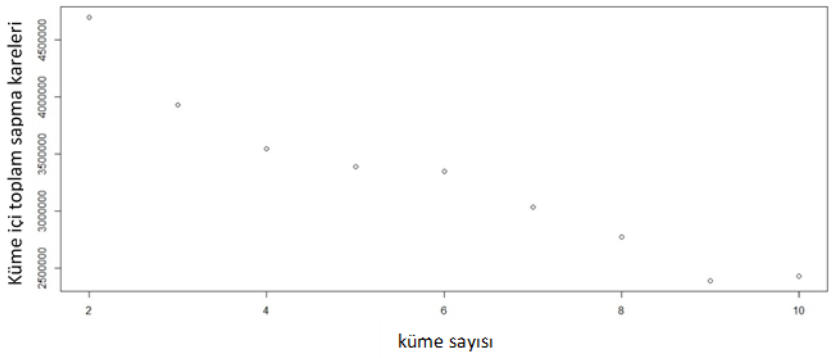


Şekil 3: k-ortaylar algoritması ile kümelerin aldığı ortalama silüetler değerleri



Şekil 4: k-ortalamalar algoritması ile elde edilen kümelere ait küme içi toplam sapma kareleri değerleri

k-prototipler algoritması ile elde edilen kümelerin küme için toplam sapma kareleri Şekil 5'te verilmiştir. Veri setini dört kümeye ayırmak uygundur. k-prototipler algoritmasının ortaya çıkardığı segmentlere ait özellikler ise Tablo 6'da sunulmuştur.



Şekil 5: k-prototipler algoritması ile elde edilen kümelere ait küme içi toplam sapma kareleri değerleri

Tablo 3: KAMILA algoritması ile elde edilen segmentlerdeki istatistikler

	Müşteri	K.altı	Temiz.	Elbise	Akşam	İçecek	Meyve	Alet	Atışt.	Tutar	Sayı
Veri Seti	999	592	416	14	714	322	243	113	425	34 TL	8
Küme 1	186	177	156	11	182	117	98	51	152	96 TL	23
Küme 2	813	415	260	3	532	205	145	62	273	20 TL	5

Tablo 4: k-ortaylar algoritması ile elde edilen segmentlerdeki istatistikler

	Müşteri	K.altı	Temiz.	Elbise	Akşam	İçecek	Meyve	Alet	Atışt.	Tutar	Sayı
Veri Seti	999	592	416	14	714	322	243	113	425	34 TL	8
Küme 1	261	123	91	4	0	57	47	18	83	13 TL	3
Küme 2	432	366	294	10	408	222	165	73	312	57 TL	14
Küme 3	306	103	31	0	306	43	31	22	30	18 TL	3

Tablo 5: k-ortalamalar algoritması ile elde edilen segmentlerdeki istatistikler

	Müşteri	K.altı	Temiz.	Elbise	Akşam	İçecek	Meyve	Alet	Atışt.	Tutar	Sayı
Veri Seti	999	592	416	14	714	322	243	113	425	34 TL	8
Küme 1	224	0	5	1	177	40	32	16	49	12 TL	3
Küme 2	280	280	27	2	192	43	16	23	65	23 TL	5
Küme 3	376	293	271	10	316	229	188	65	292	59 TL	15
Küme 4	119	19	113	1	29	10	7	9	19	22 TL	3

Tablo 6: k-prototipler algoritması ile elde edilen segmentlerdeki istatistikler

	Müşteri	K.altı	Temiz.	Elbise	Akşam	İçecek	Meyve	Alet	Atışt.	Tutar	Sayı
Veri Seti	999	592	416	14	714	322	243	113	425	34 TL	8
Küme 1	274	207	29	2	80	74	48	18	70	11 TL	3
Küme 2	152	142	138	7	149	103	93	41	123	100TL	24
Küme 3	240	190	59	3	221	82	48	28	192	32 TL	9
Küme 4	333	53	190	2	264	63	54	26	40	22 TL	4

KAMILA algoritması ile elde edilen iki kümedeki müşterilere ait kümeleme analizine girdi olan değişken değerleri incelendiğinde 186 müşterinin yer aldığı kümede ortalama 96 TL kadarlık alışveriş yapılmış ve müşterilerin sepetinde ortalama 23 ürün yer almıştır. Bu rakamlar ikinci kümedeki değerlere göre çok yüksektir. Ayrıca birinci kümedeki müşterilerin tamamına yakını kahvaltılık, temizlik malzemesi, akşam yemeği için çeşitli gıdalar ve atıştırmalık ürünler satın almıştır. İkinci kümede yer alan müşterilerin tamamına yakınının herhangi bir ürün grubunu satın alması söz konusu değildir. Firmanın pazarlama faaliyetleri (katalog postalama, sms, e-mail atma vb.) için yükleneceği maliyet düşünüldüğünde, birinci segmentteki alışveriş adetleri ve tutarları göz önüne alınırsa, bu müşterilerini tercih etmesi gerekmektedir. Bu müşterilerin pazarlama sonrası yapacakları ziyaretlerde harcanan maliyeti karşılayacağı beklentisi yüksektir.

Diğer algoritmaların ortaya çıkardığı segmentler incelendiğinde k-ortalamalar algoritmasının ortaya çıkardığı üçüncü segment, k-ortaylar algoritmasının ortaya çıkardığı

üçüncü segment ve k-prototipler algoritmasının ortaya çıkardığı ikinci segment dikkat çekmektedir. Fakat her üç algoritmanın çıkardığı segmentler birbirlerinden tam olarak ayırt edilememektedir. K-ortaylar algoritmasının ortaya çıkardığı 1.segment ile 3.segment, k-ortalamalar algoritmasının ortaya çıkardığı ikinci segment ile dördüncü segment ve son olarak k-prototipler algoritmasının ortaya çıkardığı 1.segment ile 4.segment yapı olarak birbirlerine benzemektedir.

Özellikle k-prototiplerin ortaya çıkardığı *ikinci* segmentte de KAMILA algoritmasında olduğu gibi müşterilerin tamamına yakını kahvaltılık, temizlik malzemesi, akşam yemeği için çeşitli gıdalar ve atıştırmalık ürünler satın almıştır. Dolayısıyla bu iki algoritmanın kümelemelerinin uyum içerisinde olduğu düşünülebilir. Kümelemelerin karşılaştırılması bölümünde bahsedildiği üzere aşağıdaki tabloda (Tablo 7) iki kümelemenin uyumu karşılaştırılmıştır. Genel çerçevede herhangi bir uyum veya görüş birliği (clusterings agreement) olmasa da üzerinde durulan segmentlerde (kamila algoritmasının ortaya çıkardığı birinci segmentteki 186 müşteri ile k-prototipler algoritmasının ortaya çıkardığı ikinci segmentteki 152 müşteri) toplam 146 müşteri vardır ve bu sayının firma için olumlu ve beklenen bir durum olduğu söylenebilir. Çünkü her iki algoritmanın da altın olarak isimlendirebileceğimiz segmentteki müşterileri bulabildiği sonucuna varılabilir.

Tablo 7: KAMILA ve k-prototipler algoritmalarının ortaya çıkardığı kümelerin kontenjans tablosu

Segmentler	**1**	**2**
1	0	274
2	146	6
3	34	206
4	6	327

Aşağıdaki tabloda ise (Tablo 8), kullanılan algoritmaların ortaya çıkardığı kümelemelerin birbirleriyle olan uyumları Düzeltilmiş Rand Değeri ile hesaplanmıştır. Görüldüğü gibi kümelemeler arasında güçlü bir görüş birliğinden söz etmek güçtür.

Tablo 8: Bütün algoritmaların birbirleri arasındaki uyumu

D. Rand indeksi	***k-ortalamalar***	***k-prototipler***	***k-medoidler***	***kamila***
k-ortalamalar	-	0,21	0,38	0,38

k-prototipler	0,21	-	0,17	0,29
k-medoidler	0,38	0,17	-	0,18
kamila	0,38	0,29	0,18	-

Sonuç

Büyük Veri, Nesnelerin İnterneti ve Endüstri 4.0 ile birlikte hayatımıza giren yeni kavramlar Veri Bilimi açısından birçok farklı kaynak aracılığıyla farklı tipteki verileri bir araya getirebilme fırsatı olarak düşünülebilir. Günümüz gerçek hayat verileri genellikle karma yapıda; hem sürekli hem de kesikli değişkenleri içermektedir. Dolayısıyla farklı ölçeklerle ölçülmüş değişkenler için veri analizi konusu günümüzde çok önem kazanmıştır.

Bu çalışmada karma yapıdaki verilere kümeleme analizinin hangi algoritmalarla uygulanabileceği gösterilmiş ve geçmişte ve günümüzde bu alanda yapılmış çalışmalar detaylı bir şekilde ele alınmıştır. Uygulama bölümünde, yeni geliştirilmiş olan ve ileride birçok çalışmada kullanılacağı öngörülen KAMILA algoritması ile bu algoritmadan daha önce kullanılagelen diğer algoritmalar kullanılmış ve uygulamaya konu olan süpermarket firmasının pazarlama bölümündeki karar verme süreçlerine yönelik bazı çıkarsamalarda bulunulmuştur. Buna göre firmanın analizi yapılan 999 müşterisini KAMILA algoritmasının ortaya çıkardığı iki segmente ayırması uygun görülmektedir. Birinci kümedeki müşterilerin tamamına yakını kahvaltılık, temizlik malzemesi, akşam yemeği için çeşitli gıdalar ve atıştırmalık ürünler satın almıştır. İkinci kümede yer alan müşterilerin tamamına yakınının herhangi bir ürün grubunu satın alması söz konusu değildir. Firmanın pazarlama faaliyetleri (katalog postalama, sms, e-mail atma vb.) için yükleneceği maliyet düşünüldüğünde, birinci segmentteki alışveriş adetleri ve tutarları göz önüne alınırsa, bu müşterilerini tercih etmesi gerekmektedir. Bu müşterilerin pazarlama sonrası yapacakları ziyaretlerde harcanan maliyeti karşılayacağı beklentisi yüksektir.

KAMILA algoritmasından önce karma yapıdaki veri setlerini kümeleme analizine tâbi tutmak için kullanılagelen k-ortalamalar, k-prototipler ve k-ortaylar algoritmalarının performansları da KAMILA algoritması ile karşılaştırılmıştır. Bu üç algoritmanın ortaya çıkardıkları segmentler incelendiğinde bazı segmentlerin yapı olarak birbirine benzediği, algoritmaların diğerlerinden farklı özellikleri bulunan segmentleri tam olarak ortaya çıkaramadığı gözlemlenmiştir. Bu bakımdan KAMILA algoritmasının daha iyi sonuç verdiğini söylemek mümkündür.

Daha önceki benzer çalışmalardan farklı olarak bu çalışmada karma tipteki veri kümesi kullanılmış, müşterilerin sadece satın alma işlemleri değil aynı zamanda alışveriş sepetindeki toplam ürün sayısı ve alışverişin toplam tutarı da veri setine dâhil edilmiştir. Gelecek çalışmalarda çok daha fazla gözlem ve daha fazla değişken ile karşılaştırmalı bir uygulama yapılması ve uygulama sonuçlarının pazarlama uzmanlarına yorumlatılması planlanmaktadır.

Kaynakça

Aggarwal, C. C. (2015). Data mining: The textbook, Switzerland: Springer.

Cui, H., Zhang, K., Fang, Y., Sobolevsky, S., Ratti, C., & Horn, B. K. (2017). A Clustering Validity Index Based on Pairing Frequency. IEEE Access, 5, 24884-24894.

Erilli, N. A. (2009). Kümeleme Analizine Bulanık Yaklaşım Algoritmaları ve Uygulamaları (Yüksek Lisans Tezi). On Dokuz Mayıs Üniversitesi Fen Bilimleri Enstitüsü, Samsun.

Everitt, B., & Hothorn, T. (2011). An introduction to applied multivariate analysis with R. Springer Science & Business Media.

Foss, A., Markatou, M., Ray, B., & Heching, A. (2016). A semiparametric method for clustering mixed data. Machine Learning, 105(3), 419-458.

Foss, A. H., Markatou, M., & Ray, B. (2019). Distance Metrics and Clustering Methods for Mixed-type Data. International Statistical Review, 87(1), 80-109.

Gan, G., Ma, C., & Wu, J. (2007). Data clustering: theory, algorithms, and applications (Vol. 20). Siam.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. Biometrics, 857-871.

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. Journal of intelligent information systems, 17(2-3), 107-145.

Hennig, C., & Liao, T. F. (2013). How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification? Journal of the Royal Statistical Society: Series C (Applied Statistics), 62(3), 309-369.

Hsu, C. C., & Chen, Y. C. (2007). Mining of mixed data with application to catalog marketing. Expert Systems with Applications, 32(1), 12-23.

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.

Dubes, R. C., & Jain, A. K. (1988). Algorithms for clustering data: Prentice-Hall, Inc.

Ji, J., Bai, T., Zhou, C., Ma, C., & Wang, Z. (2013). An improved k-prototypes clustering algorithm for mixed numeric and categorical data. Neurocomputing, 120, 590-596.

Kalaycı, Ş. (2010). SPSS Uygulamalı Çok Değişkenli İstatistik Teknikleri. Ankara: Asil Yay. Dağıtım.

Kassambara, A. (2017). Practical guide to cluster analysis in R: Unsupervised machine learning, (Vol. 1): STHDA.

Rousseeuw, P. J., & Kaufman, L. (1990). Finding groups in data. Hoboken: Wiley Online Library.

Liu, Y., Li, Z., Xiong, H., Gao, X., & Wu, J. (2010). Understanding of internal clustering validation measures: 2010 IEEE International Conference on Data Mining (pp. 911-916). IEEE.

Maulik, U., & Bandyopadhyay, S. (2002). Performance evaluation of some clustering algorithms and validity indices. IEEE Transactions on pattern analysis and machine intelligence, 24(12), 1650-1654.

Morlini, I., & Zani, S. (2010). Comparing approaches for clustering mixed mode data: An application in marketing research. In Data Analysis and Classification (pp. 49-57). Springer, Berlin, Heidelberg.

R Development Core Team (2013). R: A language and environment for statistical computing. R Foundation For Statistical Computing, Vienna, http://www.R-project.org

Saitta, S., Raphael, B., & Smith, I. F. (2008). A comprehensive validity index for clustering. Intelligent Data Analysis, 12(6), 529-548.

Sokal, R. R., & Rohlf, F. J. (1962). The comparison of dendrograms by objective methods. *Taxon*, *11*(2), 33-40.

Starczewski, A. (2017). A new validity index for crisp clusters. Pattern Analysis and Applications, 20(3), 687-700.

Swenson, E. R., Bastian, N. D., & Nembhard, H. B. (2016). Data analytics in health promotion: Health market segmentation and classification of total joint replacement surgery patients. Expert Systems with Applications, 60, 118-129.

Tatlıdil, Hüseyin (1992). Uygulamalı Çok Değişkenli İstatistiksel Analiz. Ankara: Engin Yayınları.

Theodoridis, S., & Koutroubas, K. (1999). Feature generation II. Pattern recognition, 2, 269-320.

Tibshirani, R., & Walther, G. (2005). Cluster validation by prediction strength. Journal of Computational and Graphical Statistics, 14(3), 511-528.

Tomašev, N., & Radovanović, M. (2016). Clustering evaluation in high-dimensional data. In Unsupervised Learning Algorithms (pp. 71-107). Springer, Cham.

Wu, J., Xiong, H., & Chen, J. (2009, June). Adapting the right measures for k-means clustering. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 877-886). ACM.

Yu, W., Qiang, G., & Xiao-li, L. (2006, October). A kernel aggregate clustering approach for mixed data set and its application in customer segmentation. In 2006 International Conference on Management Science and Engineering (pp. 121-124). IEEE.

ZAKI, Mohammed J. ve Wagner MEIRA JR. (2014), Data mining and analysis: fundamental concepts and algorithms: Cambridge University Press.

EK 1: Kamila Algoritması Kullanılarak Yapılan Analizde Kullanılan Kodlar

```
veri <- read.csv("binarymat.csv",header=FALSE,sep=";")

conInd <- c(9,10)
```

```
conVars <- veri[,conInd]

conVars <- data.frame(scale(conVars))

catVarsFac <- veri[,c(1,2,3,4,5,6,7,8)]

catVarsFac[] <- lapply(catVarsFac, factor)

catVarsDum <- dummyCodeFactorDf(catVarsFac)

kamRes <- kamila(conVars, catVarsFac, numClust=2:10, numInit=10,calcNumClust =
"ps",numPredStrCvRun = 10, predStrThresh = 0.5)

summary(kamRes)
```