Müşterilerin GSP Analizi Kullanarak Kümelenmesi
Customer Clustering using RFM Analysis

Muhammet Pakyürek[1], Mehmet Selman Sezgin[1], Sedat Kestepe[1], Büşra Bora[1], Remzi Düzağaç[1], Olcay Taner Yıldız[2]
[1]Etstur Veri Bilimi ve Analitiği Bölümü
{muhammet.pakyurek, selman.sezgin, sedat.kestepe, busra.bora, remzi.duzagac}@etstur.com
[2]Işık Üniversitesi Bilgisayar Mühendisliği Bölümü olcay.yildiz@isikun.edu.tr

***Özetçe* —Bu çalışma ile mevcut misafir ve rezervasyon verisi kullanılarak doğal öbeklenmeleri tespit ederek misafir davranışları tespit ettik. Ayrıca verilen hizmetleri ve satış stratejilerini bu davranışlara göre özelleştirdik. *K*-ortalama ile kişileri öbekledikten sonra bu mevcut öbeklenmeleri sağlayan temel karakteristikler karar ağacı yaklaşımı ile çıkartılmıştır. Bu karakteristiklerin kişinin ürün alma kanalı, belirli ürün tercihleri, rezervasyon süresi, sezonsal tercihi vb. olduğu tespit edilmiştir. Bu karakteristiklerin her öbeklenmede ciddi değişiklikler göstermiş olması çözümün genel olarak doğru olduğunun ve bu karakteristiklerin başarılı bir şekilde seçildiğini göstermektedir. Bu çalışma, grup karakteristiklerine uygun kampanyalar ve ürün paketleri oluşturulmasında önemli bir rol oynamaktadır.**

***Anahtar Kelimeler*—Kümeleme, GSP Analizi, Müşteri Segmentasyonu, K-ortalama Algoritması**

***Abstract*—In this study, customers' behaviors are determined by detecting natural clusterings using existing reservation and customer data. We also customize their services and sales strategies according to these behaviors. The basic characteristics that provide these existing heuristics have been extracted by the decision tree approach after the K-means is implemented. It is determined that these characteristics are customer's product acquisition channel, specific product preferences, reservation periods, seasonal preference, etc. The fact that these characteristics show significant changes in each clusters indicates that the solution is generally successful and that these characteristics are successfully selected. This work plays an important role in creating campaigns and product packages appropriate for these groups' characteristics.**

***Keywords*—Clustering, RFM Analysis, Customer Segmentation, K-means algorithm**

I. Giriş

K-ortalama algoritması veri madenciliği dünyasında yaygın kullanılan kümeleme algoritmalarından birisidir. Kümeleme algoritmaları ile Sınıflandırma algoritmaları arasında bir takım farklılıklar bulunmaktadır. Sınıflandırma algoritmaları bir etikete göre sınıflandırma yapmaktadır. Kümeleme algoritmaları ise otomatik olarak verileri daha küçük kümelere ya da alt kümelere ayırmaya yarayan denetimsiz öğrenme algoritmalardır.

Bu bildiride, müşterilerin mevcut MİY(Müşteri İlişkileri Yönetimi) verisinden çıkartılan GSP(Güncellik, Sıklık, Parasallık) özellikleri *K*-ortalama algoritması kullanılarak oluşan doğal kümelenmelerin bulunması, kişilerin kategorilendirilmesi ve her kategorinin karakteristiklerinin çıkartılıp yorumlanması amaçlanmıştır.

Bu bildiride, oluşan doğal kümelenmeler GSP analizindeki hedef kitlelere göre ayrıştırılarak her kümelenme kendi içinde de 2 ayrı gruba ayrılmıştır (Hedef kitle ve diğerleri gibi). Oluşan her bir ayrı öbek için karakteristikler başarılı bir şekilde elde edilmiştir.

Bölüm 2'de literatürde yapılan benzer akademik çalışmalardan, Bölüm 3'te kullanılan verilerden, Bölüm 4'te problem tanımı ve çalışmanın çözüm için yaklaşımının detaylarından, Bölüm 5'te ise elde edilen sonuçlardan bahsedilecektir.

II. Akademik Literatür

Geçmiş çalışmalarda *K*-ortalama algoritması ve GSP özelliklerine göre veri içindeki kümelerin bulunması farklı açılardan yorumlanmış ve farklı veri kümeleri üzerinde uygulanmıştır.

E. P. Xing ve diğerlerinin çalışması (2003), *K*-ortalama algoritması ile kümeleme yaparken hangi özelliklerin seçileceğinin kümeleme başarısına etkisini ortaya koyarak GSP özelliklerinin önemine ve belirleyiciliğine vurgu yapmıştır [1]. Merkez belirlemede tahmin edilebilirliğin önemini ortaya koyan B.Bahmani ve diğerleri (2012) *K*-ortalama algoritmasını *K*-ortalama++ ile birlikte kullanmıştır [2].

K. Wagstaff ve diğerleri de (2001) *K*-ortalama algoritmasını, verilere ilişkin sahip olunan ön bilgilerin kullanımına el verecek şekilde değiştirerek her ne kadar öğreticisiz olmaktan çıkarsalar da bu değişiklik sayesinde çok yüksek başarı oranları elde etmişlerdir [3]. Müşteri taleplerinde örüntü belirlemek için AÖM (Ardışık Örüntü Madenciliği) kullanan Y. L. Chen ve diğerleri (2009), GSP destekli çalışmalarında klasik çalışmalarına göre daha keskin örüntüler elde etmiştir [4].

C. H. Cheng ve diğerleri ise (2009) GSP özellikleri ile *K*-ortalama uygulamasını RS (Rough Set) teorisi ile desteklemişler. Çalışma, önerilen yöntemin karar ağaçları ve yapay sinir ağları kullanılarak yapılan kümelemeye göre yüksek başarı oranları elde ettiği sonucunu göstermiştir [5]. GSP'nin VTM (Vaka Tabanlı Muhakeme) ile birlikte ağ erişiminde şüpheli kullanıcıların tespiti için kullanıldığı H. K. Kim ve diğerlerinin çalışmasında (2010), GSP fikri yeniden yorumlanmıştır [6].

D. L. Olson ve diğerleri (2009) çalışmalarında GSP'nin zayıf olduğu bir çalışma örneği sergilemişler. GSP, logistic

978-1-5386-1501-0/18/$31.00

regression, karar ağaçları, yapay sinir ağları teknikleri kullanarak müşteri sınıflandırma çalışması yapan D. L. Olson ve diğerleri, GSP kullanarak elde edilen sonucun doğruluk oranının diğerlerinden düşük olduğunu keşfetmiştir [7]. J. A. ve diğerleri çalışmalarında (2007) GSP'yi CHAID (Chi-Square Automatic Interaction Detection) ve Logistic Regression'a karşı test etmişler. Çalışmada iki farklı veri kümesi kullanmışlar. GSP ile CHAID arasında yaptıkları kıyaslamada, veri kümesi üzerinde yapılan bir ön çalışmanın CHAID ile yapılacak analizi GSP'den üstün kıldığını, ön hazırlık uygulanmamış veri kümesiyle yapılan değerlendirmede ise GSP ile CHAID metodlarının benzer doğruluk oranlarında sonuç verdiğini gözlemlemişlerdir [8].

III. Veri tanitimi

Demografik bilgiler (isimsizlestirilmiş), otel tercihleri, rezervasyon sayıları, tarihleri ve rezervasyon kanalı (web, acente..) gibi veriler, muhtelif sistemlerde ve bu sistemler kaynak alınarak oluşturulmuş veri havuzlarında bulunmaktadır. Bu çalışmada, bu tür veriler kullanılmıştır. Ayrıca yaptığı rezervasyonların miktarı, sıklığı ve güncelliği göz önünde tutularak GSP analizi yapılmıştır. Ayrıca bu mevcut verilerden, analizlerde daha anlamlı ve isabetli sonuçlar elde edilebilmesi için istatistik ve matematiksel (ortalama satış, aylara göre satış dağılımları) kavramlar kullanılarak yeni öznitelikler de elde edilmiştir (Tablo I).

A. GSP Analizi

GSP analizi, müşterilerin şirket ile ilgili aktiviteleri baz alınarak yapılan bir analizdir. Bu analiz tekniğini kullanarak kişilerin şirket için değerleri belirlenir. GSP analizi için kullanılan özellikler şunlardır:

Güncellik : Yapılan en son alışverişten bu yana geçen ay sayısı

Sıklık : Bireysel bazda toplamda yaptığı alışveriş sayısı

Toplam harcanan para : Bir kişinin harcadığı toplam para (Bu analizde geçmiş her sene için %10 faiz uygulanmıştır. Böylece eski verilerin günümüzdeki olası değeri hesaplanarak paraların güncel halleri arasında daha nesnel ve anlamlı bir karşılaştırma elde edilmiştir.)

Bu özniteliklerin dağılımına bakılarak harcanan para 4 kademede, sıklık 4 kademede ve güncellik ise 3 kademede (aktif, potansiyel ve kayıp) değerlendirilmiştir. Bu kademeler 1, 2, 3, 4 gibi en düşükten en yükseğe doğru sıralanmıştır. Bu çalışmadaki deney sonuçları aktif kişiler için elde edilmiştir. Aktif kişilerde hedef kitle, sıklıkta 4. kademe ve harcanan parada 2., 3., 4. kademeler; sıklıkta 3. kademe ve harcanan parada 3., 4. kademeler; sıklıkta 2. kademe ve harcanan parada 4. kademe olarak seçilmiştir.

Yapılan kümeleme çalışmasında değerli kitlelerin karakteristiklerinin daha iyi anlaşılabilmesi için GSP analizine göre müşteride odaklanması gereken hedef kitle belirlenmiştir. Bu hedef kitle GSP analizinde değerlilik kriterleri hesaplanırken tüm kişilerin %20'sinin bu alan içinde kalması esas alınarak aralıklar belirlenmiştir. Her öbeklenme grubu için hedeflenmesi gereken kitleler belirlenip öbeklenmeler kendi içinde de hedef kitle ve diğerleri şeklinde ikiye ayrılmıştır. Her iki durumun karakteristikleri ayrı ayrı oluşturulmuştur.

TABLO I: Veri Öznitelikleri ve Açiklamalari

Kullanılan Öznitelikler	İçerdiği Veritabanı Alanları
Demografik bilgiler (İsimsizlestirilmiş)	Yaşanan şehir, ilçe, semt, cinsiyet, yaşı, meslekleri
Toplam ürün satın alma sayısı	kişinin toplam rezervasyon sayısı
Tek seferde harcanan ortalama para	Yapılan rezervasyonların ortalama tutarı
Tek seferde harcanan en alt ve en üst tutarlar	Kişinin yaptığı en alt ve en üst rezervasyon tutarları
Grup şeklinde alınan bir ürün ise ürünü alan müşteri grubunu tanımlayan oranlar	Rezervasyonlara katılan ortalama yetişkin ve çocuk sayısı
Satın alınan ürün gruplarının sayıları ve satın alınma oranları	Değişik ürün grubu satışlarının sayısı ve toplam satılan ürünlere oranı
Kişinin aktif alışveriş yaptığı süre	Kişinin aktif olarak alım yaptığı sürenin yıl cinsinden karşılığı
Aylara göre satın alma oranı	Dönemsel davranışları belirlemek için kişilerin her ay hangi oranda alışveriş yaptığı bilgisi
Alışverişler arasında geçen ortalama süre	İki rezervasyon satışı arasında geçen ortalama süre
Ürünlerin detayları ile ilgili oranlar ve en üst, en alt değerler	Bir kişinin rezervasyonlarının ortalama konaklama süresi, konaklama süresinde kişi başı harcanan ortalama para ve bunlarla ilgili minimum, maksimum değerler
Kampanyaların tercihi ile ilgili oranlar	Kişilerin kampanyaları kullanma oranı ve kişilerin rezervasyon aldıktan ortalamada ne kadar süre sonra tatile gittiği bilgisi
Satış kanallarının kişi tarafından tercih edilme oranları	İnternet sitesi ve turizm acenteleri gibi satın alma kanallarından alınan rezervasyon sayısının toplam rezervasyon sayısına oranı

IV. Problem ve Çözüm

Veri Tanıtımı'nda anlatılan verilerin değişik teknikler kullanılarak incelenmesi esnasında doğal kümelenmelerin mevcut olduğu gözlemlenmiştir. Bu çalışmanın amacı, misafir verisinde oluşan doğal kümelenmelerin, veri içerisinde bulunan önemli özellikler ve GSP skorlama ile K-Ortalama kullanılarak tespit edilmesidir. Ayrıca bu çalışma, kümelerdeki genel karakteristiklerin karar ağacı kullanılarak yapılan gözlemlerle örtüştüğünü ortaya koymaktadır.

Uygulanan çözümde sırasıyla önce veri ön işleme adımları sonrasında, şifrelenmiş veri üzerinden K-ortalama ile öbekleme ve sonrasında ham veri üzerinde karakteristiklerin çıkarılması için Karar ağacı uygulanmıştır. Bu karar ağacının neticesinde her bir öbeklenme için oluşan a farklı ayrışmaların olduğu gözlemlenmiştir.

A. Kümeleme için Ön İşlemler

1) Verinin Karar Ağacı ile şifrelenmesi: Verinin ham haliyle yapılan kümeleme sonuçlarının Tablo II'de görüldüğü üzere düşük olduğu gözlenmiş olup; başarımın arttırılması amacıyla veri karar ağacı ile tekrardan şifrelenmiştir.

2) Boş verilerin silinmesi: Boş hücre sayısı yüksek olan veriler, yapılan analizlerin yanlış noktalara yakınsamasına, yanlış çıkarımların yapılmasına neden olmaktadır. Bu çalışmada, boş hücre sayısı mevcut kolon sayılarının %10' ununa tekabül eden satırlar elenmiştir.

3) Uç değerli verilerin çıkarılması: Uç değerli veriler genelde *K*-ortalama gibi mevcut verinin metoidlere olan uzaklığını esas alan kümeleme algoritmalarında kümelemenin yanlış merkezlerde oluşmasına neden olmaktadır. Uç değerli verilerin elenmesi için çeyreklik hesaplamaları kullanılmıştır.

4) Silinmeyen boş verilerin doldurulması: Literatürde işbirlikçi filtre (collabrative filter), makine öğrenmesi (machine learning) gibi yöntemler kullanılarak boş verilerin doldurulduğu çalışmalar bulunmaktadır. Bu çalışmada, sürekli veriler için ortalama değeri, kategorik veriler için en çok kullanılan veri kullanılmıştır.

5) Verinin normalleştirilmesi: Veri normalleştirme hem hesaplamaları kolaylaştıran hem de algoritmaların daha yüksek hız ve verimde sonuca yakınsamasını sağlayan bir işlemdir. Bu çalışmada minimum - maksimum normalleştirmesi kullanılmıştır.

B. K-ortalama

Algoritma istatistiksel olarak benzer nitelikteki kayıtları aynı gruba sokar. Bir elemanın yalnızca bir kümeye ait olmasına izin verilir. Algoritmanın isminde yer alan *k* harfi, küme sayısını belirtir. Algoritma, örnekleri kümelerken bir taraftan da hata hesaplamada yaygın olarak kullanılan karesel hata'yı en aza indirgeyecek *k* küme sayısını arar. Verilen *n* sayıdaki veri kümesi *k* tane kümeye bu hata fonksiyonunun sonucunu en aza indirgeyecek şekilde yerleştirilir. Burada küme benzerliği, kümedeki değerlerin ortalamaya yakınlıkları ile ölçülür. Kümenin merkezinde yer alan değer kümenin temsilci değeridir ve medoid olarak adlandırılır.

Algoritma temel olarak 4 aşamadan oluşur:

1) Küme merkezleri belirlenir.
2) Merkez dışındaki örnekler küme merkezlerine olan mesafelerine göre kümelere atanır.
3) Yapılan atamaya göre yeni merkezler belirlenir. (veya eski merkezler yeni merkeze kaydırılır)
4) Kümeler arası eleman değişimi bitene kadar veya belirli bir eşik değerinin altında değişim yakalanana kadar 2. ve 3. adımlar tekrarlanır.

Algoritmanın avantajlı olduğu noktalar, uygulaması basit, hızlı sonuç alınabilir olması ve öbeklenme sonuçlarının yorumlanmasının basit olmasıdır.

Eksik olduğu yönler ise **küme sayısının başlangıçtan belirleniyor olması** ve **küme merkezlerinin başlangıç değerlerinin yanlış bir şekilde seçilme ihtimalinin yüksek olmasıdır**. Diğer bir sorun da uç değerli verilerin bu kümelenmenin yanlış noktalarda oluşmasına neden olmasıdır.

Doğru küme sayısının bulunabilmesi için öncelikle kümelenmelerin doğru şekilde oluştuğunu ölçebilen bir metriğe ihtiyaç vardır. Bu konuda, özellikle kesin referans bir veri varsa kullanılabilecek çeşitli metrikler bulunmaktadır: Türdeşlik (homogeneity), eksiksizlik ve V-ölçüsü (completeness and V-measure), Fowlkes-Mallows dereceleri (Fowlkes-Mallows scores), düzenlenmiş benzerlik indeksi (Adjusted Rand index). Fakat problemin doğası gereği doğal öbeklenmeler çıkarılmak istendiğinden kesin referans veri bulunmamaktadır. Bu durumda öne çıkan metrikler Calinski-Harabasz indeks ve Silüet Katsayısı'dır (Silhoutte Coefficient). Bu çalışmada silüet katsayısı kullanılmıştır.

Silüet Katsayısı
$$s_i = \frac{b_i - a_i}{\max(b_i, a_i)} \quad (1)$$

a_i: Mevcut kümede olan nesnelere ortalama uzaklık

b_i: Mevcut kümeye en yakın kümedeki nesnelere olan ortalama uzaklık

görüldüğü üzere s_i -1 ile +1 arasında değişmektedir. İdeal durumda beklenen a_i'nin 0'a yakın çıkması ve bu durumda s_i'nin ise 1'e yaklaşmasıdır. Bu yüzden de s_i, +1'e ne kadar yakın ise kümelenmenin o kadar iyi olduğu düşünülebilir.

Küme merkezleri için başlangıç değeri olarak verilen değerlerin optimize edilebilmesi için *K*-ortalama++ çalışması [9] kullanılmıştır. Bu algoritma *K*-ortalama küme merkezlerini ilklendirirken:

1) Tekdüze dağılım olarak rastgele bir c_1 küme merkezi olarak seçilir.
2) Sonraki küme merkezi, halihazırda seçili olan küme merkezlerine en uzak olacak şekilde seçilir.
3) 2. adım *K* tane küme merkezi oluncaya kadar uygulamaya devam edilir.

Aşırı uçlardaki verilerin ayıklanması için çeyreklik hesaplaması kullanılmıştır. Buradaki kriter:

$$\Delta = 1.5 * (Q_3 - Q_1) \quad (2)$$

Q_1, Q_3: Veri, bir öznitelik için ortalama değerden ikiye bölündüğünde, her bölümün ortalama değerleridir.

Çeyreklikler arası ağırlıklı farkı temsil eden Δ değeri ile uç noktalar tespit edilebilmektedir. Uç değerli noktalar bu durumda şu şekilde hesaplanır.

$$ucdeger = veri[i] < Q_1 - \Delta \quad \cup \quad veri[i] > Q_3 + \Delta \quad (3)$$

Öbeklenmelerin ham veri üzerindeki karakteristiklerinin Karar ağacı kullanarak nasıl çıkarıldığı Deney ve Sonuçlar kısmında açıklanmıştır.

V. Deney ve Sonuçlar

Yapılan deneylerde farklı *K* değerleri ve farklı kümeleme algoritmaları için silüet katsayıları hesaplanmış olup ilgili algoritma ve *K* değerinin karşısına değeri konulmuştur.

TABLO II: Ham Veri İle Kümeleme Performans Kıyaslaması

K	GKM	*K*-ortalama	*K*	GKM	*K*-ortalama
1	-0,132	0,041	**6**	-0,012	0,002
2	0,057	0,209	**7**	0,003	0,003
3	0,096	0,036	**8**	0,019	-0,006
4	0,095	0,015	**9**	-0,014	-0,006
5	0,006	0,006	**10**	0,001	-0,007

Öncelikle öbeklenmenin verinin doğal haliyle nasıl olduğunu gözlemlemek için K-ortalama ve GKM (Gaussian Karışımı Modeli) kullanılmış olup sonucu Tablo II'de görüldüğü gibi çıkmıştır.

Deneylerde daha yüksek sayıda K değerleri için sonuçlar elde edilmiş olup; sonuçların K=10'a kadar olan değerlerden çok farklı olmadığı gözlenmiştir. Fakat deneylerin küme sayısı 10 kadar olan kısmı yeterli miktarda tekrar edilmiş olup tüm deneylerin ortalaması burada verilmiştir. Bundan sonraki verilen tüm veriler en az 10 defa tekrar edilmiş olup ortalama değerleri bu çalışmada paylaşılmıştır.

Tablo II'de görüldüğü üzere verinin ham haliyle yapılmış olan öbeklemenin performansı kötü çıkmıştır. Bunun temel nedeni, sürekli verilerin öbeklenme sınırlarının belirli ve geniş bir şekilde tanımlanmasını olumsuz etkilemesidir. Ayrıca bozuk ve uç değerli veriler de öbeklenmenin bozulmasına neden olmaktadır. Bu yüzden ham veriyi rastgele karar ağaçlarıyla şifrelenmiş, ardından tekil değer ayrışımı tekniği uygulanmıştır. Böylece verideki süreklilikten kaynaklanan belirsizlik ve tamamen olmasa da bozuk ve uç değerli verilerin etkisi kaldırılmıştır. Ham veri 26 öznitelikten (kolon) 43 özniteliğe çıkmış olup, silüet katsayısı metriğine göre kümelenme başarıları Tablo III'de verilmiştir.

TABLO III: KÜMELEME PERFORMANS KARŞILAŞTIRMASI (KARAR AĞACI ŞİFRELEMESİ SONRASI)

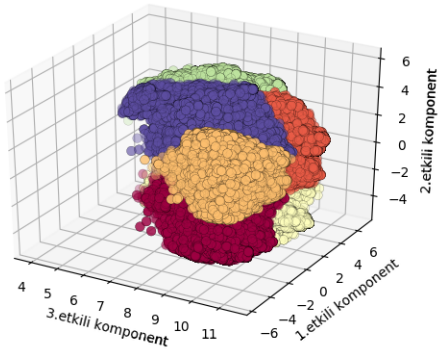
K	GKM	K-ortalama	K	GKM	K-ortalama
2	0,162	0,159	**15**	0,189	0,239
3	0,190	0,190	**16**	0,198	0,255
4	0,194	0,219	**17**	0,190	0,259
5	0,157	0,201	**18**	0,215	0,263
6	0,203	0,217	**19**	0,203	0,266
7	0,160	0,215	**20**	0,216	0,265
8	0,175	0,233	**21**	0,189	0,264
9	0,166	0,237	**22**	0,220	0,249
10	0,183	0,238	**23**	0,228	0,271
11	0,179	0,241	**24**	0,216	0,256
12	0,203	0,248	**30**	0,227	0,277
13	0,203	0,257	**40**	0,225	0,281
14	0,222	0,256	**50**	0,215	0,289

Tablo III'de görüldüğü üzere karar ağacı şifrelemesi kullandıktan sonra kümelenmedeki başarı artmıştır. Ayrıca K-ortalamanın, GKM'den daha iyi performans gösterdiği gözlemlenmiştir. Şifrelenmiş veride, bozuk ve uç değerli verilerin etkisini daha da azaltmak için bir sonraki deneyde tüm ön işlemler uygulanmış ve K-ortalama algoritması kullanılmıştır. Bu sefer deneyde küme sayısı 5'ten 30'a 5'er 5'er artırılmıştır. En iyi değerlerin 5-10 arasında olduğu tespit edilmiştir. Tam olarak en iyi performansın olduğu küme sayısını bulabilmek için 5 ile 10 arasındaki tüm değerler ile tekrar öbeklendirilmiştir.

TABLO IV: ŞIFRELENMIŞ VE İŞLENMİŞ VERİDEKİ SİLÜET DEĞERLERİ

K	K-ortalama	K	K-ortalama
5	0,363	**5**	0,362
10	0,335	**6**	0,352
15	0,235	**7**	0,366
20	0,208	**8**	0,351
25	0,198	**9**	0,346
30	0,186	**10**	0,335

Tablo IV'de görüldüğü üzere en uygun kümeleme sayısının 7 olduğu tespit edilmiştir. Bu kümelenme karakteristiklerinin ham veri üzerinde neler olduğunu görmek için karar ağacı uygulanmştır. Karar ağacı uygulanırken 'gini indeks'teki azalma miktarı belirlenen bir limitin altına düştüğünde ayrıştırma sonlandırılmıştır. Böylece daha kapsamlı ve tutarlı karakteristiklerin oluşturulması sağlanmıştır. Karar ağacındaki ayrışmalarda kullanılan gerçek özniteliklerde çarpıcı değerlere sahip olanlar kullanılarak kümeler isimlendirilmiştir. Örneğin tatil köyü alan erken rezervasyoncular, son dakikacı - fırsatçı gibi isimlendirmeler kullanılmıştır. Veriler, Şekil 1'de görüldüğü üzere öbeklenmiştir.



Şekil 1: Müşteri Gruplarının Öbeklenmesi

VI. TARTIŞMA

Bundan sonraki çalışmalarda verilerin geliştirilmesi ve daha karmaşık öbeklenmeleri yapan rastgele karar ormanı, derin öğrenme algoritmaları kullanarak daha kapsamlı ve ayrıntılı öbeklenmelerin keşfedilmesi planlanmaktadır.

KAYNAKLAR

[1] E.Xing, A.Ng, M.Jordan, and S.Russell. "Distance metric learning, with application to clustering with side-information". In Proceedings of the Conference on Advances (NIPS)(2003)

[2] B.Bahmani, B.Moseley, A.Vattani, R.Kumar, S.Vassilvitskii "Scalable K-Means++", VLDB Endow 5(7), 622–633 (2012)

[3] K. Wagstaff, C. Cardie, S. Rogers, S. Schroedl "Constrained K-means Clustering with Background Knowledge", (2001)

[4] Y.L. Chen, M.H. Kuo, S.Y. Wu, K. Tang, (2009). "Discovering recency, frequency, and monetary (RFM) sequential patterns from customers purchasing data, Electronic Commerce" Research and Applications, (2009)

[5] C.H. Cheng, Y.S. Chen, "Classifying the segmentation of customer value via RFM model and RS theory", Expert Systems with Applications, (2009)

[6] H. K. Kim, K. H. Im, S. C. Park, "DSS for computer security incident response applying CBR and collaborative response", Expert Systems with Applications, 2010

[7] D.L. Olson, Q. Cao, C. Gu, D. Lee, "Comparison of customer response models", Service Business, 2009

[8] J. A. McCarty, M. Hastak, "Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression", Journal of Business Research, (2007)

[9] D. Arthur, S. Vassilvitskii, "k-Means++: the advantages of careful seeding", in Proc. of SODA, (2007)

[10] Y-L. Chen, M-H. Kuo, S-Y. Wu, K. Tang, "Discovering recency, frequency, and monetary (RFM) sequential patterns from customers purchasing data", Electronic Commerce Research and Applications, (2009)