T.C.
DOKUZ EYLÜL ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
EKONOMETRİ ANABİLİM DALI
EKONOMETRİ PROGRAMI
YÜKSEK LİSANS TEZİ

PARÇACIK SÜRÜ OPTİMİZASYONU VE VERİ MADENCİLİĞİ YÖNTEMLERİYLE KÜMELEME ANALİZİ: İZMİR'DE KURULU BİR FİRMANIN MÜŞTERİLERİNİN KÜMELENMESİ UYGULAMASI

Ayşegül CENGER

Danışman
Doç. Dr. Mehmet AKSARAYLI

İZMİR - 2018

TEZ ONAY SAYFASI

YEMİN METNİ

Yüksek Lisans Tezi olarak sunduğum “Parçacık Sürü Optimizasyonu Ve Veri Madenciliği Yöntemleriyle Kümeleme Analizi: İzmir'de Kurulu Bir Firmanın Müşterilerinin Kümelenmesi Uygulaması” adlı çalışmanın, tarafımdan, akademik kurallara ve etik değerlere uygun olarak yazıldığını ve yararlandığım eserlerin kaynakçada gösterilenlerden oluştuğunu, bunlara atıf yapılarak yararlanılmış olduğunu belirtir ve bunu onurumla doğrularım.

Tarih
…/…/2018
Ayşegül CENGER
İmza

ÖZET

Yüksek Lisans Tezi

Parçacık Sürü Optimizasyonu Ve Veri Madenciliği Yöntemleriyle Kümeleme Analizi: İzmir'de Kurulu Bir Firmanın Müşterilerinin Kümelenmesi Uygulaması

Ayşegül CENGER

Dokuz Eylül Üniversitesi
Sosyal Bilimler Enstitüsü
Ekonometri Anabilim Dalı
Ekonometri Programı

Günümüzde firmaların kurumsallaşması için birtakım yöntemlerle müşterilerini iyi tanıması, müşteri özelliklerine göre hareket etmesi ve çalışanlarına bu doğrultuda görevler vererek, alanlarında uzmanlaşmalarını sağlaması gerekmektedir. Ancak bunu sağlamak müşteri sayısı çok olan şirketlerde geleneksel yöntemlerle kolay olmamaktadır. Yapılan tezde, müşteriler hem Veri Madenciliği Yöntemlerinden K-ortalamalar ve İki Aşamalı Algoritma Yöntemleri ile hem de modern sezgisel teknik olan Parçacık Sürü Optimizasyonu Yöntemi ile kümelenmiştir. Bu yöntemlerle ilgili kavramlar açıklanmış ve kapsamlı bir literatür araştırması yapılmıştır. K-ortalamalar ve İki Aşamalı Algoritma Yöntemleri için SPSS Clementine 10. 1 programı tercih edilmiştir. Parçacık Sürü Optimizasyonu algoritması ise kümelenmeye uyarlanarak MATLAB programında bu amaçla kullanılmıştır. Algoritmada amaç fonksiyonu olarak Dunn İndeksi kullanılmış ve küme sonuçları elde edilmiştir.

Anahtar Kelimeler: Veri Madenciliği, Kümeleme Yöntemleri, Parçacık Sürü Optimizayonu ile Kümeleme, Müşteri Kümeleme

ABSTRACT

Master's Thesis

Clustering Analysis with Particle Swarm Optimization and Data Mining Methods: Application of a Cluster of Customers of a Company Established in İzmir

Ayşegül CENGER

Dokuz Eylül University

Graduate School of Social Sciences

Department of Econometrics

Econometrics Program

Nowadays, for the institutionalization of companies, it is necessary to provide specialization in their fields by giving them a good understanding of the customers in a number of methods, acting according to customer characteristics and giving their employees the tasks in this direction. However, this is not easy with traditional methods in companies with a large number of customers. In this thesis, the customers were clustered with both K-means and Two-step Algorithm Methods from Data Mining Methods as well as the modern heuristic technique, Particle Swarm Optimization Method. The concepts related to these methods have been explained and a comprehensive literature search has been done. SPSS Clementine 10. 1 has been preferred for K-means and Two-Stage Algorithm Methods. The Particle Swarm Optimization Algorithm has been used for this purpose in the MATLAB program, adapted to the cluster. Dunn Index is used as the objective function in the algorithm and cluster results are obtained.

Keywords: Data Mining, Clustering Methods, Clustering with Particle Swarm Optimization, Customer Clustering.

PARÇACIK SÜRÜ OPTİMİZASYONU VE VERİ MADENCİLİĞİ YÖNTEMLERİYLE KÜMELEME ANALİZİ: İZMİR'DE KURULU BİR FİRMANIN MÜŞTERİLERİNİN KÜMELENMESİ UYGULAMASI

İÇİNDEKİLER

ÜÇÜNCÜ BÖLÜM
VERİ MADENCİLİĞİ

DÖRDÜNCÜ BÖLÜM
KÜMELEME ANALİZİNDE K-ORTALAMALAR VE İKİ AŞAMALI ALGORİTMA YÖNTEMLERİ

BEŞİNCİ BÖLÜM
UYGULAMA

KISALTMALAR

T.C.	Türkiye Cumhuriyeti
SS	Standart Sapma
SPSS	Statistical Packages for the Social Sciences
PSO	Parçacık Sürü Optimizasyonu
ÖS	Önem Seviyesi
ORT	Ortalama
MATLAB	Matrix Laboratory

TABLOLAR LİSTESİ

ŞEKİLLER LİSTESİ

GİRİŞ

Günümüzde oldukça gelişen teknoloji sayesinde, geleneksel yollarla çözülemeyen birçok problem çözülebilir hale gelmektedir. Veri miktarında artışın olduğu durumlarda ise veriler işlenmediği sürece değersiz görünen veri yığınları ortaya çıkmaktadır. Bu durumda veri madenciliği devreye girmektedir. Veri madenciliği sayesinde veri tabanı sistemleri içerisinde gizli kalmış bilgilerin çekilmesi sağlanmaktadır. Bu da her alanda avantaj haline gelmektedir.

Veri madenciliği tekniklerinden biri olan kümeleme tekniği ise birçok alanda kullanılması oldukça faydalı bir tekniktir. Bu teknik, elde olan bir grup değişkeni önceden belirlenmiş özelliklere göre kümelere ayırma işlemidir. Birbirine benzer yapıdaki değişkenler bir araya gelerek kümeleme işlemi yapılır ve bu işlem veri setini daha iyi tanımlamaya yarar. Böylece yapılacak analizlere ortam hazırlanmış olur.

Bu çalışmada bir firmanın müşterilerini kümelemek için veri madenciliği yöntemlerinden olan İki Aşamalı Kümeleme ve K-ortalamalar Kümeleme yöntemlerinden farklı olarak sezgisel optimizasyon algoritması olan Parçacık Sürü Optimizasyonu kümeleme algoritması da kullanılmıştır. Burada amaç, kümelere ayrılan müşterilerin özelliklerini tespit ederek, satış politikalarını düzenlemek ve firmanın kurumsallaşmasını sağlamaktır.

Birinci bölümde çalışmanın amacı ve kapsamı belirtilmiş ve yapılan literatür araştırması sunulmuştur.

İkinci bölümde kümeleme algoritması olarak kullanılan Parçacık Sürü Optimizasyonunun genel tanımı yapılmış, algoritmanın özelliklerinden bahsedilmiş ve ayrıca probleme uyarlanan kümeleme algoritması hakkında bilgiler verilmiştir.

Üçüncü bölümde veri madenciliği tanımı, süreci, uygulama alanları ve teknikleri hakkında genel bilgiler verilmiştir.

Dördüncü bölümde ise veri madenciliği içerisinde yer alan kümeleme tekniklerinden uygulamada kullanılacak olan İki Aşamalı Kümeleme Tekniği ve k-ortalamalar tekniği ele alınarak ayrıntılı bilgiler verilmiştir.

Beşinci bölüm uygulamanın yapıldığı bölümdür. Bu bölümde 394 müşteri için belirlenen 11 değişken tanımlanmış ve değişkenlerin kombinasyonları alınarak oluşturulan 7 adet modelin özellikleri anlatılmıştır. Kümeleme sonuçları SPSS

Clementine 10. 1 ve MATLAB 2016a programları kullanılarak bulunmuş ve çıktılar yorumlanarak firma müşterilerinin kaç kümeye ayrıldığı, bu kümeler için hangi değişkenlerin daha önemli olduğu gibi anlamlı sonuçlar elde edilmiş ve bu sayede firmanın farklı kümelerdeki müşterilerine özgü satış politikalarının düzenlenmesi için öneriler sunulmuştur.

BİRİNCİ BÖLÜM
KAPSAM VE LİTERATÜR

1.1. AMAÇ

Yapılan çalışmada amaç, İzmir'de faaliyet gösteren bir otomasyon firmasının amaçları dâhilinde müşteri ilişkilerini daha iyi yönetebilmek ve pazar stratejilerini belirlerken müşteri istek ve ihtiyaçlarına yönelik davranabilmek için müşterilerin benzerliklerini dikkate alarak en uygun kümeleme sayısını bulmak ve kümeleri elde etmektir. Böylelikle, firma için gerekli politikalar oluşturulabilecek ve eldeki müşterilerin kaybı önlenerek kaybedilen müşterilerin ise geri kazanımı sağlanabilecektir. Bu sayede firmanın kurumsallaşma yolunda ilerleme kaydedebilmesi için imkan sağlanmış olacaktır.

Müşteri sayısının çok olduğu problemlerde bu çözüm zor bir problem haline gelmektedir. Böylesi problemlerin iş hayatında çok sık karşılaşılan problemlerden olması ve çözüm yöntemi olarak yöneylem araştırması yöntemlerinin kullanılıyor olması araştırmayı önemli kılmaktadır.

1.2. KAPSAM

Çalışmada örneklem, İzmir'de faaliyet gösteren bir otomasyon firması müşterilerinden oluşmaktadır. Veri setinin bir kısmı gerçekleşen ticari işlemler sonucu firmadan alınmış, bir kısmı ise uzman görüşü ile elde edilmiştir. Toplamda 394 müşteri vardır ve 11farklı değişken kullanılmıştır. Bu değişkenler 7 farklı model kurularak kümeleme işlemi hem Veri Madenciliği Yöntemleri ile hem de Parçacık Sürü Optimizasyonu Yöntemi ile gerçekleştirilmiştir.

1.3. LİTERATÜR ARAŞTIRMASI

Özmen, Delice ve Aydoğan (2018) çalışmalarında, Türkiye'nin en büyük 100 telekomünikasyon firmasından birine müşteri segmentasyonu uygulamıştır. Firmanın veri ambarından fatura bilgileri, çağrı detayları gibi birçok bilgi elde edilmiştir. Büyük veri boyutundan dolayı manuel analizin yapılması mümkün olmadığı için, veri madenciliği tekniğinden yararlanılmıştır. Parçacık Sürü Optimizasyonu tabanlı kümeleme tekniği ve Davies-Bouldin uygunluk fonksiyonu ile müşteri bölümleri meydana gelmiştir. Parçacık Sürü Optimizasyonu tabanlı kümeleme tekniği ve

Davies-Bouldin uygunluk fonksiyonunun müşteri bölümlemesinde etkili bir şekilde kullanılabildiği gösterilmiştir.

Alswaitti ve arkadaşları (2018) çalışmalarında, parçacık sürülere dayanan veri kümelemesi için yeni bir yaklaşım sunmuştur. Bu yaklaşımda, çalıştırma ve arama süreçleri arasındaki denge, çok boyutlu çekimsel öğrenme katsayılarının tahmini ve erken yakınsamayı gösteren yeni bant genişliği tahmin yöntemiyle ilişkili çekirdek yoğunluğu tahmin tekniğinin bir kombinasyonu kullanılarak değerlendirilmiştir. Önerilen algoritma, Dunn indeksi tarafından kümeleme kompaktlığını temsil eden 11 benchmark verisi kullanılarak diğer algoritmalar ile karşılaştırılmıştır. Önerilen algoritmanın önemi, uzman sistemler ve makine öğrenimi alanında önemli bir teknik olarak kümelemeyi ilerletmek için PSO tabanlı kümeleme algoritmalarının sınırlamalarını ele almasıdır. Bu bağlamda, önerilen algoritma, yüksek boyutlu veri kümelerine uygulandığında en son teknoloji ürünü algoritmalara kıyasla daha iyi sonuçlar elde etmektedir. Bu bulgu, özellik uzayının tüm boyutlarında parçacık hareketlerini dikkate alan çok boyutlu öğrenme katsayılarının tahmin edilmesinin önemini göstermektedir.

Doğan ve arkadaşlarının (2017) yaptıkları çalışmada, mikroskobik görüntülerin kümelenmesi için yeni bir yaklaşım olarak entropi tabanlı Parçacık Sürü Optimizasyonu uygulamış ve entropi ölçülerinin sürece etkisi karşılaştırılmıştır. Gri forma dönüştürülmüş görüntülerin uygun tek bir eşik unsurunun tespiti işlemi uygulanmıştır. Bu işlem için Tsallis entropi, tek seviyeli Kapur entropi, ve Otsu sınıflar arası varyans entropi ölçüsü tabanlı parçacık sürü optimizasyonu kullanılmış ve bulunan en iyi eşik değeri ile görüntü ikili görüntüye dönüştürülmüştür. Elde edilen kümeleme sonuçları, uzman kişilerin belirledikleri kümeleme sonuçları ile karşılaştırılmıştır. Yapılan çalışmada farklı entropi ölçülerinin Parçacık Sürü Optimizasyonu ile mikroskobik görüntülerin kümelenmesindeki etkisi gözlemlenmiştir. Sayısal sonuçlar Parçacık Sürü Optimizasyonu ile mikroskobik görüntüleri kümelemenin daha başarılı olduğunu göstermiştir.

Aksaraylı ve Pala (2017) çalışmalarında İki Aşamalı Kümeleme Algoritmasını kullanarak Uzaktan Eğitim alanında öğrencilerin tercih nedenleri, başarı durumları ve demografik değişkenleri ile kümeleme analizi gerçekleştirmişlerdir.

Izakian ve arkadaşları (2016) çalışmalarında, parçacık sürü optimizasyonu (PSO) yaklaşımı kullanılarak yörünge verilerinin kümelenmesi için otomatik bir yöntem önermiş ve yörünge verileri için en yaygın kullanılan mesafe ölçümlerinden biri olan Dinamik Zaman Eğrisi (DTW) mesafesini göz önüne almışlardır. Önerilen teknik içerisinde, kümelenme işlemi sırasında optimal küme merkezlerinin yanı sıra optimum sayıda küme merkezi bulabilmektedir. Arama alanının boyutsallığını azaltmak ve tavsiye edilen yöntemin performansını (belirli bir performans indeksi açısından) geliştirmek için, küme merkezlerinin Ayrık Kosinüs Dönüşümü (DCT) gösterimini göz önünde bulundurmaktadır. Önerilen yöntem, çeşitli küme geçerlilik indekslerini optimizasyon için amaç fonksiyonu olarak kabul edebilir. Hem yapay hem de gerçek dünyadaki veri kümeleri üzerindeki deneysel sonuçlar, önerilen tekniğin bulanık C-means, bulanık K-means ve literatürde önerilen iki evrimsel kümelenme tekniğine üstünlüğünü göstermektedir.

Armano ve Framani (2016), Parçacık sürü optimizasyonu kullanarak çok amaçlı kümeleme analizi hakkında yaptıkları çalışmalarında, kısmi kümelenmeyi, çok amaçlı bir optimizasyon problemi olarak tanımlamışlardır. Amaç, iyi ayrılmış, birbirine bağlı ve kompakt kümeler elde etmektir ve bu amaçla, veri bağlantısı ve uyum kavramlarına dayanan iki amaç fonksiyonu tanımlanmıştır. Bu bağlamda kapsamlı bir deneysel çalışma yürütülmüş ve elde edilen sonuçlar diğer dört modern kümeleme tekniğiyle karşılaştırılmıştır. Önerilen algoritmanın en uygun sayıda kümeye ulaşabileceği ve diğer yöntemlere göre daha iyi performans gösterdiği sonucuna ulaşılmıştır.

Alam ve arkadaşları (2015) makalelerinde kümeleme tekniklerinin yeterliliği ve elde edilen kümelerin doğruluğu gibi konularda sorunların yaşandığını vurgulamışlardır ve bu sorunların üstesinden gelmek için optimizasyon tabanlı tekniklere yönelmişlerdir. Parçacık sürü optimizasyonuna dayalı hiyerarşik teknikler için evrimsel parçacık sürü optimizasyonu (EPSO) ve hiyerarşik parçacık sürü optimizasyonunu (HPSO) seçmişlerdir. UCI makine öğrenme veri havuzundan ve yerel olarak bir web sunucusundan toplanılan gerçek verilerden, farklı veri kümeleri üzerinde önerilen kümeleme tekniklerini değerlendirmişlerdir. Önerilen tekniklerin performansını ölçmek için kümeler arası mesafeleri ve uygulama süresini kullanmışlardır. Değerlendirme için K-ortalamalar, PSO kümeleme, Hiyerarşik

Aglomeratif Kümeleme (HAC) ve Gürültülü Uygulamaların Yoğunluğa Dayalı Mekansal Kümelenmesi (DBSCAN) gibi karşılaştırmalı olarak kullanılan farklı kümeleme teknikleri seçilmiştir. Sonuçlar, önerilen tekniklerin belirtilen ölçütlere göre daha iyi performans gösterdiğini doğrulamaktadır.

Aher ve Metre (2014), çalışmalarında Parçacık Sürü Optimizasyonu ve kümelenme zorluklarını çözmek için varyanslarının uygulanabilirliği üzerine genel bir bakış ortaya koymuşlardır. Ayrıca çok boyutlu verileri kümelemek için Subtraktif Kümeleme tabanlı Sınır Kısıtlı Adaptif Parçacık Sürü Optimizasyonu (SCBRAPSO) algoritması olarak adlandırılan gelişmiş bir Parçacık Sürü Optimizasyonu algoritması önermişlerdir. Bu algoritmaların karşılaştırılması sonucunda boyutsallık ile, zaman karmaşıklığının veri kümesinin büyüklüğü ile doğru orantılı olduğunu gözlemlemişlerdir. Gelecekteki çalışmada, daha iyi bir performans elde ederken, algoritmanın karmaşıklığını en aza indirmeyi amaçlamışlardır.

Wan (2013), çalışmasında Tayvan'daki geniş bir alanda yer alan Shei Pa Milli Parkı'nın heyelan sorunları üzerinde durulmuştur. Baskın özellikler ve eşikleri ortaya çıkarmak için analizde, etkilenen değişkenlerin heyelan oluşumuna ilişkin entropinin hesaplanması ve heyelan duyarlılık haritalarının oluşturulmasındaki zorlukların giderilmesi şeklinde iki aşama vardır. Burada K-ortalamalar Kümeleme Analizi ile Parçacık Sürü Optimizasyonu (KPSO) kullanılmıştır. Duyarlılık haritasının genel doğruluğu, KPSO için % 86 olarak elde edilmiştir.

Çevik ve Koçer (2013) çalışmalarında, esnek hesaplama tabanlı yapıda meme kanserinin teşhisine yönelik bir yazılım oluşturmuştur. Yazılımda, Parçacık Sürü Optimizasyonu ile Yapay Sinir Ağları teknikleri birlikte uygulanmıştır. Veri tabanından meme kanseri veri kümesi temin edilmiştir. Elde edilen sonuçlara göre Parçacık Sürü Optimizasyonu ile öğrenme oluşturan Yapay Sinir Ağları tekniğinin hızı ve performansında artış gözlenmiştir.

Hatamlou (2013), çalışmasında kara delik olgusundan esinlenerek yeni bir sezgisel algoritma önermiştir. Diğer nüfus tabanlı algoritmalar gibi kara delik algoritması da çözüme, aday çözümlerin başlangıç popülasyonu ve hesaplanan amaç fonksiyonu ile başlamaktadır. Algoritmanın her yinelemesinde, en iyi aday kara delik olarak seçilir ve daha sonra yıldız adı verilen diğer adayları kendisine çekmeye başlar. Bir yıldız kara deliğe çok yaklaştığında, kara delik tarafından yutulur ve

sonsuza dek yok olur. Böylece, yeni bir yıldız rastgele oluşturulur ve arama alanına yerleştirilerek yeni bir arama başlatılır. Algoritma performansının değerlendirilmesi için, kümeleme problemi çözülmüştür. Geliştirilen algoritmanın diğer sezgisel algoritmalardan daha üstün olduğu sonucuna ulaşılmıştır.

Ortakçı ve Göloğlu (2012) çalışmalarında, kümeleme problemini çözmek için problemi bir optimizasyon problemi olarak ele alarak Parçacık Sürü Optimizasyonu ile çözmüşlerdir. Uygulanan yöntemin, hem doğru kümeleme yapabilmesi hem de küme sayısının bulunması açısından yapay veri seti üzerinde zambak veri setinden daha iyi sonuçlar verdiği görülmüştür. Yapılan çalışmanın sonucunda yapay veri seti kullanıldığında %95, zambak verileri kullanıldığında ise %70 oranında kümeleme başarısı elde edilmiştir.

Alagöz ve Kutlu (2012)'nun çalışmasında, parçacık sürü optimizasyonunun emtia piyasası ürünleri ile portföy optimizasyonu gerçekleştirilmiştir. Elde edilen sonuçlara göre Parçacık Sürü Optimizasyonu ile ortalama varyans modeli sonucunda ortaya çıkan sonuçlar benzerlik göstermiştir. Parçacık hareketlerinin etkin sınıra optimum düzeyde ulaşması ve hisse senedi ağırlıklarının yakın oranlarda bulunması sağlanmıştır.

Khan ve arkadaşları (2010), "Veri Madenciliğinde Optimizasyon için Veri Kümeleme Tekniği ile Parçacık Sürü Optimizasyonu Analizi" adlı çalışmalarında, küme analizi için Parçacık Sürü Optimizasyonu'nun kullanımını incelemeyi amaçlamışlardır. Bulanık C-ortalama kümelenmesinin etkinliği, gelişmiş performans göstererek sürüde daha fazla çeşitlilik sağlamıştır ve parçacıklar değişen çevreyi izlemek için güçlü hale gelmişlerdir.

Özsağlam ve Çunkuş (2008) çalışmalarında, Parçacık Sürü Optimizasyonu ile Genetik Algoritma ve Diferansiyel Evrim Algoritmasının performansları, test fonksiyonları oluşturularak kıyas gerçekleştirilmiştir. Elde edilen sonuçlardan, Parçacık Sürü Optimizasyonunun her iki algoritmaya göre yakınsama hızı ve performans açısından daha iyi çözümler ürettiği tespit edilmiştir.

İKİNCİ BÖLÜM
PARÇACIK SÜRÜ OPTİMİZASYONU

2.1. PARÇACIK SÜRÜ OPTİMİZASYONU TANIMI

2.1.1. Standart PSO

J. Kennedy ve R. C. Eberhart tarafından 1995 yılında geliştirilmiş olan parçacık sürü optimizasyonu, evrimsel bir algoritma olarak ortaya çıkmıştır (Kennedy ve Eberhart, 1995: 1943).

Şekil 1: Parçacık Sürü Optimizasyonu Akış Şeması

Kaynak: Özsağlam ve Çunkaş, 2008: 300.

Kuş sürüleri ve balıklar gibi sosyal hayvanların ortaya koyduğu davranışları temel alan bu yöntem sürekli optimizasyon problemleri için geliştirilmiştir. Bu

hayvanların davranış şekilleri ele alındığında yiyecek bulma esnasında devamlı etkileşimde kalarak sürüden birinin yiyecek bulması halinde sürünün diğer üyelerinin de yönlerini yiyeceğin bulunduğu tarafa çevirdikleri ve hızlarını da bu duruma göre ayarladıkları görülmüştür. Söz konusu bu etkileşim ise PSO ile modellenmiştir. Şekil 1'de PSO'nun akış diyagramı görülmektedir. Bu yöntemdeki tüm çözümlere sürü (swarm) adı verilmiştir. PSO, her ne kadar evrimsel bir algoritma olarak karşımıza çıksa da çaprazlama ve mutasyon gibi etmenler bulunmadığından optimum çözüme daha hızlı yakınsamaktadır. Sürüde bulunan kuşların her birine parçacık (particle) adı verilmektedir. Bu bağlamda her bir parçacık çözümün bir parçasını oluşturmaktadır. Örnek verilecek olursa, D (dimension) değişkenli bir problemin uygunluk fonksiyonunun hesaplanması için n adet parçacık ile başlangıç çözüm denklemi şu şekilde gösterilmektedir :

$$\begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1D} \\ X_{21} & X_{22} & \cdots & X_{2D} \\ \cdots & \cdots & \cdots & \cdots \\ X_{n1} & X_{n2} & \cdots & X_{nD} \end{bmatrix}$$

i. parçacık için uygunluk fonksiyonu aşağıdaki gibidir:

$$F_{fitness=f}(X_{i1}, X_{i2}, X_{i3}, \ldots, X_{iD})$$

Yukarıda verilen matriste parçacık olarak yer alan her satır bir çözüm anlamına gelmektedir. n parçacık için n adet çözüm bulunmaktadır. Parçacıkların hareket etmesinde iki önemli parametre bulunmaktadır. Her iterasyonda her birey için en iyisi pbest ve tüm bireyler için en iyi gbest elde edilmektedir. n adet parçacık ile çalışan bir PSO'da her iterasyonda n adet pbest bulunuyorken yalnızca bir adet gbest bulunmaktadır. Algoritmanın ilk adımında her parçacık için başlangıç değeri, pbest değeridir. Sonraki iterasyonlarda diğer parçacıkların durumuna göre konumu güncellenerek mevcut konum önceki pbest ile kıyaslanır. pbest'ten daha iyi bir çözüm olması durumunda yeni çözüm pbest olarak atanır (Erdoğmuş ve Yalçın, 2015: 16).

PSO algoritması için takip edilmesi gereken akış diyagramı şekil 2'de açıklanmıştır.

Şekil 2: PSO Algoritmasının Kod Yapısı

```
For     her parçacık için
        {başlangıç koşullamaları}
End
Do      {
        For     her parçacık için
                {uygunluk değerini hesapla eğer
                uygunluk değeri, pbest ten daha iyi
                ise; şimdiki değeri yeni pbest olarak
                ayarla}
        End
        Tüm parçacıkların bulduğu pbest değerlerinin
        en iyisini, tüm parçacıkların gbest'i
                olarak ayarla
        For     her parçacık için
        {(8) denklemine göre parçacık hızını hesapla
        (9) denklemine göre parçacık pozisyonunu
güncelle}
        End
        }
While   {iterasyon sayısı ≤ maksimum iterasyon sayısı
        veya eğitim ölçütü ≤  hedef eğitim ölçütü }
```

Kaynak: Çevik ve Koşer, 2013: 42.

Parçacık sürü optimizasyonunda türev bilgisine gerek duyulmaması, bu yöntemi klasik optimizasyon tekniklerinden farklı kılmaktadır. Diğer evrimsel algoritmalara bakıldığında parçacık sürü optimizasyonu algoritmasının gerçekleşmesi ve ayarlanması gereken parametre sayısının az olması nedeniyle uygulaması kolaydır (Çavuşlu vd, 2010: 83). PSO yöntemi, sipariş miktarının belirlenmesi, çizelgeleme problemleri, güç ve voltaj kontrolü, motor parametrelerinin belirlenmesi, tedarik seçimi gibi bir çok optimizasyon problemlerinde başarı ile kullanılmaktadır (Özsağlam ve Çunkaş, 2008: 300).

2.1.1.1. Başlangıç Değerleri

PSO performansındaki ilave değişkenlik, rastgele başlangıç konfigürasyonlarının kullanılmasından kaynaklanmaktadır. Sürüdeki bir parçacık için başlangıç pozisyonu, problem uzayının her bir boyutu boyunca tek tip rastgele bir sayı çizilerek seçilir. Arama alanının geniş kapsamını sağlamak için, parçacıklar, alan boyunca mümkün olduğunca eşit bir şekilde dağıtılacak şekilde başlatılmalıdır. Standart hedefleme yöntemi, özellikle yüksek boyutlu alanlarda bu hedefe ulaşmakta başarısız olur (Richards ve Ventura, 2004: 1).

2.1.1.2. Konum Değeri

Her parçacık, rastgele başlatılan kendi konumuna ve hızına sahiptir. Başlatmadan sonra, parçacıklar en iyi konumlarını kendi ve komşuluk deneyimlerinden öğrenmektedirler. Her parçacık *p*best ve *g*best denilen iki pozisyonu muhafaza etmek durumundadır. *p*best, parçacıkların en iyi pozisyonunu, *g*best ise tüm parçacıklar arasında en iyi küresel konumu temsil etmektedir. Her parçacığın konumu ve hızı aşağıdaki denklemlere göre güncellenir:

$$v_i\ (\mathrm{t}+1) = v_1(\mathrm{t}) + c_1 * r_1(p_{best} - n_i(\mathrm{b})) + c_2 * r_2(g_{best} - x_i(\mathrm{t}))$$

$$x_i\ (\mathrm{t}+1) = x_i\,(t) + v_i\,(\mathrm{t}+1)$$

Burada x_i konum, v_i hız, *p*best en iyi kişisel konumu ve *g*best küresel en iyi pozisyonu temsil etmektedir. Benzer şekilde, r_1 ve r_2 iki rastgele sayı ve aralıkları [0, 1] dir. C_1 ve C_2, sırasıyla, ivme katsayıları olarak adlandırılan biliş ve sosyal bileşeni temsil eden öğrenme faktörleridir (Imran vd., 2014: 2)

2.1.1.3. Hız Değeri

Her parçacığın hızı aşağıda verilen denklemle güncellenir (Blondin, 2009: 12):

v_i (t + 1) = w v_1(t)+$c_1 r_1[\hat{x}_i(t) - x_i(t)]$+$c_2 r_2[g(t) - x_i(t)]$

Formulde i parçacık indeksini, w atalet katsayısını, c_1, c_2 ivme katsayısını $0 \leq c_1$, $c_2 \leq 2$, r_1, r_2 rastgele değerleri ($0 \leq r_1$, $r_2 \leq 1$ her hız güncellemesini yeniden oluşturmaktadır. Denklem içerisindeki $v_i(t)$, t anında parçacık hızı, $x_i(t)$, t anında parçacığın konumu, $\hat{x}_i(t)$ t anında parçacığın kişisel en iyi değeri, $g(t)$ t anında sürünün en iyi çözümü anlamına gelmektedir.

2.1.1.4. Atalet Ağırlığı

PSO algoritması, sürünün taradığı alanın genişliği hakkında bilgi vermektedir. Bu bağlamda eski hız değeri, sürünün geniş bir alanı taramasını sağlar, eşitliğin ikinci kısmı ise sürünün daha dar alanı taramasını sağlar. Elde edilen verilere ek olarak algoritma mümkün olduğu kadar geniş bir alanı taramalı ve optimuma ulaşmalıdır. Ulaşılan sonuçlar bağlamında sürünün davranışı sözlü olarak ifade edilebilmektedir. Başlangıçta geniş bir alan taranmalı, sürü optimuma ulaştıkça alan küçültülerek optimum değer elde edilmelidir. Bunun gerçekleşmesi için eski hız değerinin eşitlik içinde kontrol edilmesi gerekmektedir (Altınöz ve Yılmaz, 3).

2.1.1.5. Hızlandırma Katsayıları

Hız güncelleme denkleminde hızlanma katsayıları olarak adlandırılan iki önemli parametre olan c_1 ve c_2 bulunmaktadır. c_1, bilişsel ivme katsayısı olarak adlandırılırken, c_2 sosyal ivme katsayısını temsil etmektedir. c_1'in daha yüksek değerleri, parçacıkların arama alanındaki sapmasının daha büyük olmasını sağlarken, c2'nin daha yüksek değerleri, mevcut küresel en iyi duruma (gbest) olan yakınlaşmayı göstermektedir. PSO'da arama alanının araştırılması ve kullanılması arasında daha iyi bir uzlaşma sağlamak için zaman değişkeni hızlanma katsayıları ortaya atılmıştır. c_1'in c_{1i}'den c_{1f}'ye başlangıç değerinden düşmesine izin verilirken,

c_2 c_{2i}'den c_{2f}'ye aşağıdaki gibi olan denklemler kullanılarak artırılmıştır. c_{1t} ve c_{2t} değerleri aşağıdaki gibi değerlendirilmiştir (Tripathi vd., 2007: 5039).

$$c_{1t} = (c_{1f} - c_{1i})\frac{t}{\max\_t} + c_{1i},$$

$$c_{2t} = (c_{2f} - c_{2i})\frac{t}{\max\_t} + c_{2i},$$

c_1 ve c_2 değerleri, iterasyonlar ile güncellenir. Denklemdeki r_1 ve r_2 rasgele sayıları her bir boyut ve her parçacık için bağımsız olarak üretilir (Tripathi vd., 2007:5037).

2.1.1.6. Kişisel En İyi Değer

PSO'da parçacığın bireysel deneyimi, pbest niteliğinde yakalanır, bu da şimdiye kadar elde edilen en iyi performansa karşılık gelmektedir. Mevcut çözümün en iyi çözümle karşılaştırılması sonucu ikincisi, yalnızca bu çözüme katkı sağlıyorsa değiştirilir. Pbest işlevi kişisel en iyi değeri vermektedir. (Tripathi vd., 2007:5037).

2.1.1.7. Global En İyi Değer

Gbest terimi o ana kadar elde edilmiş en iyi uygunluk değerine sahip çözümü temsil etmektedir. Çoğunlukla, çok amaçlı optimizasyon problemlerinde yer alan çoklu hedeflerin çelişmesi, tek bir optimum çözümün seçilmesini zorlaştırmaktadır. Bu sorunu çözmek için, domine edilmeyen çözümlerin bir arşivi korunmakta ve bir çözüm gbest olarak alınmaktadır (Tripathi vd., 2007:5037).

2.1.1.8. Sonlandırma Kriteri

Sonlandırma kriteri, önceden tespit edilmiş jenerasyon sayısı ya da herhangi bir uygun çözüm sunabilmektedir. Sonlandırma kriteri test edilerek kriter sağladığı durumda algoritma sonlandırılır. Ancak kriter sağlamadığı durumlarda bir sonraki

aşamaya geçilir ve yeni çözüm arayışı devam eder (Özsağlam ve Çunkaş, 2008: 301).

2.1.2. PSO Algoritması

Mevcut probleminin çözümü için uygunluk fonksiyonu ve değişken sayısı tespit edildikten sonra parçacık sürü optimizasyonu algoritmasında şu adımlar takip edilerek uygulanabilmektedir (Akbulut, 2009: 10):

1.Adım: Başlangıç populasyonu oluşturma aşamasında, probleme ait kısıtların olduğu durumlarda, populasyondaki parçacıklar çözüm uzayı içerisinde bulunmalıdır. Bu çerçevede rastgele olacak şekilde değerler kullanılarak "n" boyutlu bir optimizasyon probleminde "n" tane boyut için en yüksek ve en düşük değerlerin arasındaki bir başlangıç populasyonu oluşturulur. Oluşturulan parçacıkların $(X_1, X_2, \ldots., X_n)$ koordinatları t=0 anındaki konumlarını temsil etmektedir.

2. Adım: Bu aşamada tüm parçacıklar için uygunluk fonksiyonu hesaplanır. Başlangıç konumu, mevcut parçacığın içinde bulunduğu tek konum olması nedeniyle parçacığın en iyi konumunu oluşturmaktadır. Bu bağlamda ilk küresel en iyi, konumlar arasındaki en iyi uygunluk değerine sahip konum olarak değerlendirilmektedir.

3. Adım: "n" boyutlu bir hız vektörü, tüm parçacıklar için rastgele olarak oluşturulur.

4. Adım: Her parçacığın konumu, kendi hız değerine göre güncellenir.

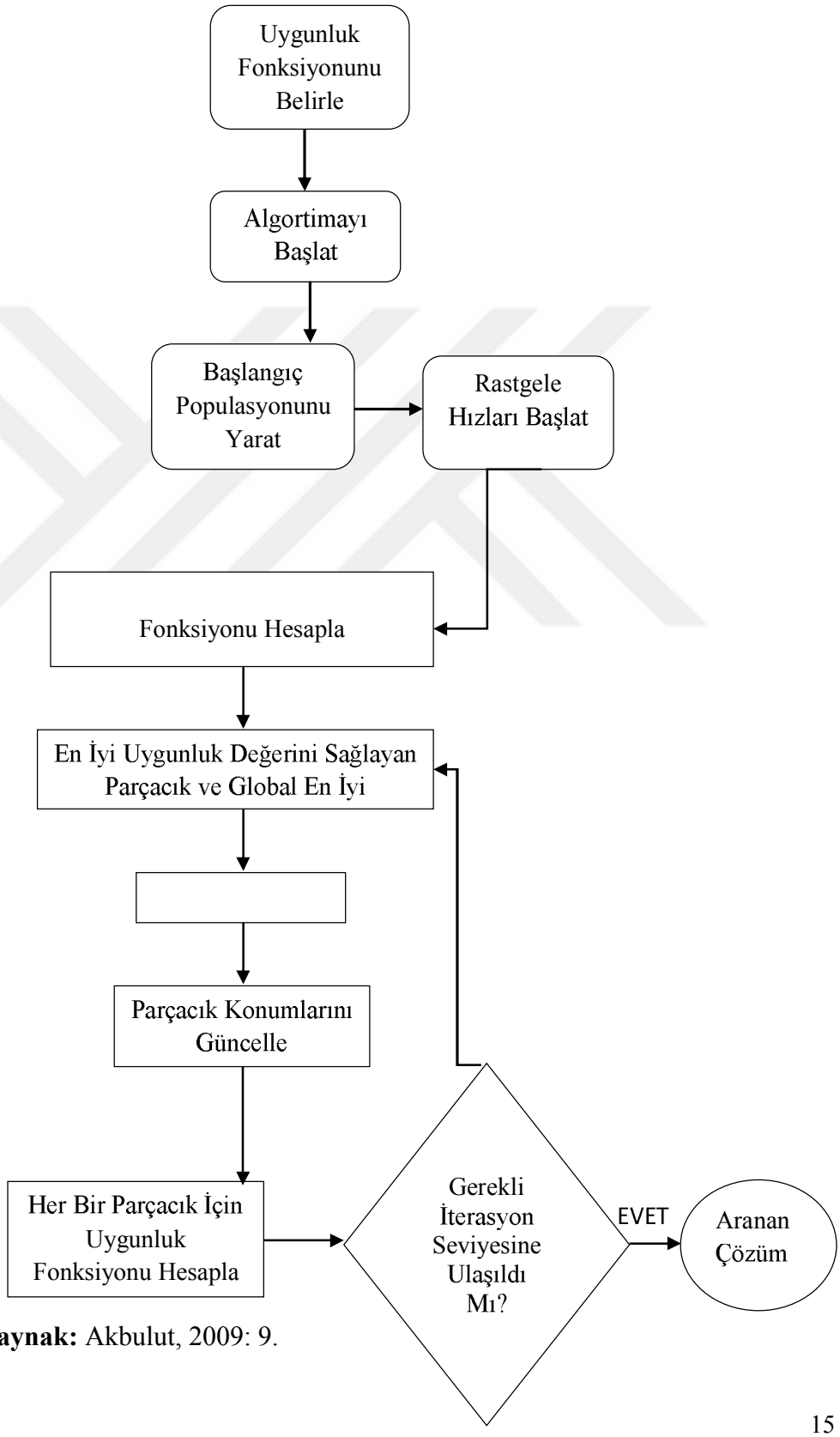
5. Adım: Parçacıkların yeni konumlarına göre uygunluk fonksiyonu hesaplanır. Bu sırada parçacığın yeni hesaplanan uygunluk değerinin, önceki kişisel en iyi uygunluk değerinden daha iyi olduğu durumda, o anki konum kişisel en iyi konum olarak atanır. Aynı şekilde tüm sürü için hesaplanmış olan uygunluk değerleri arasından bir değerin, önceki en iyi küresel değerden daha iyi olduğu durumda, o anki en iyi uygunluk değerini gösteren konum, küresel en iyi konum olarak atanır.

6. Adım: PSO hız vektörü denklemi aracılığı ile tüm parçacıklar için yeni hız vektörleri hesaplar. Böylece parçacıkların konumları güncellenmiş olmaktadır. Hız vektöründeki rastgele öğeler, algoritmayı rastsal hale getirmektedir. Belirlenen yeni konumlar kullanılarak 5. Adım tekrarlanır.

7. Adım: Durdurma koşulları sağlanana kadar 5. ve 6. adımlar tekrar edilir. Genellikle belli bir iterasyon sayısına ulaşıldığında durdurma gerçekleştirilir.

Şekil 3: Parçacık Sürü Optimizasyonu Algoritmasının Akış Şeması

Algoritmanın akış diyagramı aşağıdaki gibidir.



Kaynak: Akbulut, 2009: 9.

2.2. PARÇACIK SÜRÜ OPTİMİZASYONU İLE KÜMELEME

PSO'daki tüm parçacıklar kümeleme probleminin çözümü için potansiyel bir öneridir. Kümeleme problemlerinde de, sürüde bulunan tüm parçacıklar küme merkezlerini temsil etmektedir ve optimum küme merkezini bulmaya çalışmaktadır. Merkezlere olan uzaklıklar göz önünde bulundurularak veri setinde bulunan tüm elemanlar en yakın kümeye atanmakta ve böylece en uygun kümeleme bulunmaya çalışılmaktadır. Sürünün başlangıç popülasyonu X ; X= $\{X_1, \dots, X_i, \dots, X_{pop}\}$.

Pop = sürüdeki parçacık sayısı

Kümeleme işlemleri sırasında da her bir parçacık, oluşturulan kümelerin merkezini gösteren birer vektördür.

Her bir parçacık, $X_i = \{\overrightarrow{O_1}, \dots, \overrightarrow{O_2}, \dots, \overrightarrow{O_J}, \dots \overrightarrow{O_k}\}$.

k = oluşturulan küme sayısı

O_J *= J. kümenin ağırlık merkezi*

Veri setinde kümelenecek olan her elemanın boyutu ile ağırlık merkezinin boyutu aynıdır. Her bir boyutta gerçek değerler bulunmaktadır. Küme sayısının önceden tespit edilmediği durumlarda doğru kümelemeyi bulma ve veri setinin kaç kümeye ayrılacağı başka bir optimizasyon konusunu oluşturmaktadır (Ortakçı ve Göloğlu, 2012: 4).

Kümeleme için iki nesne arasındaki benzer uzaklıklar tanımlanmalıdır. Bu amaçla aşağıda gösterildiği şekilde Minkowski uzaklığından türeyen Öklid mesafesi kullanılmaktadır. (Zhang vd., 2010: 4762) :

$$D(o_i, o_l) = \left(\sum_{j=1}^{m} (o_{ij} - o_{lj})^r\right)^{1/r} \Rightarrow D(o_i, o_l) = \sqrt{\sum_{j=1}^{m} (o_{ij} - o_{lj})^2} \qquad (1)$$

$D(o_i, o_l)$: i. ve j. Nesneler arasındaki benzemezlik ölçüsü

o_{ij}: i. nesnenin j. özelliğinin değeri (i = 1,. . . ,n and j = 1, . . . ,m)

m boyutlu uzaydaki n adet nesnelere ait veri seti K kümeye bölünür. Böylece, kümeleme probleminin matematiksel formülasyonu aşağıdaki gibidir (Shelokar v.d, 2004,190) :

$$\text{Min} \sum_{k=1}^{K} \sum_{i=1}^{n} w_{ik} D(o_i, z_k) \quad (2)$$

$$\sum_{k=1}^{K} w_{ik} = 1 \qquad i = 1, \dots, n \quad (3)$$

$$\sum_{i=1}^{n} w_{ik} \geq 1 \qquad k = 1, \dots, K \quad (4)$$

z_k, k kümesinin merkezini, w_{ik}, eğer $w_{ik} = 1$ ise k. kümeye ait i. nesneyi gösterir.

W_{ik}^t, hesaplanması aşağıdaki gibidir:

$$W_{ik}^t = \begin{cases} 1 \text{ eğer } k = \text{argmin}_{\delta=1\dots,K} D(o_i, z_\delta^{t-1}), \; i = 1, \dots, n, k \\ 0 \text{ diğer durumlarda} \end{cases} \quad (5)$$
$$= 1, \dots, K$$

Her bir nesneyi 6.denklem kullanılarak t-1 iterasyonda güncellenen tüm küme merkezlerinden en yakın küme merkezine atayacağız. Bu 3.denklemde verilen kısıtı sağlar. Böylece t. İterasyondaki her bir kümenin merkezi aşağıdaki gibi hesaplanır:

$$\boldsymbol{Z_{kj}^t} = \frac{\sum_{i=1}^{n} W_{ik}^t o_{ij}}{\sum_{i=1}^{n} w_{ik}^t}, \; k = 1, \dots, K \; , \; j = 1, \dots, m \quad (6)$$

Aşağıdaki denklem t iterasyonda p parçacığının uygunluk değerini vermektedir:

$$f_p^t = \sigma_p^t = \sum_{k=1}^{K_p} \sum_{i=1}^{n} W_{ik}^{pt} D\left(o_i, z_k^{pt}\right) \quad (7)$$

Küme sayısının bilinmediği durumda kullanılan uygunluk fonksiyonu ise aşağıdaki gibidir:

$$f_p^t = \sigma_p^t - \min_{k \neq l} D(z_k^{pt}, z_l^{pt}) \quad (8)$$

$$\upsilon K_p^{t+1} = \upsilon K_p^t + \omega_1 \left(GK_b - K_p^t\right) + \omega_2 \left(GK_p - K_p^t\right) \quad (9)$$

$$\upsilon Z_{kj}^{p,t+1} = \upsilon Z_{kj}^{pt} + \omega_1\left(GZ_{\partial j}^{b} - Z_{kj}^{pt}\right) + \omega_2\left(GZ_{\partial j}^{p} - Z_{kj}^{pt}\right) \quad (10)$$

ω_1 ve ω_2, 0 ila 2 arasında üniform rastgele sayılar, b ise sürüdeki en iyi parçacıktır.

υK_p^t ve υZ_{kj}^{pt} , parçacıkların boyutunu gösterir.

∂, en iyi çözüm kümesini veren indeks sayısıdır ve aşağıdaki gibi hesaplanmaktadır:

$$\partial = \text{argmin}_{l=1\ldots,GK_b} \mathrm{D}\left(GZ_l^b, Z_k^{pt}\right) \text{ ya da } \partial = \text{argmin}_{l=1\ldots,GK_p} \mathrm{D}\left(GZ_l^p, Z_k^{pt}\right) \quad (11)$$

t+1 iterasyondaki p parçacığının hareketi aşağıdaki gibidir:

$$K_p^{t+1} = max(round\left(\upsilon K_p^{t+1} + K_p^t\right), 1) \quad (12)$$

$$Z_{kj}^{p,t+1} = \upsilon z_{kj}^{p,t+1} + Z_{kj}^{pt} \quad (13)$$

K_p boyutunda p parçacığının hareketinden sonra mevcut çözümün rastgele seçilen küme merkezinden yalnızca birine hareket eder.

En küçük kümeyi tanımlamak için, her kümenin S_k^{pt} boyutu aşağıdaki gibi hesaplanır:

$$S_k^{pt} = \sum_{i=1}^{n} W_{ik}^{pt} \quad k = 1, \ldots, K_p^t \quad (14)$$

Eşitlik 14' e göre $K_p^{t+1} > K_p^t$ olduğu durumda,

en büyük küme iki yeni kümeye bölünür ve bu bölme işlemi $K_p^{t+1} = K_p^t$ sağlanana kadar tekrar edilir. İlk olarak en büyük kümeye bölmek için her bir özellik için minimum ve maksimum değer bulunarak aralık tanımlanır, daha sonra yeni küme merkezi bu aralığın içinde rastgele bir şekilde belirlenir.

Parçacık Sürü Optimizasyonu ile Kümeleme algoritması aşağıdaki gibidir:

BEGIN
$t = 0$
for $p = 1$ **to** P
başlangıç parçacık p yi rastgele oluştur
her nesneyi en yakın p parçacık kümesine ata // Denklem 5'e bak
p parçacığının kümelenme merkezlerini güncelle // Denklem 6'ya bak
8. denklemi kullanarak *fp* yi hesapla
p parçacığı en iyi olan başlangıç konumunu hafızaya al
end for
while $t < maksimum\ iterasyon\ sayısı$
Deb'in kuralına göre sürünün en iyi parçacığını bul
$t = t + 1$
for $p = 1$ **to** P
K boyutunda p parçacığını taşı // Denklem 12'ye bak
while $K_p^t < K_p^{t-1}$
çıkarma prosedüründe rastgele çıkar
end while
while $K_p^t > K_p^{t-1}$
ekleme prosedüründe rastgele ekle
end while
$k = 1$ ve K_p^t arasında rastgele tam sayı
for $j = 1$ **to** m
zkj boyutunda p parçacığını taşı *zkj* // Denklem 13'e bak
end for
her nesneyi en yakın p parçacık kümesine ata // Denklem 5'e bak
p parçacığının kümelenme merkezlerini güncelle // Denklem 6'ya bak
8. denklemi kullanarak *fp* yi hesapla
Deb'in kuralına göre eğer mevcut pozisyon hafızaya alınan pozisyondan daha iyiyse, daha sonra p parçacığı en iyi mevcut pozisyonu hafızaya alır
end for
end while

Kaynak: Cura, 2012: 4.

ÜÇÜNCÜ BÖLÜM

VERİ MADENCİLİĞİ

3.1. VERİ MADENCİLİĞİNİN TANIMI

Günümüzde bilgi kavramının dönüşüm yaşaması ve bilgi yoğunluğunun artması, beraberinde yeni sistemleri getirmiştir. İşletmelerin, kurum ve kuruluşların sistemlerini oluşturmada, politikalarını yapılandırmada güvenirlik ve sağlıklı bilgi akışına ihtiyaç duyulmaktadır. Bilgi ve verilerin hayatın her alanında kendine yer bulmasıyla işletmeler ve şirketlere bilgi yoğunluğu külfet oluşturmaktadır. Bilginin toplanması, depolanması ve saklanmasının başlı başına meşakkatli bir durum olması işletmeleri veri tabanlarında veya dosyalama yapılarında çözüm bulmaya itmiştir. Biriken ve gerekli olan verilerin temini, depolanması, saklanması, ihtiyaç duyulduğunda tekrar temin edilebilmesi veri madenciliği kavramını ortaya çıkarmıştır (Ahi, 2015: 22).

Veri madenciliği, verilerin büyüklüğüne göre analiz ederek verilerin anlamlı bir yapıya dönüşümünü sağlayan bilgi telhis sürecidir. Veri madenciliği karmaşık bilgilerin arasından gerekli olan bilginin çıkarılması, yani değerli madenin kazıp çıkarılması manasından türetilmektedir**.** Veri madenciliği, veri tabanlarında çeşitli veriler arasında, gerekli görülen, yararlı bulunan, işlevsel verilerin tespit edilmesi ve kullanılması sürecini oluşturmaktadır. Veri madenciliği ile birlikte var olan sorunlara çözüm yolları meydana getirilmektedir. Karar mekanizmalarının oluşmasında kullanılan bir araç olarak da yorumlanabilmektedir (Sumathi, 2006: 9).

Veri madenciliği ile ilgili tanımlar ise şu şekildedir (Tüzüntürk, 2010: 67).

- Veri madenciliği geniş verilerden temel bilginin ortaya çıkarılma işlemidir. Diğer bir tabirle, karmaşık bir veri yapısının içinden değerli unsurların ve beklenmedik öğelerin ortaya çıkarılması bilimidir.
- Veri madenciliği, değerli verilerin ortaya çıkarılması amacıyla birçok analiz yönteminin kullanıldığı süreçtir.
- Kuonen (2004)'e göre veri madenciliği, iş akışı için verilen kararların doğru, yararlı, diğer ön görülerden farklı, algılanabilir örüntüleridir.

- Gartner Group'a göre veri madenciliği; Matematiksel ve istatistiksel yöntemlerle örüntü verilerinin ayıklanarak, ihtiyaçların bulunması ve değerli olan bilginin ortaya çıkarılması işlemidir (Larose, 2005: 2)
- Veri madenciliği, birçok veri arasında saklı olan, kıymetli, ihtiyaç duyulan ve kullanışlı bilgilerin ortaya çıkarılmasıdır (Koyuncugil, 2007: 1).
- Bilgisayar teknolojisinin kullanılması ile verilerin hızlı şekilde kaydedilmesi ve depolanması, buna bağlı birçok yardımcı araç ile büyük hacimli verilerin bilgi dâhilinde olmayan, saklı, örtük, klasik metotlarla gün yüzüne çıkarılması işlemleridir (Şentürk, 2006: 4).

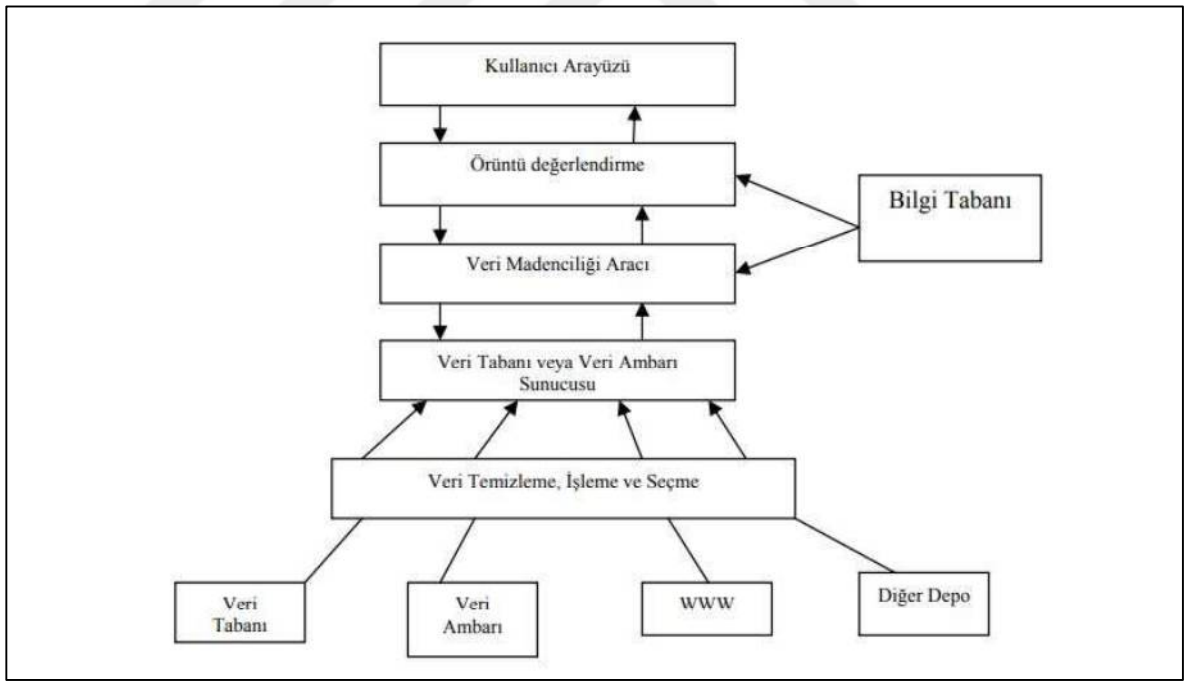
Veri madenciliği yeni keşfedilen bir istatistiki yapı değildir; veri işlemlerinin artması ve verilerin dönüştürülmesi eksikliklerinin tamamlanması durumlarını düzenlemek amacıyla ortaya çıkan bir kavramdır. Verilerin oluşturulması, tanımlanması, örüntü tanıma gibi veritabanı ile ilgili disiplinleri içermektedir. Klasik istatistiksel uygulamalar dışında veri madenciliği daha kapsamlı bir yapıya sahiptir. İstatistikçi, veri kümesi içerisinde bin veri incelerken, veri madencisi milyonlarca veri dâhilinde hareket etmektedir. İstatistiki verilerin düzenlenmesi ve hesaplanması için belirli yöntemler ve iyi rehberlikler gerekmektedir. Bu sebeple veri madenciliği tek bir yapıdan oluşmamakla birlikte birçok bileşenden meydana gelerek verilerin düzenlenmesini sağlamaktadır (Tüzüntürk, 2010: 73).

Veri madenciliği yapısı birden fazla bileşenden oluşmaktadır. Bu bileşenler şu şekilde yer almaktadır (Han ve Kamber, 2012: 10):

- **Veritabanı, Veri Ambarı veya Diğer Bilgi Depoları:** Veriler, veritabanı, veri ambarı ve diğer bilgi depolarında toplanmaktadır. Veri üzerinde birçok işlem gerçekleştirilebilmektedir. Bu işlemlerden bazıları: veri ekleme, veri temizleme ve veri bütünleme şeklindedir.
- **Veritabanı veya Veri Ambarı Sunucusu:** Veritabanı veya veri ambarı sunucusu, kullanıcının yapacağı işlere yönelik gerekli olan verinin veritabanından çıkartılma işlemidir.
- **Bilgi Temeli:** Araştırma için gerekli olan bilgilerin veya ilgili örüntülerin ele alındığı kısmın temelini oluşturmaktadır. Niteleyiciler ve onların değerini organize edebilmek için kullanılmaktadır.

- **Veri Madenciliği Motoru:** Veri madenciliği motoru, veri madenciliğinin önemli analiz kısımlarını, sınıflandırma ve kümeleme analizi için esas olan öğelerden biridir.
- **Örüntü Değerlendirme:** Örüntü değerlendirme, gerekli olan veya farklılık arz eden örüntülerin açığa çıkartılması işlemidir. Uygulanan veri madenciliği yöntemine dayalı olarak, veri madenciliği modülü dâhilinde de bulunabilmektedir.
- **Grafiksel Kullanıcı Arayüzü:** Grafiksel Kullanıcı Arayüzü aşaması, veri madenciliği ile veri madenciliğini kullanan arasında ilişkiyi oluşturan kısımdır. Kullanıcı veri tabanı ile ilişkili bağlantılar gerçekleştirebilmektedir. Bu bağlantı, kullanıcının veri madenciliğinin oluşturulmasında katkı sağlamak amacıyla bilgi girişi yapması şeklindedir. Bu bileşen ile kullanıcı, veri ambarı ve farklı yapıdaki örüntüleri görüntülemelerini ve taramalarını da sağlamaktadır.

Şekil 4: Veri Madenciliğinin Tasarımı



Kaynak: Özdemir, Aslay ve Çam, 2010: 253.

Veri madenciliğinin işletmeler tarafından ürünün hizmet alanında kullanılma amaçları aşağıda yer almaktadır (Özmen, 2001: 3):

- Bir işletme, Pazar stratejilerini belirlerken müşteri istek ve ihtiyaçlarına yönelik davranmaktadır. Rakip şirketleri tercih eden müşteriler ve müşterilerin neden rakip şirketi tercih ettikleri veri madenciliği ile tespit edilmektedir. Bu sayede müşterilerin kayıp edilmemesi ve kayıp edilen müşterilerin ise geri getirilmesi sağlanabilmektedir.
- Müşterinin ürünü hangi sebeple tercih ettiği, hangi ürünlere daha fazla yöneldiği ve memnuniyet durumları tespit edilebilmekte ve bu tespitlere yönelik hareket edilebilmektedir.
- Karlı müşteriler tespit edilerek, müşterilere özel kampanyalar gerçekleştirilmektedir. Bu sayede masraflı olan müşterilerin masrafları azaltılabilmektedir.
- Ürün tanıtımı için hedef kitlenin tespiti ve hedef kitleye ulaşılmasını sağlamaktadır.

3.2. VERİ MADENCİLİĞİ SÜRECİ

3.2.1. Verilerin Hazırlanması (Veri Önişleme)

Veri önişleme, veri madenciliğinin hazırlanması olarak tanımlanmaktadır. Verinin hazırlanması, analiz edilmesi bir süreçtir. Bu süreçte iyi bir analiz ve ayıklama işlemlerinin yapılması gerekmektedir. Veri tabanlarının hazırlanması ve düzenlenmesinde dikkat edilmesi gereken hususlar ise şu şekilde sıralanmaktadır (Özmen, 2003: 37-39):

- Amaca uygun olmayan tüm veriler çıkartılmalıdır,
- Eksik veya hatalı bir şekilde yapılan işlemler verilerden ayrıştırılmalıdır,
- Eksik şekilde girilmiş verilerin sistematik hataları oluşturup oluşturmadığı tespit edilmelidir,
- Sürekli olarak tekrar eden ve aynı anlamı taşıyan veriler kontrol edilip silinmelidir,
- Eklenecek yeni bir verinin zaman harcamaya değer olup olmadığı tespit edilerek işlemlerin buna göre yapılması gerekmektedir.

3.2.1.1. Veri Temizleme

Verilerin analiz edilebilmesi ve veri madenciliğin gerçekleşmesi amacıyla kullanılan yöntemdir. Veri temizleme, eksik ve tamamıyla farklı verilerin uyumsuzluğunu gidermektedir. Verilerin karmaşık bir yapıda olması güvensiz sonuçları beraberinde getirmektedir. Bu da kullanışsız verilerin oluşmasına neden olmaktadır. Temizleme ile değerler ayrıştırılmaktadır. Veri kodlarında hatalar da bulunabilmektedir. Veri transferinde, veri kodlarında bulunan yanlışlara gürültü adı verilmektedir. Gürültüyü kaldırmak için ise kutulama(binning), kümeleme, regresyon teknikleri kullanılmaktadır. Eksik değerler ise şu şekilde temizlenebilmektedir (Oğuzlar, 2003: 71):

1. Eksik değere sahip kayıtlar atılabilir.
2. Eksik değerlerin yerine, değişkenin ortalaması kullanılabilir.
3. Aynı gruptaki tüm veriler için değişkenin ortalaması kullanılabilir.
4. Eldeki verilere dayanarak en uygun değer, eksik değerlerden bulunup kullanılabilir.

3.2.1.2. Veri Birleştirme

Veri madenciliğinde, verilerin düzenli bir yapıya dönüşmesi için veri küpü, düz dosyalar şeklinde birleştirme işlemleri gerçekleştirilmelidir. Bu noktada verilerin birleştirilmesi en önemli husustur. Verilerin birleştirilmesinde eksik verilerin meydana gelmemesi önemlidir. Diğer bir önem arz eden konu ise indirgemedir. Değişkenler arasında meydana gelen tutarsızlık veri setinde fazlalıkları doğurmaktadır. Fazlalıkların tespit edilmesi gerekmektedir. Veri birleşiminde diğer bir husus da ölçek ve kodların uyuşmamasıdır. Verilerin farklılık göstermesi tehdit oluşturabilmektedir (Akın, 2008: 62).

3.2.1.3. Veri Dönüştürme

Modelin daha sağlam bir yapıya kavuşması için verilerin içinde yer alan bazı değişkenlerin sürekli yapıdan sayısal bir aralığa, kategorik bir yapıdan sayısal bir aralığa dönüştürme işlemidir (Akman, 2010: 8).

Veri dönüştürmede en çok kullanılan yöntem ise normalleştirme yöntemidir. Normalleştirme yöntemi, verilerin 0,0-1,0 gibi aralıklara ölçeklenmesini sağlamaktadır. Bu işleme normalizasyon adı da verilmektedir. Normalizasyon işleminin gerçekleştirilmesinde kullanılan teknikler aşağıda yer almaktadır (Bayer, 2011: 23-24):

- **Min-Maks Yöntemi:** Ham verilerin doğrusal şekilde normalize edilmesi yöntemidir. 'Min' bir verinin sahip olabileceği en küçük değeri, 'maks' ise verinin sahip olabileceği en büyük değeri tanımlamaktadır. Bu değerler 0-1 aralığında dönüştürülmektedir. Yöntemin formülü ise şu şekildedir:

$$S' = \frac{\mathrm{S} - \mathrm{min}}{\mathrm{maks} - \mathrm{min}} \left(yeni_{maks} - yeni_{min}\right) + yeni_{min}$$

S': Verinin normalize edilmiş değeri

S : Verinin orijinal değeri

- **Z Skor Yöntemi:** Z skor yönteminde ortalama ve standart sapma değerleri kullanılmaktadır. Z skor yönteminin formülü şu şekildedir:

$$S' = \frac{\mathrm{S} - \mathrm{ort}}{\sigma}$$

σ: Standart sapma

Ort: İlgili alanın ortalaması

- **Ondalık Ölçme Yöntemi:** Gerekli olan verinin ondalık kısmında gerçekleştirilen değişiklik ile veri normalize edilmektedir. Formülü aşağıda yer almaktadır:

$$S' = \frac{\mathrm{S}}{10^a}$$

a: maks(|S'|)< 1 olacak şekilde en küçük tamsayı

3.2.1.4. Veri İndirgeme

Büyük veri setlerinde veri madenciliği işlemleri gerçekleştirilirken, yapılan tüm işlemler maliyet oluşturabilmektedir. Ya da doğru sonuca ulaşmak için verilerin azaltılması gerekebilmektedir. Bu gibi durumlarda sonuca etkisinin az olduğu öğeler her ne kadar uygulamaya alınmış olsa bile öğe sayısı yapılacak işleme göre biri veya birden fazlası çıkartılarak azaltma işlemi yapılabilmektedir (Stepaniuk, 2008: 43).

3.3. VERİ MADENCİLİĞİ UYGULAMA ALANLARI

Veri madenciliğinin uygulama alanları ise geniş bir alana yayılmaktadır. Yoğun olarak bankacılık, sağlık, eğitim, güvenlik, pazarlama ve sigortacılık sektörlerinde kullanılmaktadır. Veri madenciliğinin sektörleri ve kullanım amaçları ise aşağıda yer almaktadır (Akman, 2010: 10):

- **Sağlık ve Farmakoloji:** İlaç oluşturma, yapılandırma, hastalığın bulunmasında, hastalığın iyileştirilme aşamalarının oluşturulmasında kullanılmaktadır.
- **Biyoloji:** DNA sıra analizi ile genlerde bulunan rahatsızlıkların tespitinde kullanılmaktadır.
- **Pazarlama Yönetimi:** Müşterilerin özelliklerinin analiz edilmesi, gerekli özelliklerin tespit edilmesi, hangi ürünü en fazla satın alabileceklerinin analiz edilmesi, pazarlamaya yönelik kampanyaların planlanmasında, var olan müşterilerin tutulması ve yeni müşterilerin elde edilmesi aşamasında, pazar sepeti ve çapraz satış analizlerinde, müşteriler ile işletme arasındaki ilişkinin kurulmasında kullanılmaktadır.
- **Bilişim ve Mühendislik:** İnternet üzerinden gerçekleştirilen işlemler, internet üzerinden yapılan usulsüzlükler ve dolandırıcılıkların tespiti, bilgisayar ağlarına yetkisiz girişlerin tespit edilmesi, kimlik tespitinde parmak izi okuma ve yüz tarama tekniklerinin uygulanmasında, yapay zekâ yapılandırmalarında kullanılmaktadır.
- **Bankacılık:** Finansal veriler arasındaki farklılarının tespiti ve gizli korelasyonların bulunmasında, kredi kartı üzerinden gerçekleştirilen

dolandırıcılık işlemlerinin tespit edilmesinde, kredi başvuru ve taleplerinin oluşturulması ve değerlendirilmesinde, usulsüz işlemlerin, hataların tespitinde, risk analizlerinin oluşturulması ve risk değerlendirmelerin yapılmasında kullanılmaktadır.

- **Meteoroloji ve Atmosfer Bilimleri:** Bölgelere yönelik hava durumu, iklim değişikleri, günlük veriler, hava durumu tahminleri gerçekleştirmek amacıyla kullanılmaktadır.
- **Sigortacılık:** Poliçe talebinde bulanabilecek müşterilerin analizi, sigortacılık alanında yapılan dolandırıcılık işlemlerinin tespitinde ve riske sahip müşterilerin analizinde kullanılmaktadır.
- **Perakendecilik:** Satış durumları veri analizleri, alış verişlerin analizleri, ürün tedariği ve ürünlerin mağazalardaki konumlarına yönelik durumunu en iyi hale getirmek amacıyla kullanılmaktadır.
- **Borsa:** Hisse senedi fiyat tahmini, piyasaların tespiti için, alım ve satımların en iyi hale dönüştürülmesinde kullanılmaktadır.
- **Endüstri:** Ürünün kalitesinin analizinde, ürünlerin üretim ve dağıtımlarını en uygun seviyeye ulaştırmak amacıyla veri madenciliği analizi kullanılmaktadır.
- **Telekomünikasyon:** Hizmetin kalite düzeyinin yükseltilmesi, iyileştirilmesi, hatlardaki yoğunluğu ve telefon ile gerçekleştirilen dolandırıcılıkların tespitinde veri madenciliği kullanılmaktadır.

Veri madenciliği kullanım alanları oransal olarak Tablo 1'de yer almaktadır. Görüldüğü üzere en çok orana % 38.3 ile CRM/ Müşteri Analizleri sahiptir. CRM/ Müşteri Analizlerini sırasıyla bankacılık, dolandırıcılık tespiti, finans, doğrudan pazarlama vb. alanlar takip etmektedir.

Tablo 1: Veri Madenciliği Uygulama Alanları

2008 yılında veri madenciliğini uyguladığınız endüstri/alanlar nelerdir?	
CRM/ Müşteri Analizleri (41)	% 38.3
Bankacılık (34)	% 31.8
Dolandırıcılık Tespiti (21)	% 19.6
Finans (18)	% 16.8
Doğrudan Pazarlama (15)	% 14.0
Diğer (14)	% 13.1
Yatırırım/Borsa kararları (14)	% 13.1
Kredi kartı Skorlama (14)	% 13.1
Telekomünikasyon13)	% 12.1
Perekandecilik(13)	% 12.1
Reklam (13)	% 12.1
Biyoteknoloji/Genetik (12)	% 11.2
Bilim (11)	% 10.3
Sigortacılık (11)	% 10.3
Sağlık (10)	% 9.3
İmalat (9)	% 8.4
E-ticaret (8)	% 7.5
Web Kullanım Madenciliği(8)	% 7.5
Sosyal Politikalar/Anket Analizi(8)	% 7.5
Tıp/Farmakoloji (8)	% 7.5
Güvenlik/Anti-terör (6)	% 5.6
Web içerik madenciliği (6)	% 5.6
Kamu/Askeri uygulamalar (4)	% 3.7
Seyahat (3)	% 2.8
Junk e-posta / Anti-spam tespiti (3)	% 2.8
Eğlence /Müzik (3)	% 2.8
Sosyal Ağlar(2)	% 1.9

Kaynak: Akman, 2010: 11.

3.4. VERİ MADENCİLİĞİ TEKNİKLERİ

Veri madenciliğinde problem durumuna göre birçok istatistiki çözüm tekniği bulunmaktadır. Bu teknikler birçok alanda yoğun bir şekilde kullanılmakta, problemin tespiti yapılmaktadır.

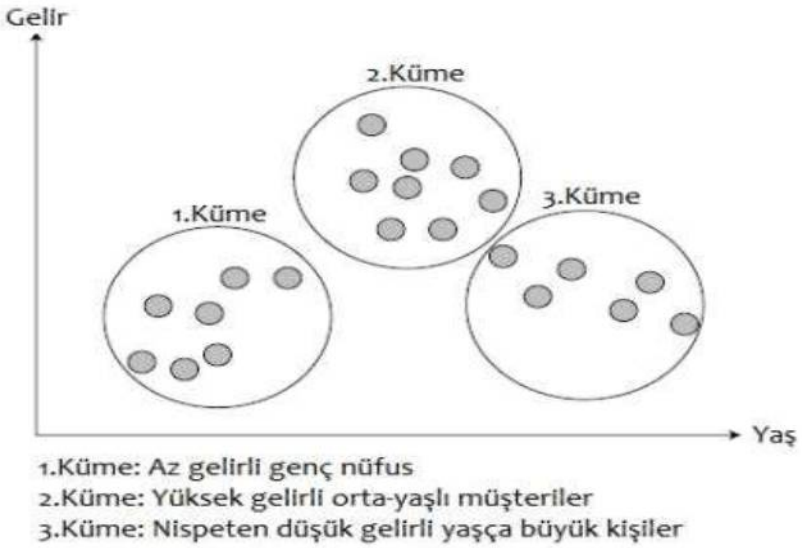
Şekil 5: Veri Madenciliği Yöntemleri

Kaynak: Akçay, 2014: 61.

3.4.1. Kümeleme Teknikleri

Kümeleme teknikleri, veri madenciliğinde sıklıkla kullanılan önemli bir öğedir. Veri tabanlarında meydana gelen karmaşık yapının anlamlı algılara ulaşması kolay olmamaktadır. Var olan verileri ayıklamak, ayrıştırmak, çözümlemek gerekmektedir. Dağınık yapının bir bütünlüğe kavuşması için yöntem ve teknikler kullanılmalıdır. Kümeleme yöntemi ile birbirine benzer veriler bir araya getirilerek homojen kümelere dönüştürülmektedir (Getoor, 2008: 85; akt. Akçay, 2014). Kümeleme yönteminde amaç, veriler arasındaki benzerlik ve mesafelerin tespit edilmesi ve gruplanması böylelikle kolayca işlenebilmesidir (Akpınar, 2000: 6).

Şekil 6: Bir Kümeleme Yöntemi Örneği



Kaynak: Akçay, 2014: 68**.**

Kümeleme analizi X veri matrisinde

$$\times' = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n^1} & \cdots & x_{np} \end{bmatrix}_{np}$$ Gözlemler

şeklinde bulunan birim veya değişkenleri, aynı alt gruplara bölmeye yardımcı olan yöntemler grubudur. Kümeleme analizinin kullanımında iki amaç bulunmaktadır. Bu iki amaç; verileri indirgemek ve özet bilgiler temin edebilmek şeklinde olduğu için kümelemeler uzaklık veyahut benzerlikler ile gerçekleştirilmektedir. Nicel verilerin uzaklığını hesaplayan yöntemlerin beşi aşağıda gösterildiği gibidir: (Tüzüntürk, 2010: 77):

1. Öklit Uzaklığı

$$d(i,j) = \sqrt{\sum_{k=1}^{p} \left(X_{ik} - X_{jk}\right)^2}$$

2. Karasel Öklit Uzaklığı

$$d(i,j) = \sum_{k=1}^{p} \left(X_{ik} - X_{jk}\right)^2$$

3. Mahalanobis Uzaklığı

$$d(x_i, x_j) = D^2 = (\bar{x}_i - \bar{x}_j)'S^{-1}(\bar{x}_i - \bar{x}_j)$$

4. Hotelling Uzaklığı

$$T^2 = \frac{n_1, n_2}{\mathrm{n}} (\bar{x}_i - \bar{x}_j)'S^{-1}(\bar{x}_i - \bar{x}_j)$$

5. Manhattan Uzaklığı

$$d_M(i,j) = \sum_{k=1}^{p} |x_{ik}, x_{jk}|$$

Uygulamanın tamamı nicel değişkenden oluşmamaktadır. Bir kısmı nitel değişkenlerden meydana gelmektedir. Nitel verilerin uzaklıklarının hesaplanması için ise şu işlem uygulanmaktadır:

$$d(x_i, x_j) = \frac{1}{p}\sum_{k=1}^{p} |x_{ik} - x_{jk}| \qquad W_k = \begin{Bmatrix} 1 & Nicel\,Veri\,ise \\ 1/k.değ.dağ & Nitel\,Veri\,ise \end{Bmatrix}$$

Uzaklık matrisinin kullanılması iki kümeleme yöntemi ile gerçekleşmektedir. Küme sayısına göre yöntem kullanılmaktadır. İlk yöntem şu formüller ile hesaplanmaktadır:

1. Tek Bağıntı (En yakın komşuluk)

$$d_{k(i,j)} = \mathrm{Min}\,(d_{ki}, d_{kj})$$

2. Tam Bağıntı (En uzak komşuluk)

$$d_{k(i,j)} = \mathrm{Max}\,(d_{ki}, d_{kj})$$

İkinci yöntem ise hiyerarşik bir yöntem değildir. Küme sayısına önceden karar verilmiştir.

1. k-ortalama tekniği

$$W_a = \frac{1}{n}\sum_{i=1}^{p} min|x_i - a_{jn}|^2$$

2. En çok olabilirlik

$k = (n/2)^{1/2}$

3.4.1.1. Bölümleyici Teknikler

Veri madenciliğinde geniş ve dağınık veri setlerinin kullanılır bir yapıya dönüşmesini sağlamaktadır. Veri madenciliği bilginin keşfedilmesi ve ihtiyaç doğrultusunda kullanılmasını amaçlamaktadır. Bu amaç doğrultusunda yapılandırılmış bölümleyici kümeleme algoritmaları ise verileri n tane giriş parametresi şeklinde kabul eden, veriyi k tane kümeye bölerek hareket eden algoritmalardır. Bu teknikte kümeyi merkez noktaları temsil eder. Bölümleyici algoritmalar verilerin kullanımını sağladığı ve iyi sonuçlara ulaşımı gerçekleştirdiği için sıkça kullanılan algoritmalardandır. Yaygın şekilde kullanılan bölümleyici algoritmalar şu şekilde sıralanmaktadır; k-ortalama, k-medoid, clara ve clarans, c-means algoritmalarıdır (Alzand ve Karacan, 2014: 57).

3.4.1.2. Hiyerarşik Teknikler

Kümelerin aşamalı ve sıralı olmasını sağlayan yöntemdir. Bu yöntem gruplayıcı ve bölücü olarak ikiye ayrılmaktadır. Gruplayıcı yöntemde başlangıç aşamasında her gözlem bir küme olarak kabul edilmektedir. Yakın iki küme birleştirilerek tek kümeye indirilmektedir. Böylelikle küme sayısı her aşamada bir azaltılır. Dendogram ya da ağaç grafiği olarak adlandırılan tekniklerle bu süreç gösterilebilir. Bölücü yöntemde ise gruplayıcı yöntemin aksine tüm gözlemlerden oluşan büyük kümeler meydana getirilmektedir. Her gözlem tek bir küme haline gelene kadar ayıklanmaktadır (Everitt vd., 2011: 96).

3.4.1.3. Yoğunluk Tabanlı Teknikler

Yoğunluk tabanlı tekniklerde uzaklığa dayalı küme seçimi yerine yoğunluğa bağlı küme seçimi gerçekleştirilmektedir. Yoğunluğa dayalı teknikte yoğunluklu alanlar küme tabanlarında bulunurken, seyrek alanlar gürültülü veriler ve küme

sınırlarında yer almaktadır. Farklı büyüklükteki kümelerde kullanılabilecek bir yöntemdir (Bircan ve Çam, 2016: 89).

3.4.1.4. Izgara Tabanlı Teknikler

Bu yöntemde veriler, ızgaralara benzer bölümlerden oluşmaktadır. Kümeleme uygulaması her ızgarada ayrı şekilde gerçekleşmektedir. Izgara tabanlı teknikte verileri, bölge odaklı sorgulamakta daha kesin fakat gayri resmi olarak tanımlamaktadır. Yoğunluk, toplam alan vb. belirli koşulları sağlayan bölgelerin seçimini talep etmektedirler. Bu yöntemin avantajı bağımsız işlemlerin hızla gerçekleştirilmesi ve bağımsız kümelerin oluşmasıdır (Wang H. ve Yu Z., 1997: 186).

3.4.1.5. Model Tabanlı Teknikler

Model tabanlı yöntemlerde, her küme için ayrı model oluşturulmaktadır. Veri için en uygun olan model bulunmaktadır. Veriler uzaydaki dağılım durumlarında da modelleme gerçekleştirebilmektedir. Bir takım yöntemlerle küme sayısı otomatik olarak bulunabilmektedir. Bu işlem esasında gürültü ve uç noktalar dikkate alınarak gerçekleştirilmektedir (Albayrak, 2009: 14).

3.4.2. Sınıflandırma Teknikleri

Veri madenciliğinde kullanılan en temel uygulama olan sınıflandırma tekniği, kategorik olarak oluşturulmuş verilerin sonuçlarını tahmin etmek amacıyla kullanılmaktadır. Sınıflandırma yönteminde, tanımlanan değer yeni bir sınıf içinde yer almaktadır. Bu değerler "eğitim verisi" olarak adlandırılmaktadır. Bu işlemin gerçekleştirilmesi için öğenin özelliklerinin iyi bilinmesi gerekmektedir (Ergüden ve Erşahin, 2018: 37). Sınıflandırma modeli üç aşamada uygulanmaktadır (Bayer, 2011: 26):

- Birinci aşamada; her öğenin sınıfı tespit edilmektedir. Veri oluşumunda kullanılan kümeye öğrenme kümesi denilmektedir. Eğer öğelerin özellikleri

bilinmiyorsa bu öğelere denetimsiz öğrenme denilmektedir. Sınıf etiketlerinin bilinmesi ise denetimli öğrenmeyi oluşturmaktadır.

- İkinci aşamada; veriler uygulamaya alınır ve test aracı rastgele bulunmaktadır. Öğrenme kümesi ve sınıf etiketi karşılaştırılmaktadır. Modelin doğruluğu sınıflandırılmış test kümesi ile toplam test kümesinin verilerinin oranları ile oluşturulmaktadır.
- Üçüncü aşamada; kullanımından sonra bilinmeyen "veri sınıf etiketi" tahmini gerçekleştirilmektedir.

3.4.2.1. Diskriminant Analizi

Eski bir teknik olan Diskriminant analizi ilk işleyişini 1936 yılında gerçekleştirmiştir. Diskriminant analizi tıp, biyoloji ve sosyal bilimler alanında sıklıkla kullanılan bir tekniktir. Verileri önceden belirlenen sınıflara atarak öğrenme kümesi oluşturulmaktadır. Öğrenme kümesine bağlı olarak doğrusal fonksiyonel kümeler oluşturulmaktadır. Yeni gözlemlerde ise tüm fonksiyonlar hesaplanarak yeni gözlem en yüksek sınıfa dahil edilmektedir (Edelstein, 1999; akt. Aytaş, 2013: 29).

3.4.2.2. Naive Bayes

Naive Bayes analizi, olasılık hesaplama temelinde hareket etmektedir. Daha önceden belirlenen alanlara yerleştirilmiş verilerin kullanılması sonucunda yeni verinin hangi alana dahil edilmesi gerektiğini hesaplayan yöntemdir. Bu yöntem ile belirsizlik taşıyan durumun modeli oluşturularak belirli sonuçlara ulaşım sağlanmaktadır (Ergüden ve Erşahin, 2008: 52). Naives Bayes tekniği ile metin belgelerinin sınıflandırmasında, sağlık alanında yoğun olarak kullanılmaktadır (Bayer, 2011: 36).

3.4.2.3. Karar Ağaçları

Karar ağacı modeli, sınıfları bilinen örnek verilerden tümevarım şeklinde öğrenilen karar yapısıdır. Bu karar yapısı ağaç şeklinde oluşmaktadır. Karar ağacı

yöntemi ile karar verme yönergeleri eşliğinde büyük kayıtlar küçük kayıt gruplarına bölünmektedir (Sun ve Li, 2008: 2).

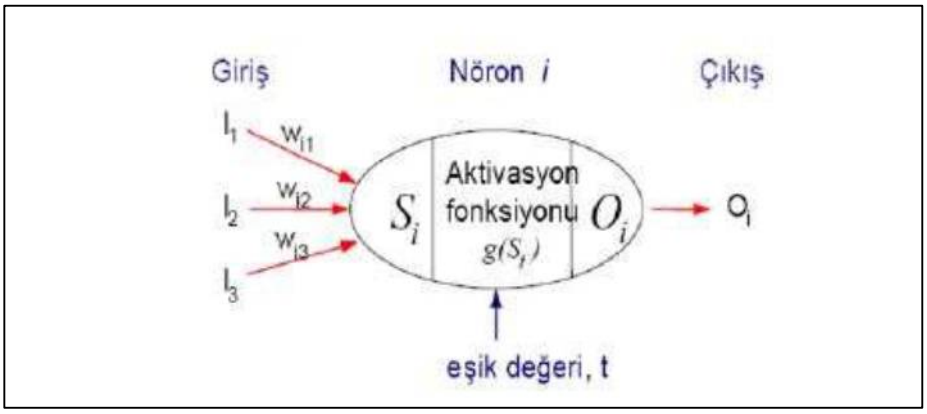
Karar ağacı analizinin yoğun olarak kullanıldığı alanlar aşağıda yer almaktadır (Albayrak ve Yılmaz, 2009: 40):

- Belli bir grubun üyelerinin arasına dahil olacak kişilerin belirlenmesinde,
- Olayların yüksek, orta, düşük risk grupları şeklinde kategorize edilmesinde,
- Parametrik modellerin oluşturulmasında kullanılacak değişkenlerin en önemlisinin tespit edilmesinde,
- Oluşabilecek olayların önceden tahmin edilmesinde,
- Belirli alt grupların kendilerine has özelliklerin ifade edilmesinde,
- Kategorilerin bütün hale getirilmesi ve sürekli olan değişkenlerin kesikli değişken yapıya dönüştürülmesinde kullanılmaktadır.

3.4.2.4. Yapay Sinir Ağları

Yapay sinir ağları, bilgi işleme sistemlerinin biyolojik sinir ağları şeklinde dizilimine verilen isimdir. Yapay sinir ağları, eski bir bilgi işleme sistemi olup 1942 yılında McCulloch ve Pitt tarafından temelleri atılmıştır. McCulloch ve Pitt hücre ağlarını geliştirerek yapay sinir ağlarının başlangıcını oluşturmuşlardır (Keskenler ve Keskender, 2014: 11). Yapay sinir ağları, sinir hücrelerinin birleşimi ve katmanlar dâhilinde dizilimi şeklinde oluşturulmuştur. Yapay sinir ağları, bilgileri hesaplama ve işleme özelliğini paralel dağılmış ağlarından ve öğrenebilme, genelleme gibi yeteneklerinden almaktadır. Bu özellikleri ile yapay sinir ağları karmaşık yapıdaki problemleri kolayca çözebilmektedir (Şekeroğlu, 2010: 51).

Şekil 7: Yapay Sinir Ağları



Kaynak: Şekeroğlu, 2010: 52.

Yapay sinir ağları şu özellikleri sebebiyle etkin kullanım sağlamaktadır (Şekeroğlu, 2010: 54: Yavuz, 2012: 7):

- Yapay sinir ağları doğrusal bir yapıya sahip değildir. Bu sebeple bütün ağlara yayılım gerçekleştirmektedir. Karmaşık yapıdaki problemlerin çözümü ise bu özelliği sebebiyle daha da kolay bir hal almaktadır.
- Hücreler arasında doğru bağlantılar kurulduğunda bağlantılara uygun şekilde istenilen davranışı gösterebilmektedir.
- Yapay sinir ağları dahil olduğu problemi öğrendikten sonra verilen eğitim esnasında karşılaşmadığı örneklerine de istenilen tepkiyi gösterebilmektedir.
- İlgilendiği problemde meydana gelen değişikliklere göre kendisini ayarlamaktadır.
- Yapay sinir ağları problemin değişimine göre ağırlıklarını değiştirebilmektedir. Bu nedenle problemdeki değişiklikleri algılayabilir, değişimlere göre tekrar eğitilebilmektedir.
- Yapay sinir ağları birbiri ile bağlantılı paralel ağlara sahiptir. Her bağda bilginin bulunması ve bütün ağlara dağıtılmış olması bazı hataları meydana getirmektedir. Bu hatalar bazı hücreleri etkisiz bırakabilmektedir. Bu hata doğruyu risk boyutunda etkilememektedir. Hatayı tolere etme yeteneğine sahiptir.

3.4.2.5. Kaba Kümeler

Kaba kümeler, klasik küme yapısının dışında, kümelerin tanımlanabilmesi için elemanların bilgilerine ulaşılması gerektiğini savunan yaklaşımdır. Nesnelerin aynı şekilde tabirlendirilmesi nesnelerin aynı veya benzer olduğunu göstermektedir. Nesnelerin ayırt edilememesi kısmında nesneler kümelendirilmektedir. Bu küme şekli bilginin temelini oluşturduğu için elemanter küme adı verilmektedir. Elemanter kümelerin her birinin isimlendirilmemesi kaba kümeleri meydana getirmektedir. Her bir adlandırılmayan kaba küme, kendinin veyahut tümleyenin elemanları olarak sınıflandırılmaktadır (Binay, 2002: 23).

Kaba kümeler yaklaşımında muğlak kavram kesin kavram ile karakterize edilerek alt ve üst yaklaşımlar meydana getirilmektedir. Alt yaklaşım, kavram dahilindeki tüm unsurların içinde barındırmaktadır. Üst yaklaşım, kavrama yakın olan, kavrama ait olduğu düşünülen tüm öğeleri dahil etmektedir. Bu iki yaklaşım arasındaki farklar sınır alanlarını meydana getirmektedir (Pawlak ve Skowron, 1994: 72; akt. Aydoğan ve Gencer, 2007).

3.4.2.6. Genetik Algoritmalar

Genetik Algoritmalar, evrimsel hesaplama metotları içerisinde yer almaktadır. Genetik algoritmaların temelinde biyoloji bulunmaktadır. Genetik algoritmalar evrim yasalarından esinlenerek oluşturulmuştur. Veri madenciliğinde kümeleme ve öngörme problemlerinde kullanılmaktadır. Verilere en uygun olan modelin tespitinde kullanılmaktadır. Genetik algoritmalar modellerin tanımlamasına ve uygunluğuna bağlı olarak farklılık göstermektedir. Genetik algoritmalar hem değişik yapıdaki verileri işleyebilmekte ve yapay sinir ağları ile çalışma gerçekleştirebilmektedir. Kısa sürede uygun sonuçlar alabilmek amacıyla genetik algoritmalara başvurulmaktadır (Dunham, 2003).

Genetik algoritmaların işlem adımları ise şu şekilde sıralanmaktadır (Çalışkan, Yüksel ve Dayık, 2016: 22):

- Arama ağındaki bileşenler dizi şeklinde kodlanmaktadır. Bu kodlama durumunda uygunluk fonksiyonu oluşturulur. Sorunların iyileştirilmesinde yeni çözümler ortaya koymaktadır.
- İstenilen bireyler olasılık unsurlarına göre rastlantısal çoğalma işlemine sokulmaktadır.
- İstenilen bireyler çaprazlama, mutasyon işlemine sokulmaktadır.
- Hata oranı düşürülene kadar bu adımlar tekrar uygulanmaktadır.
- Adım sayısı, istenilen kuşak adedine kadar devam eder. Kuşak sayısına ulaşıldığında ise işlem sonlandırılmaktadır.

Şekil 8: Genetik algoritmanın işlem basamakları

Kaynak: Badem, 2007: 12.

3.4.2.7. Regresyon Analizi

Sıklıkla kullanılan veri madenciliğinde yapılan modelleme uygulamasıdır. Bağımlı bağımsız değişkenler arasındaki ilişkilerin açıklanmasıdır (Hosmer ve Lemeshow, 2000: 33). İlişkinin matematiksel değerler şeklinde oluşturulması ve açığa çıkan denkleme regresyon denklemi de denilmektedir. Regresyon denklemi ve matematiksel olarak açıklanması ise şu şekilde yapılmaktadır: Bağımsız

değişken(ler)in katsayıları olan parametreler ($\theta 1, \ldots, \theta n$), bağımsız değişken(ler)in bağımlı değişken üzerindeki etkisinin bir ölçüsünü göstermek ve bağımlı değişkenin beklenen ile tahmin edilen değeri arasındaki fark hata terimi (e) olmak üzere, $y = F(x, \theta) + e$ eşitliğini sağlayan en iyi F fonksiyonu için θ değerlerinin araştırılması sürecidir (Şık, 2014: 49).

3.4.3. Birliktelik Kuralları Teknikleri

Birliktelik kuralları, veri madenciliğinde en sık kullanılan tekniktir. Birliktelik kuralları; bilinen grup ile aynı özelliği taşıyan fakat farklı grupta yer alan verilerin tespitinde ve aralarındaki bağın kurulumunda kullanılmaktadır. Birliktelik kuralı ile verilerin özellikleri ortaya koyulurken, aynı zamanda temel veriler arasındaki istatistiksel korelasyonun ifadesidir. Birliktelik kuralları 1990'lı yıllarda marketlerde ürün ve müşteri tanıma sistemleri, satış hareketleri, verilerin bilgisayarlara aktarılmasını gerektiren işlemlerde, müşteri-ürün ilişkisinin oluşturulmasında kullanılmaktadır. Birliktelik kuralı sayesinde ürünler arasında birliktelik sağlanan müşteride diğer ürünü de alma isteği doğmaktadır (Şen, 2008: 24). Bir veri setindeki bazı özelliklerin bir kümesi $x = \{x_1, \ldots, x_n\}$ olmak üzere bu özellikler kullanılarak bir y kuralının c güven ve s doğruluk seviyesinde öngörülmesi kısaca $x = \{x_1, \ldots, x_n\} \Rightarrow y[c, s]$ şeklinde gösterilmektedir (Şık, 2014: 57).

DÖRDÜNCÜ BÖLÜM
KÜMELEME ANALİZİNDE K-ORTALAMALAR ve İKİ AŞAMALI ALGORİTMA YÖNTEMLERİ

4.1. K-ORTALAMALAR TANIMI

k-ortalamalar algoritması, giriş uzayının "k" adet merkez ile ifade edilmesini sağlayan bir yöntem olarak tanımlanmaktadır (İşler ve Narin, 2012: 26). k-ortalamalar kümeleme analizi, MacQueen tarafından 1967 yılında geliştirilen ve bir veri kümesini k adet gruba ayırmak için yaygın biçimde kullanılan bir yöntemdir (Wagstaff, Cardie, Rogers ve Schroedl, 2001: 577).

k-ortalamalar yöntemi, kümeleme problemlerinde en basit denetimsiz öğrenme tekniği olarak nitelendirilmektedir. Algoritmada "n" adet nesneden meydana gelen veri kümesini (X), giriş değişkeni olarak verilen k ($k \leq n$) adet kümeye bölmek esas mantıktır. Burada amaç, yapılan bölümleme sonucunda oluşan kümelerin; küme içindeki benzerliklerinin maksimum, kümeler arasındaki benzerliklerinin minimum olmasını sağlamaktır. K-ortalamalar yönteminin performansı "k" küme sayısına, başlangıç küme merkez değerlerine ve benzerlik ölçüm kriterine bağlıdır (Çalışkan ve Soğukpınar, 2008: 121).

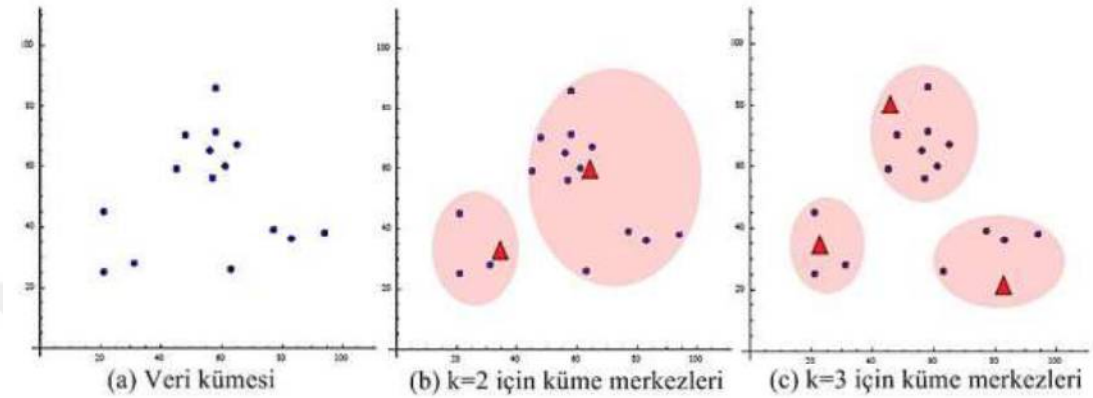
Algoritma, k ortalamalarının bulunması temeline dayanarak bir ortalama değerler seti ile başlamakta ve noktaların merkeze uzaklıkları doğrultusunda sınıflandırma yapılmaktadır. Ardından kümeye atanan gözlemler kullanılarak küme ortalamaları yeniden hesaplanmaktadır. Daha sonra tüm gözlemler, yeni ortalamalar kümesine göre yeniden sınıflandırılmaktadır. Küme ortalamaları, ardışık işlem aşamaları arasında çok fazla değişiklik yapmayıncaya kadar bu adım tekrarlanmaya devam edilmektedir. Son olarak kümelerin ortalamaları bir kez daha hesaplanarak gözlemler kalıcı kümelere atanmaktadır (Norusis, 2011: 374).

4.2. K ORTALAMALAR YÖNTEMİNDE VERİ GÖSTERİMİ

k ortalamalar algoritmasında, başlangıçta seçilen k sayısına ilişkin örnek veri dağılımları Şekil 9'da gösterilmiştir. Bir veri setinde k sayısının 2 seçilmesi

durumunda kümeleme sonucunda elde edilecek kümeler, bu kümelerin merkezleri üçgenlerle temsil edilerek gösterilmiştir. k sayısının 3 seçilmesi durumunda ise ortaya çıkan durum da k sayısının kümeleme işlemini nasıl değiştirdiğini göstermektedir (Özkan, 2013: 6).

Şekil 9: k Sayısı Değişimine Göre Verilerin Gösterimi



k ortalamalar yönteminin sadeliği ve hızına rağmen bazı dezavantajları bulunmaktadır. Bunlardan en belirgin olanı, çözümün başlangıç merkezlerinin seçimine bağlı olmasıdır. Başlangıç merkezlerinin uygun seçilmemesi, yerel minimum değerinin zayıf olmasına yol açabilir. Bu nedenle, başlangıç problemini atlatmak için genellikle birden fazla rastgele yeniden başlatma gerçekleştirilir. Genellikle yeniden başlatma ile geri dönen çözümler, özellikle geniş bir araştırma alanına (örneğin birçok kümelenme ve boyut) sahip problemler için, iyi olandan çok kötü olanlara doğru ulaşılan amaç değeri açısından önemli ölçüde farklılık göstermektedir. Bu nedenle, iyi bir yerel minimum bulma olasılığını arttırmak için algoritmanın birçok defa çalışması gereklidir (Tzortzis ve Likas, 2014: 2506).

4.3. K-ORTALAMALAR YÖNTEMİNDE MATEMATİKSEL GÖSTERİM

k-ortalamalar yöntemi; C sayı kümelerini, veri seti X ve N sayı özellik vektörüne ait D sayı değişkenlerinden ayırarak sınıflandırmaktadır. Bu yöntemde C sayı kümesi merkezlerinin belirlenmesi ile işlem başlatılmakta ve veri kümesinin her bir üyesi benzerlik ölçütü ile en yakın kümeye atanmaktadır. Atama sonrası kümelerin merkezleri yeniden hesaplanmakta ve bazı küme üyeleri başka bir kümeye

atanabilmektedir. Bu işlem, değişim sonlanana kadar tekrarlanmaktadır. Burada, N vektör sayısı, d değişken sayısı olmak üzere, bir probleme ait X veri seti aşağıdaki gibi tanımlanabilir.

$X=\{X_k|k=1,2,\dots N\}$

Bu veri kümesinde, k. özellik vektörü şu şekilde yazılabilir:

$X_k=\{X_{k1},X_{k2}\dots\dots\dots X_{kd}\}, X_k \in R^d$

k-ortalamalar yönteminde veri seti alt kümelere ayrıldığında, aşağıdaki denklikte verilen amaç fonksiyonunun minimizasyonu hedeflenmektedir. Bu denklemin minimuma indirgenmesi, her bir özellik vektörü ile en yakın küme merkezi arasındaki uzaklığın minimuma indirilmesini gerektirmektedir. Böylelikle aynı kümede toplanan benzer yapılandırılmış verilerin sağlanması gerçekleştirilir (Fırat, Dikbaş, Koç ve Güngör, 2010: 1610).

$$J(S:X)=\sum_{i=1}^{C}\sum_{k=1}^{N} d_{ik}^2(x_k,s_i)$$

k ortalamalar algoritmasında her nesnenin merkeze uzaklıklarının hesaplanması amacıyla dört farklı formül kullanılmaktadır. Bu formüller Öklid uzaklığı-Öklid uzaklığının karesi, City-block (Manhattan) uzaklık formülü ve Chebychev uzaklığı formülüdür (Demiralay ve Çamurcu, 2005: 5; Rao ve Srivinas, 2006: 39). Yukarıdaki denklemde uzaklık ölçümü için Öklid uzaklığı kullanıldığında;

$$\text{uzaklık}(x,y)=\left\{\sum_i (x_i-y_i)^2\right\}^{1/2}$$

$$d_{ik}^2=\left\|x_k^{(i)}-s_i\right\|^2$$

$s_i=\frac{1}{N}\sum_{k=1}^{N} x_k^{(i)}$ formülü oluşur.

Sonraki aşamada Öklid uzaklık matrisinin matematiksel gösterimi yer almaktadır.

$$\begin{bmatrix} d(1,1) & \cdots & d(1,C) \\ \vdots & \ddots & \vdots \\ d(N,1) & \cdots & d(N,C) \end{bmatrix}$$

Belirtilen matriste C küme sayısını, S küme merkezlerini içeren matrisi ($S=\{s_1,s_2\dots s_C\}$), s_i i. kümenin merkezini, $x_k^{(i)}$ i. kümenin k. özellik vektörünü, d_{ik}^2 i.

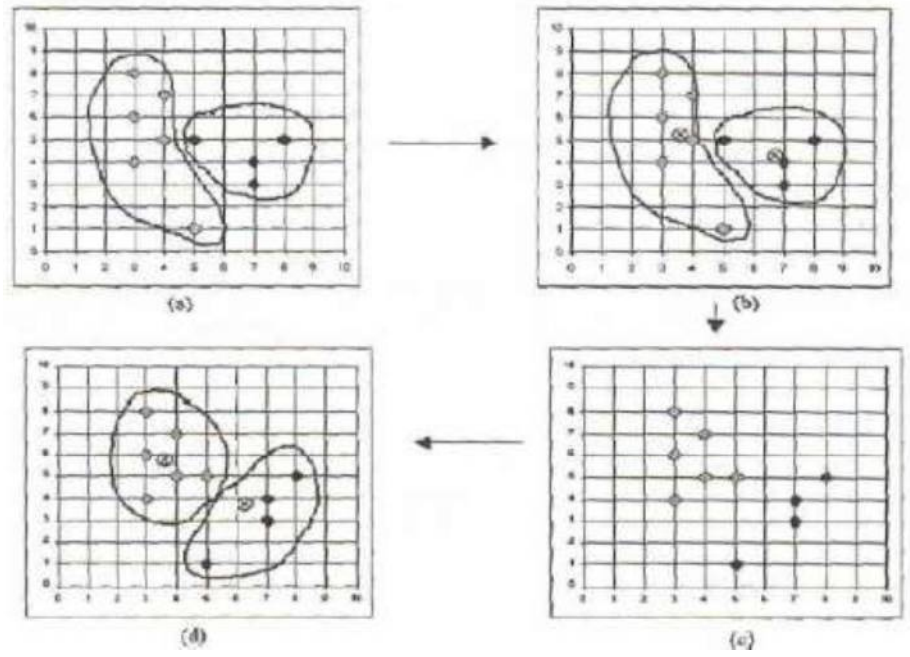
kümenin k. özellik vektörü ile i. kümenin merkezi arasındaki uzaklığı ifade etmektedir (Fırat, Dikbaş, Koç ve Güngör, 2010: 1611).

4.4. K ORTALAMALAR YÖNTEMİNİN AŞAMALARI

K-ortalamalar algoritmasında bulunan işlem aşamaları şu şekilde sıralanmaktadır (Altınışık ve Yıldırım, 2013: 12; Bora, Gupta ve Kahn, 2015: 197):

- Öncelikle iki farklı yöntemle başlangıç küme merkezleri tespit edilmektedir. Bu yöntemlerden birincisinde, nesnelerin içerisinden küme sayısını ifade eden "k" tane rastgele nokta seçilmektedir. İkinci yöntemde merkez noktalar, tüm nesnelerin ortalamasının alınması ile belirlenmektedir.
- İkinci aşamada seçilen küme merkezi ile tüm nesneler arasındaki uzaklıklar hesaplanmaktadır. Ulaşılan uzaklık değerleri doğrultusunda tüm nesneler, "k" tane küme arasından kendisine en yakın olan kümeye yerleştirilmektedir.
- Üçüncü aşamada, oluşturulan kümelerin yeni merkezleri o küme içindeki tüm nesnelerin ortalamaları ile değiştirilmektedir.
- Son aşamada kümelerin merkez noktaları sabit olana kadar ikinci ve üçüncü aşamalar tekrar edilmektedir.

Şekil 10: k Ortalamaları Kümeleme Aşamaları

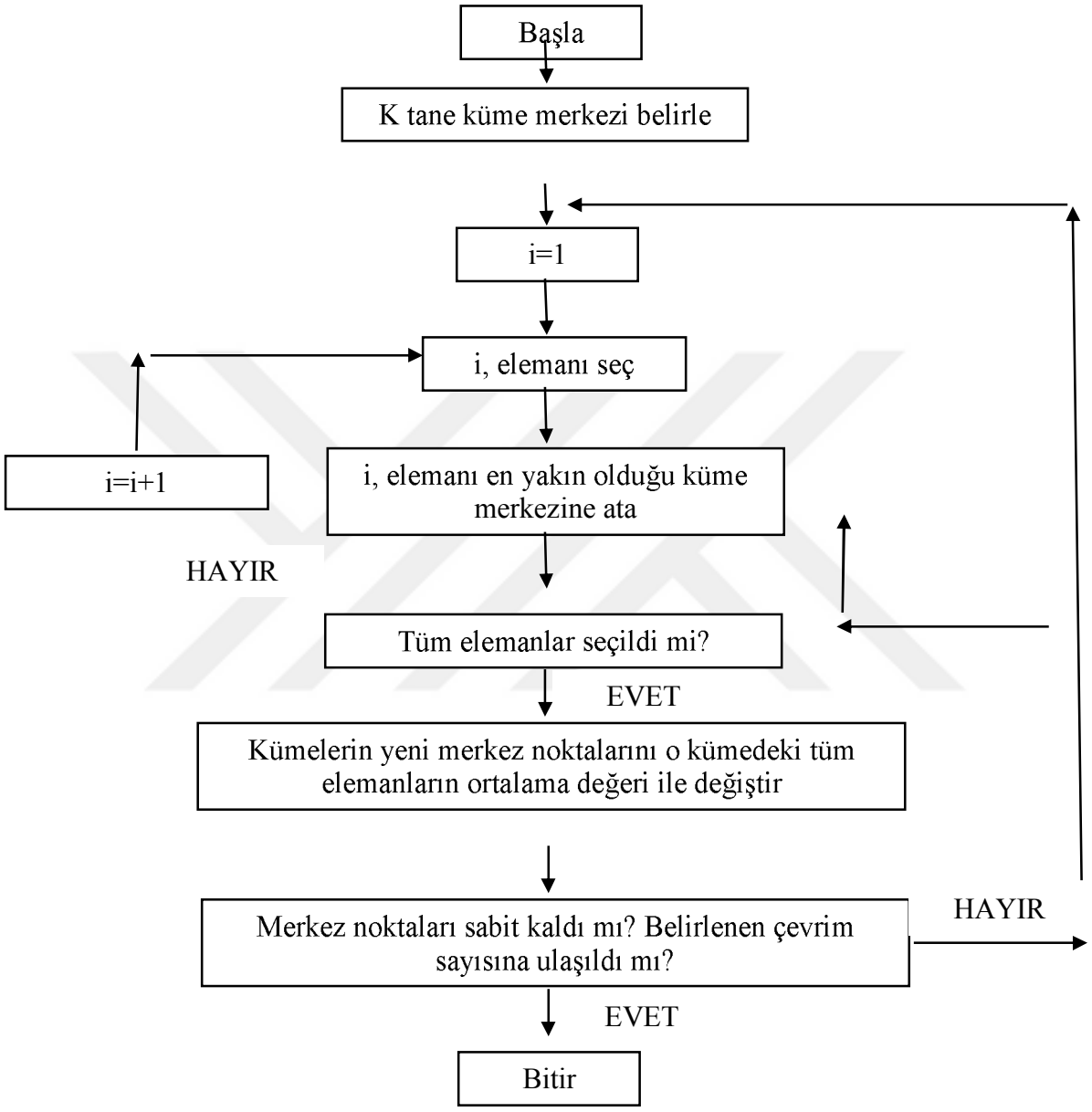


Merkezlere ilk değerlerin rastgele atanmasının ardından merkez değerlerinin güncellenmesi için kullanılan yöntemler Batch metodu ve online metodudur. Batch metodunda giriş kümesine ilişkin her nesnenin hangi merkeze yakın olduğu hesaplanmaktadır. Aynı merkeze yakın nesnelerin ortalamaları alındığında, merkeze ait değer güncellenmekte ve durma koşulu sağlanana kadar işlem sürdürülmektedir. İkinci yöntem olan online metodu ise, giriş kümesi içerisinden bir nesne seçerek bu nesnenin merkezlere uzaklığını hesaplamaktadır. Nesnenin en yakınındaki merkez tespit edilerek merkez değeri güncellenmekte ve tüm nesneler için bu işlem gerçekleştirilmektedir. Merkez değerinin güncellenmesi, nesne ile merkez arasındaki uzaklığın her adımda azalan bir öğrenme katsayısıyla çarpılmasıyla yapılır. Böylece ilk güncellemede merkezin yer değişimi büyük, güncellemelerin sayısı arttıkça yer değişimi küçük olmaktadır. Diğer bir ifade ile merkezler yakınsamaktadır. İlk metoda benzer şekilde durma koşulu sağlanana kadar her nesne için işlem tekrarlanmaktadır (İşler ve Narin, 2012: 26).

4.5. K-ORTALAMALAR ALGORİTMASI

K ortalama algoritmasına yönelik akış şeması Şekil 11'de gösterilmiştir.

Şekil 11: k ortalama yöntemine ilişkin algoritma akış şeması



Algoritma "k" adet küme merkezinin belirlenmesiyle başlamakta ve veri setinden "i" elemanının seçilmesiyle devam etmektedir. Elemanın en yakın olduğu merkeze atanmasının ardından bu işlem tüm elemanlar için gerçekleştirilmektedir. Eğer tüm elemanlar seçilmediyse algoritma "i" elemanının seçilmesi aşamasına geri dönmektedir. Tüm elemanlar seçildikten sonra kümelerin yeni merkez noktaları,

kümedeki tüm eleman ortalaması ile değiştirilir. Bu işlemden sonra merkez noktaların sabit kalmaması durumunda algoritma ilk aşamaya geri dönmektedir. Eğer merkez noktalar sabit kalırsa algoritma tamamlanmaktadır (Özkan, 2013: 5).

4.6. İKİ AŞAMALI KÜMELEME ANALİZİ TANIMI

İki Aşamalı Kümeleme Analizi, küme sayısının bilinmediği durumda hem sürekli hem de kategorik verileri içeren veri setlerine uygulanabilmektedir. K-ortalamalar tekniği gibi geleneksel kümeleme analizlerinde büyük veri setlerinin oluşturduğu problemler iki aşamalı teknik ile giderilebilmektedir. Geleneksel yöntemlerin uygun sonuçlar vermesi yalnızca büyük verilerin küçük gruplara bölünmesi ile gerçekleştirilirken, İki Aşamalı yöntemde veri setlerinin büyük olması, tüm gözlemlerin eş zamanlı analiz edilmesine engel oluşturmamaktadır (Giray, 2016: 16).

İki Aşamalı Kümeleme Analizi, Punj ve Steward (1983) tarafından geliştirilmiş ve hiyerarşik özellikte olmayan bir kümeleme analizidir. Bu analiz, Ward'ın Minimum Varyans Yöntemi ile "K-ortalamalar" yöntemlerinin kombinasyonundan meydana gelen hibrid bir kümeleme analizidir (Özdemir ve Orçanlı, 2012: 12).

4.7. İKİ AŞAMALI YÖNTEMDE MATEMATİKSEL GÖSTERİM

İki Aşamalı Kümeleme analizinde gereksinim duyulan uzaklık ölçüleri kapsamında, bir ya da daha fazla kategorik değişkenin mevcut olması durumunda log-olabilirlik ölçüsü kullanılmaktadır. Bu ölçü kullanılırken değişkenlerin birbirlerinden bağımsız oldukları kabul edilmektedir. Tüm değişkenlerin sürekli olduğu durumlarda kullanılan Öklid uzaklığı ölçüsü ise, gözlemlerin en küçük Öklid uzaklığına sahip kümede gruplanması sağlanmaktadır (Ceylan, Gürsev ve Bulkan, 2017: 478).

Log-olabilirlik uzaklığını hesaplamak için, sürekli değişkenlerin normal dağılımlara sahip olduğu ve kategorik değişkenlerin multinomial dağılımları olduğu ve değişkenlerin birbirinden bağımsız olduğu varsayılmaktadır. i ve j kümeleri

arasındaki uzaklık aşağıdaki formüllerle belirlenmektedir (Bacher, Wenzig ve Volger, 2004: 4; Şchiopu, 2010: 68).

$$d(i,j) = \xi_i + \xi_j - \xi_{<i,j>}$$

$$\xi_s = -N_s\left(\sum_{k=1}^{K^A}\frac{1}{2}\log(\hat{\sigma}_k^2 + \hat{\sigma}_{sk}^2) + \sum_{k=1}^{K^B}\hat{E}_{sk}\right)$$

$$\hat{E}_{sk} = -\sum_{l=1}^{L_k}\frac{N_{skl}}{N_s}\log\frac{N_{skl}}{N_s}$$

Yukarıdaki formüllerde d (i, j) i ve j kümeleri arasındaki uzaklığı, <i, j> i ve j kümelerinin kombinasyonu ile oluşan kümeyi temsil eden indeksi, K^A sürekli değişkenlerin toplam sayısını, K^B kategorik değişkenlerin toplam sayısını, L_k k. kategorik değişken için kategorilerin sayısını, N_s s kümesindeki veri kayıtlarının toplam sayısını, N_{skl} k kategorik değişkeninin l kategorisini aldığı s kümesindeki kayıtların sayısını, N_{kl} l kategorisini alan k kategorik değişkenindeki kayıtların sayısını, $\hat{\sigma}_k^2$ tüm veri seti için k sürekli değişkeninin tahmini varyansını, $\hat{\sigma}_{sk}^2$ ise j kümesindeki k sürekli değişkeninin tahmini varyansını ifade etmektedir (Şchiopu, 2010: 69).

Kümelerin sayısını otomatik olarak belirlemek amacıyla iki aşama kullanılmaktadır. Birinci aşamada BIC (Schwarz'ın Bayesçi Bilgi Ölçütü) veya AIC (Akaike'nin Bilgi Ölçütü) göstergesi, belirtilen aralıktaki her bir küme sayısı için hesaplanmaktadır. Ardından bu gösterge, kümelerin sayısı için bir başlangıç tahmini bulmak için kullanılmaktadır (Şchiopu, 2010: 69).

$$AIC\ (J) = -2\sum_{j=1}^{J}\varepsilon j + 2mj$$

$$BIC\ (J) = -2\sum_{j=1}^{J}\varepsilon j + mj\log(N)$$

$$m_j = J\left\{2K^A + \sum_{k=1}^{K^B}L^K - 1\right\}$$

Denklemlerde j kümelerin sayısını, k sürekli değişken kümelerin toplam sayısını, N ise toplam gözlem sayısını temsil etmektedir. Son denklemde ise K^A toplam sürekli değişken sayısını, K^Btoplam kategorik değişken sayısını, L^Kise k

kategorik değişkenindeki kategori sayısını belirtmektedir (Ceylan, Gürsev ve Bulkan, 2017: 479).

Gözlemlerin küme içerisine atanması, aykırı değerlerin dönüşümü yoksa, gözlemleri en yakın kümeye atayarak yapılmaktadır. Aykırı bir değer varsa log-olabilirlik uzaklığı kullanılmaktadır. Burada kümeden uzaklık, $C = \log(V)$, $V = (\prod_k R_k) * (\prod_m L_m)$ olarak formüle edilen kritik değerden küçükse en yakın kümeye atama yapılırken, diğer durumlarda gözlem aykırı olarak sınıflandırılmaktadır. Eğer Öklid uzaklığı $C = 2.\sqrt{\frac{\sum_{l=1}^{K^A} \sigma_{k,l}^2}{K^A}}$ olarak formüle edilen kritik değerden küçükse, diğer durumlarda gözlem aykırı olarak sınıflandırılmaktadır (Trpkova ve Tevdovski, 2009: 306).

4.8. İKİ AŞAMALI YÖNTEMİN AŞAMALARI

İki Aşamalı Kümeleme Analizinde uygulanan aşamalar ön kümeleme, aykırı değerlerin belirlenmesi ve kümelerin oluşturulması olarak sıralanmaktadır (Arıkan, 2017: 52). Bazı kaynaklarda ise İki Aşamalı Yöntem ön kümeleme ve kümeleme olmak üzere iki grupta incelenmektedir (Li ve Sun, 2018: 3).

Ön kümeleme aşamasında tüm verilerdeki durumlar taranmakta ve bazı eşik uzaklık kriterleri temelinde ön kümeler oluşturulup oluşturulamayacağını belirlemek amacıyla nesneler arasındaki log-olabilirlik uzaklığı ölçülmektedir (Li ve Sun, 2018: 3).

Ön kümeleme aşamasının amacı, uzaklık matrisinin boyutunu azaltmaktır. Bu amaç doğrultusunda, düğümlerden (node) ve yapraklardan oluşan "Küme Özellik Ağacı" kullanılmaktadır. Her gözlem, taşıdığı değer doğrultusunda kök düğümden başlamakta ve alt düğümlerden kendine en yakın değeri içeren yapraklara atanmaktadır. Gözlemin atandığı yaprak, küme merkezinden farklı yapıya sahipse yeni yapraklar oluşturularak ağaç tamamlanmaktadır. Diğer bir ifadeyle her gözlem, uzaklık ölçüsü temelinde önceden oluşturulan ön kümelere atanmakta veya gözlem yeni bir küme meydana getirmektedir. Sürekli değişkenlerde uzaklık matrisleri, iki küme merkezi arasındaki uzaklığı veren Kareli Öklid Uzaklığı kullanmaktadır (Arıkan, 2017: 52).

Küme Özellik Ağacı oluşturma sürecinde algoritma, atipik değerlerin (aykırı değerlerin) çözülmesini mümkün kılan, isteğe bağlı bir adımı uygulamaya koymaktadır. Aykırı değerler, herhangi bir kümeye uygun olarak sığmayan gözlemler olarak kabul edilmektedir. SPSS'te bir yapraktaki gözlemler, gözlem sayısının Küme Özellik Ağacındaki en büyük yaprak girişinin belirli bir yüzdesinden daha az olduğunda aykırı sayılmakta; bu yüzde 25 olarak varsayılmaktadır. Küme Özellik Ağacını yeniden oluşturmadan önce, prosedür doğrultusunda potansiyel aykırı değerler aranmakta ve sistem bunları bir kenara koymaktadır. Küme Özellik Ağacı yeniden oluşturulduktan sonra prosedür, ağaç boyutlarını arttırmadan bu değerlerin ağaçta olup olmadığını kontrol etmektedir. Son olarak, herhangi bir yere uymayan değerler aykırı sayılmaktadır. Küme Özellik Ağacı, mümkün olan maksimum boyutu aşarsa, eşik uzaklığını artırarak mevcut Küme Özellik Ağacına göre yeniden oluşturulmaktadır. Yeni oluşturulan Küme Özellik Ağacı daha küçüktür ve yeni gözlem girişlerini mümkün kılmaktadır (Şchiopu, 2010: 67-68).

Kümeleme aşamasında, ön kümeleme aşamasında meydana gelen alt kümeler, kümeleme algoritması kullanılarak optimal sayıda kümeye bölümlenmektedir. Alt kümeler, tek bir küme olarak kabul edilmekte ve tüm veriler tek bir küme içine yerleştirilene kadar minimum uzaklığı karşılayarak bir küme içerisinde birleştirilmektedir. İstenen sayıda küme önceden belirlenmemişse, algoritma Schwarz'ın Bayesçi Kriteri (BIC) veya Akaike Bilgi Ölçütü (AIC) kullanarak en uygun küme sayısını otomatik olarak belirleyebilir (Li ve Sun, 2018: 3).

$$AIC\ (J) = -2\sum_{j=1}^{J} \varepsilon j + 2mj$$

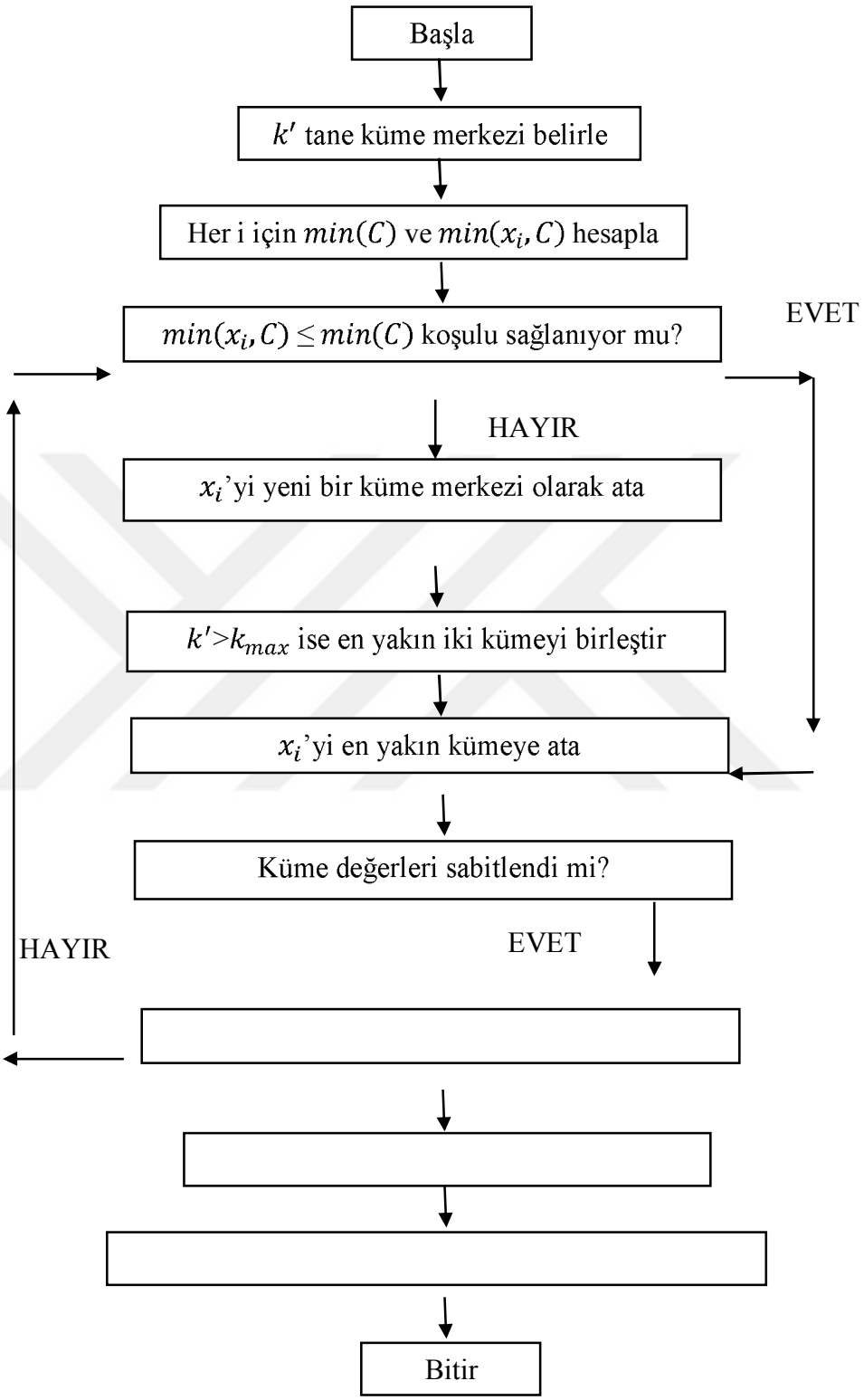
$$BIC\ (J) = -2\sum_{j=1}^{J} \varepsilon j + mj\log(N)$$

Denklemlerde j kümelerin sayısını, k sürekli değişkenlerin kümelerinin toplam sayısını, N ise toplam gözlem sayısını temsil etmektedir ($mj = 2KJ$).

4.9. İKİ AŞAMALI YÖNTEM ALGORİTMASI

İki Aşamalı Kümeleme Analizinde küme algoritması temel olarak modifiye k ortalamalar aşaması ve aykırı değer bulma aşaması olarak ikiye ayrılmaktadır. Algoritmanın ilk aşamasının birinci basamağı, k' başlangıç değerlerinin rastgele küme merkezi olarak seçilmesi ile başlamaktadır. İkinci basamakta her i için minimum C ve (x_i, C) hesaplanmaktadır. Eğer $min(x_i, C) \leq min(C)$ koşulu sağlanırsa altıncı basamağa atlanmaktadır. Dört, beş ve altıncı basamaklarda sırasıyla x_i'nin yeni bir küme merkezi olarak atanması, k'>k_{max} ise en yakın iki kümenin birleştirilmesi ve x_i'nin en yakın kümeye atanması gerçekleştirilmektedir. Yedinci basamakta ise küme değerleri stabilize olana kadar ikinci aşamaya geri dönülmektedir. Algoritmanın ikinci aşamasının ilk basamağı, bir F evreninde C merkezlerini veri setini içeren Küme Özellik Ağacının oluşturulması ile başlamaktadır. İkinci basamakta F'de yer alan ağaçtan en uzun kenarı kaldırarak yeni alınan iki alt ağaçla değiştirilmektedir. Üçüncü basamakta ise küçük alt ağaçtaki kümeler, aykırı olarak kabul edilmektedir (Jiang, Tseng ve Su, 2001: 695-696). Açıklanan basamakları içeren algoritma Şekil 12'de yer almaktadır.

Şekil 12: İki aşamalı Yönteme İlişkin Algoritma Akış Şeması



BEŞİNCİ BÖLÜM
UYGULAMA

5.1. UYGULAMA GİRİŞ

Yapılan çalışmanın amacı, faaliyet bölgesi İzmir olan bir otomasyon firmasının, kurumsallaşma yolunda ilerlemesini sağlamak ve dolayısıyla müşteri ihtiyaçlarına yönelik pazar stratejileri belirlemek için müşteri benzerliklerini temel alarak en uygun küme sayısını bulmak ve kümeleri belirlemektir. Böylece her kümenin istatistikî özelliklerine bakılarak müşterilere özgü satış politikaları geliştirilebilmekte ve firmanın kârında artış, aynı zamanda müşteri davranışlarını analiz etmek için kullanılan zaman maliyetinde azalış sağlanabilmektedir. Aynı zamanda rakip firmayı tercih eden müşterilerin hangi yönden tercih ettiği bilgisine ulaşılarak kaybedilen müşterinin geri kazanımı söz konusu olabilmektedir. Bu amaç doğrultusunda, verilerin bir kısmı gerçekleşen ticari işlemler sonucu firmadan alınmış, bir kısmı ise uzman görüşü ile elde edilmiştir. Toplamda 394 müşteri vardır ve 11farklı değişken kullanılmıştır. Bu değişkenler kullanılarak farklı kombinasyonlar ile 7 farklı model oluşturulmuştur.

Çalışmada SPSS Clementine 10. 1 ve MATLAB 2016a programları kullanılmıştır. Clementine' de veri madenciliği yöntemlerinden İki Aşamalı ve K-ortalamalar yöntem tercih edilmiştir. İki aşamalı kümeleme metodunun tercih edilme sebebi, küme sayısının bilinmediği durumda uygulanabilir olmasıdır. K-ortalamalar metodunda ise küme sayısı önceden araştırmacı tarafından belirlenmelidir. PSO ile MATLAB programında her model için çıkan sonuçlardaki küme sayıları, K-ortalamalar metodunun küme sayısı olarak kullanılmaktadır.

5.2. VERİ SETİNDE KULLANILAN DEĞİŞKENLER

Müşteri özelliklerini yansıtan değişkenler belirlenip, firmanın satış uzmanı tarafından 394 müşteri için istenilen bilgiler hücrelere doldurulmuştur.

Kullanılan değişkenler aşağıdaki gibidir:

Tablo 2: Veri Setindeki Değişkenlere Ait Özellikler

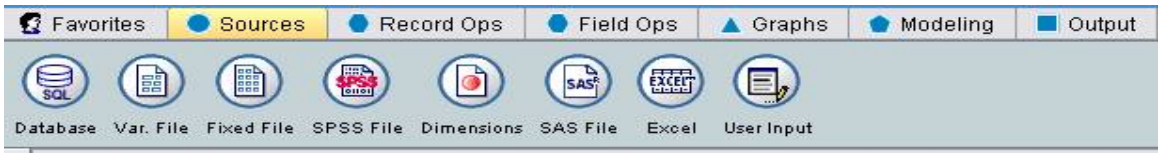
DEĞİŞKENLER	AÇIKLAMA
toplam_potansiyel_2016	Müşterinin 2016 yılı içindeki alım kapasitesini gösterir.
satis_hedefi_2016	Firmanın 2016 yılı için hedefledikleri satışı gösterir.
rakip_firma_gucu	"Zayıf, Orta, Güçlü" şeklinde kategorik olarak belirlenmiş değişkendir. Rakip firmanın rekabet sınıfını belirtir.
rekabet_marka	"Zayıf, Orta, Güçlü" şeklinde kategorik olarak, rekabetin marka odaklı değerlendirilmesi ile belirlenmiştir.
rekabet_fiyatlama	"Zayıf, Orta, Güçlü" şeklinde kategorik olarak, rekabetin fiyatlama odaklı değerlendirilmesi ile belirlenmiştir.
rekabet_kalite	"Zayıf, Orta, Güçlü" şeklinde kategorik olarak, rekabetin kalite odaklı değerlendirilmesi ile belirlenmiştir.
rekabet_hizmet	"Zayıf, Orta, Güçlü" şeklinde kategorik olarak, rekabetin hizmet odaklı değerlendirilmesi ile belirlenmiştir.
inovatifmi	Müşterinin yenilikçi olup olmadığını gösteren değişkendir.
yenilikci_sinifi	"Az, Orta, Yüksek" şeklinde kategorik olarak belirlenmiş değişkendir.
firma_kurumsallik_puani	"0 ile 100" arasında uzman görüşü tarafından müşteriye verilen kurumsallık puanını içeren değişkendir.
odeme_gucu_puani	"0 ile 100" arasında uzman görüş tarafından müşteriye verilen ödeme gücü puanını içeren değişkendir.

5.3. VERİ SETİNİN HAZIRLANMASI VE UYGULAMA AŞAMALARI

PSO' da değişkenlerin sürekli olması gerekmektedir. Bu nedenle kullanılan algoritma, çözüme girmeden önce normalleştirme işlemi yaparak, verilerin (0,1) aralığında olmasını sağlamaktadır. İki aşamalı kümeleme yönteminde ise firma verileri Excel'de Min-Maks normalizasyonu yöntemi ile normalleştirilmiştir. Dolayısıyla SPSS Clementine 10. 1 programında "Type" düğümünde veri türleri

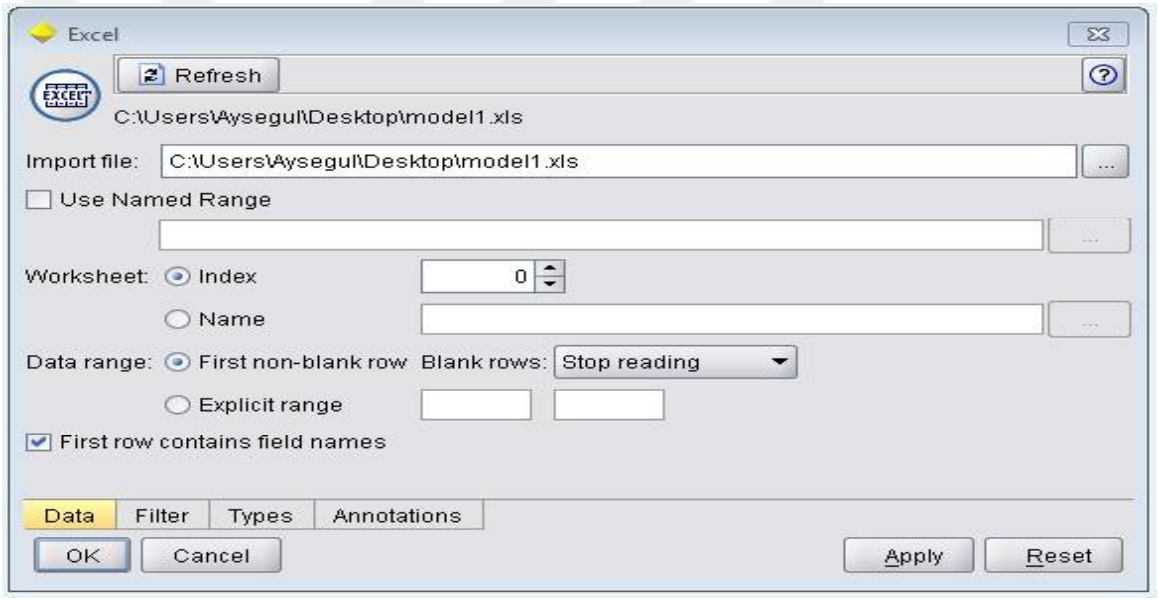
belirlenirken "Range" olarak seçilmiş, müşterinin inovatif olup olmadığını gösteren değişken ise binary olduğu için "Flag" değişken olarak belirlenmiştir. Bu değişkenlerin Clementine programına aktarımı Şekil 13'de gösterilmiştir.

Şekil 13: "Sources"dan "Excel" Düğümü Seçimi



"Excel" düğümü seçilir. "Import File" seçilerek kullanacak olduğumuz veri seti programa yüklenir.

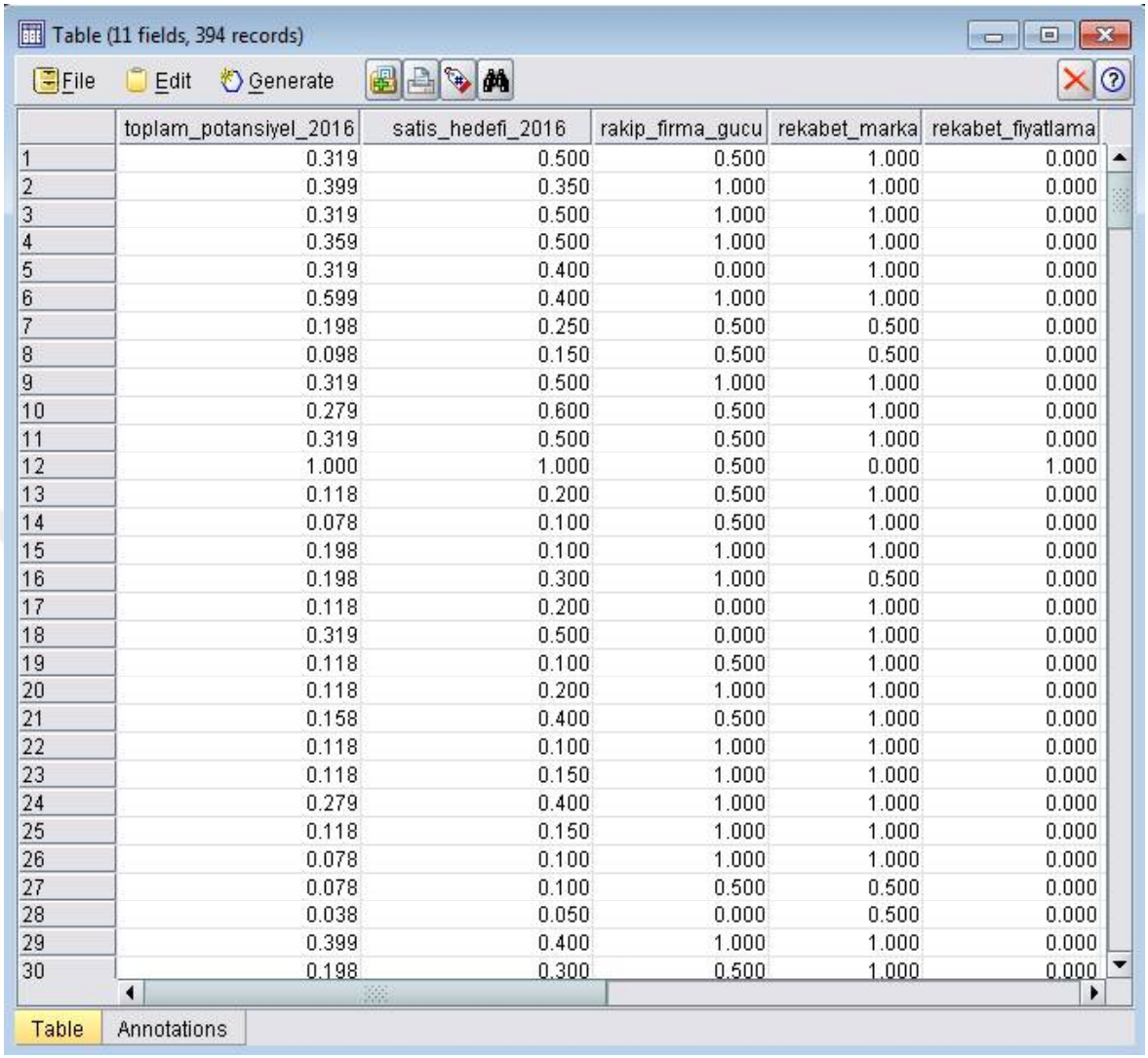
Şekil 14: Clementine' de "Excel" Düğümü ile Veri Setine Sağlanan Erişim



Şekil 15: Veri Kontrolü için "Table" Düğümü

"Table" düğümü seçilir.

Şekil 16: Table" Düğümü ile Verilerin Görünümü



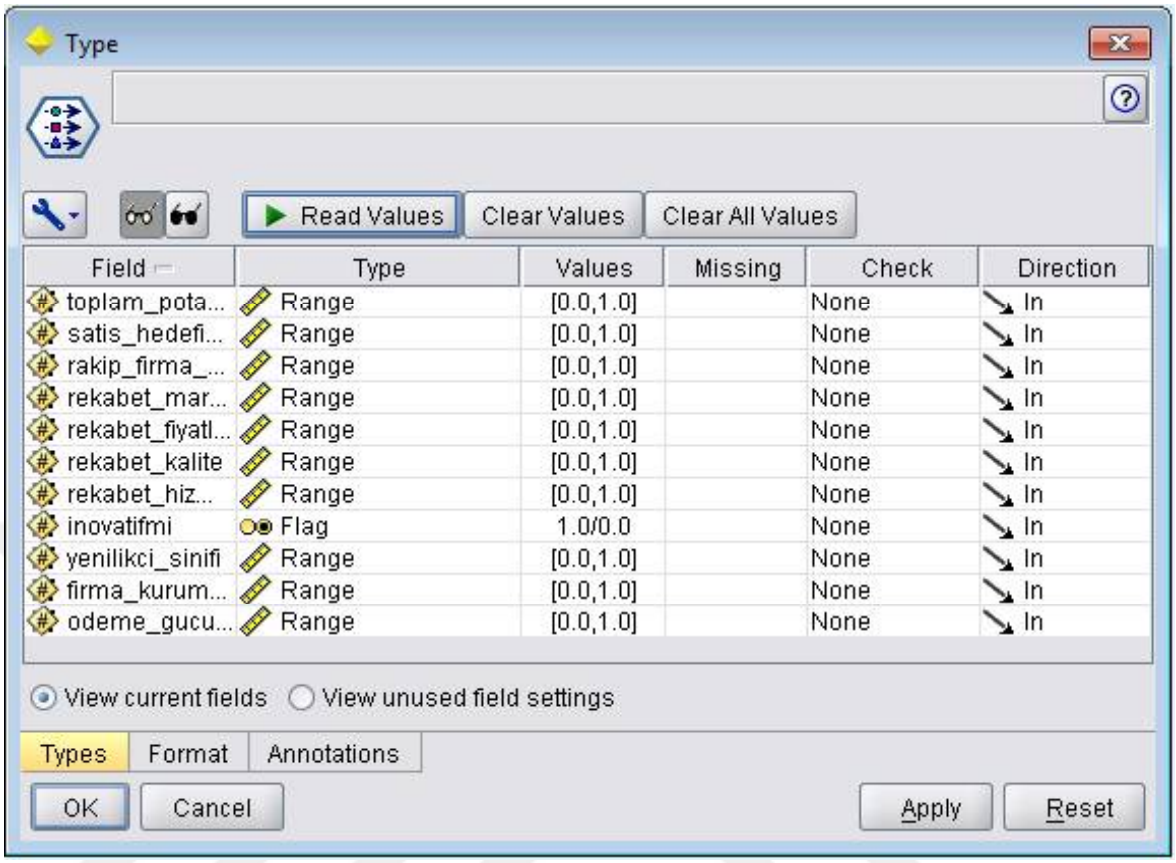
	toplam_potansiyel_2016	satis_hedefi_2016	rakip_firma_gucu	rekabet_marka	rekabet_fiyatlama
1	0.319	0.500	0.500	1.000	0.000
2	0.399	0.350	1.000	1.000	0.000
3	0.319	0.500	1.000	1.000	0.000
4	0.359	0.500	1.000	1.000	0.000
5	0.319	0.400	0.000	1.000	0.000
6	0.599	0.400	1.000	1.000	0.000
7	0.198	0.250	0.500	0.500	0.000
8	0.098	0.150	0.500	0.500	0.000
9	0.319	0.500	1.000	1.000	0.000
10	0.279	0.600	0.500	1.000	0.000
11	0.319	0.500	0.500	1.000	0.000
12	1.000	1.000	0.500	0.000	1.000
13	0.118	0.200	0.500	1.000	0.000
14	0.078	0.100	0.500	1.000	0.000
15	0.198	0.100	1.000	1.000	0.000
16	0.198	0.300	1.000	0.500	0.000
17	0.118	0.200	0.000	1.000	0.000
18	0.319	0.500	0.000	1.000	0.000
19	0.118	0.100	0.500	1.000	0.000
20	0.118	0.200	1.000	1.000	0.000
21	0.158	0.400	0.500	1.000	0.000
22	0.118	0.100	1.000	1.000	0.000
23	0.118	0.150	1.000	1.000	0.000
24	0.279	0.400	1.000	1.000	0.000
25	0.118	0.150	1.000	1.000	0.000
26	0.078	0.100	1.000	1.000	0.000
27	0.078	0.100	0.500	0.500	0.000
28	0.038	0.050	0.000	0.500	0.000
29	0.399	0.400	1.000	1.000	0.000
30	0.198	0.300	0.500	1.000	0.000

"Table" düğümü, Excel düğümüne bağlanıp çalıştırılarak veri kontrolü yapılır. İstenilen veriler görüntülenebiliyorsa, erişim sağlanmış olur. Şekil 16'daki verilere erişimin sağlandığı görülmektedir.

Şekil 17: "Field Ops"dan "Type" Düğümü Seçimi

"Type" düğümü seçilerek değişkenlerin türleri belirlenir ve değerler programa okutulur.

Şekil 18: "Type" Düğümü ile Değişkenlerin Türlerinin Belirlenmesi



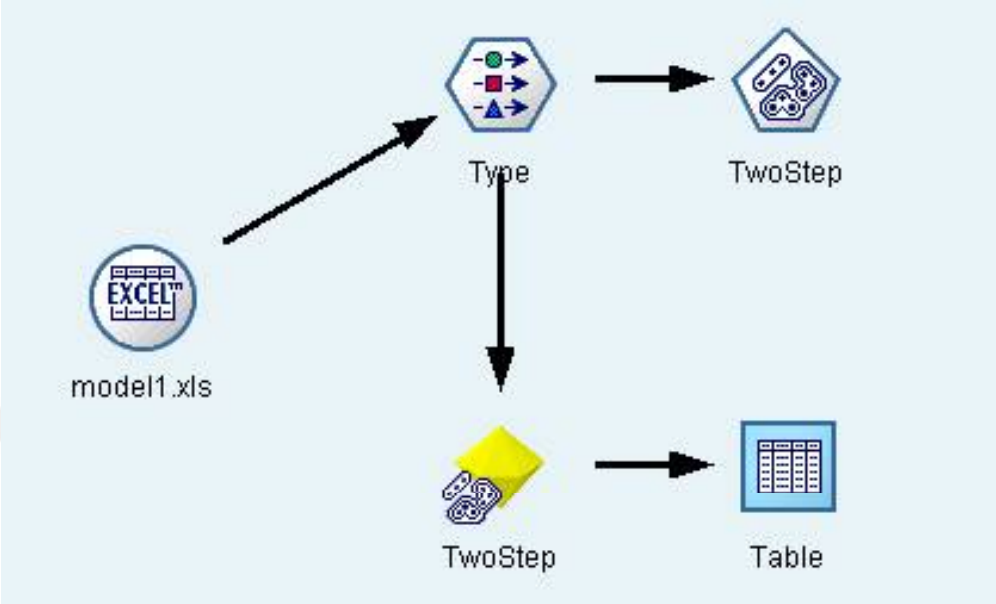
Şekil 19: "Modeling'den Kullanılacak Yöntemin Seçimi (K-ortalamalar, İki Aşamalı)

Bu uygulamada İki aşamalı ve K-ortalamalar kümeleme yöntemleri kullanılmıştır.

Program, yöntem seçildikten sonra çalıştırılarak küme sayısı ve değişkenlerin önem seviyesini, ortalamasını ve standart sapmasını vermektedir. Tekrar "Table" düğümüyle bağlandığında ise hangi müşterinin hangi kümeye ait olduğu bilgisine ulaşılmaktadır.

Model diyagramı aşağıdaki gibidir:

Şekil 20: SPSS Clementine 10.1' de model diyagramı görüntüsü



5.4. MODELLER

Yapılan çalışmada 7 farklı model kullanılmıştır. Kullanılan modeller aşağıdaki gibidir:

Model 1: Tüm değişkenlerin ele alındığı modeldir. Kümeleme müşteri özelliklerinin tümü ele alınarak yapılmıştır.

Model 2: Rekabete dayalı olarak belirlenen modeldir. Bu nedenle modele sadece "rakip_firma_gucu, rekabet_marka, rekabet_fiyatlama, rekabet_kalite, rekabet_hizmet" değişkenleri alınmıştır.

Model 3: Firma yapısına göre belirlenen modeldir. Bu nedenle modele sadece "inovatifmi, yenilikci_sinifi, firma_kurumsallik_puani, odeme_gucu_puani" değişkenleri alınmıştır.

Model 4: İş potansiyeline göre belirlenen modeldir. Bu nedenle modele sadece "toplam_potansiyel_2016, satis_hedefi_2016" değişkenleri alınmıştır.

Model 5: İkinci ve üçüncü model değişkenlerinin birleştirilmesi ile oluşturulan modeldir. Kümeleme, hem rekabet hem de firma yapısı göz önüne alınarak gerçekleştirilmiştir.

Model 6: İkinci ve dördüncü model değişkenlerinin birleştirilmesi ile oluşturulan modeldir. Kümeleme, hem rekabet hem de iş potansiyeli göz önüne alınarak gerçekleştirilmiştir.

Model 7: Üçüncü ve dördüncü model değişkenlerinin birleştirilmesi ile belirlenmiştir. Kümeleme, hem firma yapısı hem de iş potansiyeli göz önüne alınarak gerçekleştirilmiştir.

Tablo 3: Model Tablosu

DEĞİŞKEN	Türü	M1	M2	M3	M4	M5	M6	M7
toplam_potansiyel_2016	Nümerik	✓			✓		✓	✓
satis_hedefi_2016	Nümerik	✓			✓		✓	✓
rakip_firma_gucu	Ordinal	✓	✓			✓	✓	
rekabet_marka	Ordinal	✓	✓			✓	✓	
rekabet_fiyatlama	Ordinal	✓	✓			✓	✓	
rekabet_kalite	Ordinal	✓	✓			✓	✓	
rekabet_hizmet	Ordinal	✓	✓			✓	✓	
inovatifmi	Nominal	✓		✓		✓		✓
yenilikci_sinifi	Ordinal	✓		✓		✓		✓
firma_kurumsallik_puani	Nümerik	✓		✓		✓		✓
odeme_gucu_puani	Nümerik	✓		✓		✓		✓

Seçilen modeller ayrı ayrı kullanılarak uygulamaya tabi tutulmuş ve küme sayıları belirlenmiştir.

5.4.1.Modellerin PSO Kümeleme Algoritması ile Çözümü

Ele alınan problemde Aladağ vd. (2012) tarafından önerilen, Pala ve Aksaraylı (2017) tarafından da kullanılan modifiye edilmiş PSO kullanılmıştır. Bu algoritma sosyal ve bilişsel parametrelerin zamana bağlı değişimine dayanmaktadır.

Kümelemede her bir model için PSO kümeleme algoritması Cura (2012)'de olduğu gibi birbirinden bağımsız 5'er kere çalıştırılmış ve maksimum iterasyon sayısı 200 olarak belirlenmiştir. Parçacık sayısı ise Ortakçı ve Göloğlu (2012)'nun

çalışmasındaki gibi 20 olarak belirlenmiş, küme sayısı minimum 2, maksimum 10 olacak şekilde ayarlanmıştır.

Amaç fonksiyonu olarak Dunn İndeksi kullanılmıştır. Fonksiyon aşağıdaki gibidir(Omran v.d,2006: 338):

$$D = \min_{k=1,\ldots,K} \left\{ \min_{kk=k+1,\ldots,K} \left(\frac{dist(C_k, C_{kk})}{\max_{a=1,\ldots,k} diam(C_a)} \right) \right\}$$

$dist(C_k, C_{kk})$, C_k ve C_{kk} kümeleri arasında benzemezlik fonksiyonudur.

$$dist(C_k, C_{kk}) = \min_{u \in C_k, w \in C_{kk}} d(u, w)$$

$d(u,w)$, u ve w arasındaki Öklid uzaklığını, $diam(C)$ ise kümelerin çaplarını vermektedir.

$$diam(C) = \max_{u,v \in C} diam(u, w)$$

$diam(C)$, herhangi kümedeki birbirine en uzak iki elemanın en büyük Öklid Uzaklığını verir.

Kullanılan algoritmada küme merkezleri arasındaki uzaklık ele alınmıştır.

Model 1:

Model 1' in PSO kümeleme algoritması ile MATLAB programında çözdürülmesi sonucunda müşteriler 3 kümeye ayrılmıştır. Birinci kümeye müşterilerin %20'si (79), ikinci kümeye %57'si (224), üçüncü kümeye ise %23'ü (91) atanmıştır. Birinci kümeye atanan müşteriler ödeme gücü genellikle %80' in üzerinde olan ve toplam potansiyeli yüksek olan müşterilerden oluşmaktadır. Bu nedenle bu kümedeki müşterilere yapılan satışlarda ödemenin zamanında yapılacağı düşüncesiyle hareket edilerek, imtiyaz sağlanabilir. İkinci kümede ise kalite açısından rakip firmanın zayıf olduğu müşteriler kümelenmiştir. Bu kümedeki müşterilerin ise aldıkları muadil üründen vazgeçip, firmadan alım yapabilmeleri için ürünün kalitesi ile ilgili daha çok bilgilendirme yapılarak, satış potansiyeli arttırılabilir. Üçüncü kümeye atanan müşterilere baktığımızda ise müşterilerin yenilikçi bir yapıya sahip oldukları görülmektedir. Bu müşteriler ürünler hakkında

bilgilendirilirken teknolojik gelişmelerden, ürünün geliştirilen özelliklerinden bahsetmek müşteriyi alış davranışına itebilir.

Model 2:

Rekabet odaklı oluşturulan Model 2' nin PSO kümeleme algoritması ile MATLAB programında çözdürülmesi sonucunda müşteriler 5 kümeye ayrılmıştır. Birinci kümeye müşterilerin %10'u (41), ikinci kümeye %31'i (121), üçüncü kümeye %12'si (48), dördüncü kümeye %16'sı (62), beşinci kümeye ise %31'i (122) atanmıştır. Birinci kümede marka ve kalite açısından zayıf, hizmet açısından güçlü rakip firmaların olduğu müşteriler bulunmaktadır. Bu müşterilere, satılan ürünün marka ve kalite açısından daha iyi olduğunu göstermekle kalmayıp, hizmet konusunda da rakip firmanın üzerine çıkmak şirketin satışlarını pozitif yönde etkileyecektir. İkinci kümede hizmet açısından rakip firmanın zayıf olduğu müşteriler bulunmaktadır. Bu kümedeki müşterilere uygulanan satış politikasında, satış öncesi ve sonrası hizmetlerin optimum seviyede tutulması gerekmektedir. Üçüncü kümede marka ve kalite açısından zayıf, fiyatlama ve hizmet açısından güçlü bir rakip firmanın bulunduğu müşteriler kümelenmiştir. Bu müşterilerin toplam potansiyelleri ve dolayısıyla firmanın müşterilere satış hedefleri düşüktür. Muadil ürünü tercih etme sebepleri maliyetin düşük olması ve aynı zamanda hizmet konusunda iyi olmalarıdır. Dördüncü kümede sadece fiyatlama açısından güçlü olan rakip firmanın bulunduğu müşteriler yer almaktadır. Bu müşteriler diğer değişkenler dikkate alınmadan sadece fiyat odaklı hareket etmektedirler. Beşinci kümede ise alım potansiyeli, ödeme gücü ve firma kurumsallık puanı yüksek olan, her açıdan güçlü bir rakip firmanın bulunduğu, firma için önem arz eden müşteriler kümelenmiştir. Bu müşterilere uygulanacak satış politikası üzerinde ayrıca çalışılmalıdır.

Model 3:

Firma yapısı odaklı oluşturulan Model 3' ün PSO kümeleme algoritması ile MATLAB programında çözdürülmesi sonucunda müşteriler 3 kümeye ayrılmıştır. Birinci kümeye müşterilerin %27'si (107), ikinci kümeye %65'i (256), üçüncü kümeye %8'i (31) atanmıştır. Birinci kümede inovatif olmayan müşteriler kümelenmiştir. Genellikle bu müşterilerin alım alışkanlıkları değişkenlik göstermemektedir ve geçmiş yıllara bakılarak ihtiyaçları bulunup, o yönde bir satış politikası düzenlenmelidir. İkinci kümede inovatif olup, firma kurumsallık ve

ödeme gücü puanları yüksek olan firmalar kümelenmiştir. Üçüncü kümede ise firma kurumsallık ve ödeme gücü puanları genellikle 50'nin altında olan müşteriler kümelenmiştir. Bu da firmanın bu müşterilere olan satış hedeflerinin düşük olmasının anlamlı olduğunu göstermektedir.

Model 4:

İş potansiyeli odaklı Model 4' ün PSO kümeleme algoritması ile MATLAB programında çözdürülmesi sonucunda müşteriler 2 kümeye ayrılmıştır. Birinci kümeye müşterilerin %97'si (383), ikinci kümeye ise %3'ü (11) atanmıştır. Burada iki küme arasında belirgin farklılıklar bulunmamaktadır. Modelde değişken sayısının 2 olması çözümde optimum kümeleme sayısının bulunmamasına sebep olmuş olabilir.

Model 5:

Rekabet ve firma yapısı odaklı Model 5' in PSO kümeleme algoritması ile MATLAB programında çözdürülmesi sonucunda müşteriler 2 kümeye ayrılmıştır. Birinci kümeye müşterilerin %35'i (138), ikinci kümeye ise %65'i (256) atanmıştır. Birinci kümedeki müşterilerin marka ve kalite açısından "güçlü", fiyatlama ve hizmet açısından "orta" olan rakip firmaları tercih ettiği görülmektedir. Fakat aynı zamanda bu müşteriler, toplam potansiyel ve satış hedeflerinin yüksek olduğu müşterilerdir. Buradan bu müşterilerin bazı ürünleri firmanın sattığı markadan, bazılarını ise rakip firmadan temin ettiği çıkarımı yapılmaktadır. Rakip firmadan temin edilen ürünün marka ve kalite üstünlüğü tespit edilerek, bu konuda iyileştirmeye gidilmelidir. İkinci kümedeki müşterilerin ise rakip firmayı tercih etme sebepleri yalnızca fiyatlamadır. Ürünün kalitesinden ve marka değerinden ziyade fiyat konusuna odaklanmışlardır. Ancak beklentilerinin altında bir fiyat politikası uygulanamayacağı için bu müşteriler için satış hedefleri de yüksek tutulmamaktadır.

Model 6:

Rekabet ve iş potansiyeli odaklı Model 6'nın PSO kümeleme algoritması ile MATLAB programında çözdürülmesi sonucunda müşteriler 2 kümeye ayrılmıştır. Birinci kümeye müşterilerin %69'u (272), ikinci kümeye ise %31'i (122) atanmıştır. Birinci kümedeki müşteriler, ikinci kümeye göre toplam potansiyeli daha yüksek olan müşterilerdir. Bu durumda firma birinci kümedeki müşterilere öncelik vererek bir satış politikası gerçekleştirmelidir. Böylece alım potansiyeli yüksek müşteriler

rakip firma ürünü yerine firma ürününü tercih etmiş olur ve böylece firma, satış hedeflerini gerçekleştirmiş olur.

Model 7:

Firma yapısı ve iş potansiyeli odaklı Model 7'nin PSO kümeleme algoritması ile MATLAB programında çözdürülmesi sonucunda müşteriler 3 kümeye ayrılmıştır. Birinci kümeye müşterilerin %31'i (122), ikinci kümeye %59'u (234), üçüncü kümeye ise %10'u (38) atanmıştır. Birinci küme yenilikçi sınıfı "az" olarak tanımlanan, ikinci küme "inovatif" müşterilerin olduğu, üçüncü küme ise "inovatif olmayan" müşterilerin bulunduğu kümedir. Bu modelde tek bir değişkenin baskın olduğu görülmektedir. Müşterilere satış politikaları düzenlenirken yenilikçi yaklaşımlar sunulup sunulmayacağına bu model sayesinde karar verilebilmektedir.

5.4.2.Modellerin K-ortalamalar Kümeleme Yöntemi ile Çözümü

Burada küme sayısı olarak, PSO ile kümelenme gerçekleştirildiğinde meydana gelen küme sayıları kullanılmıştır.

Model 1:

Model 1'de küme sayısı, 3 olarak önceden belirlenmiştir. Birinci kümeye müşterilerin %32'si (126), ikinci kümeye %28'i (110), üçüncü kümeye ise %40'ı (158) atanmıştır. Küme sayısı aynı verilmesine rağmen küme yoğunlukları PSO algoritması ile belirgin farklılık göstermektedir. Burada tüm değişkenler alınmasına rağmen kümeler rekabet değişkenlerine göre meydana gelmiştir.

Model 2:

Model 2'de küme sayısı, 5 olarak önceden belirlenmiştir. Birinci kümeye müşterilerin %32'si (128), ikinci kümeye %20'si (77), üçüncü kümeye %14'ü (57), dördüncü kümeye %16'sı (62), beşinci kümeye ise %18'i (70) atanmıştır. PSO algoritması ile oluşturulan kümelerle aynı özelliklere sahiptir.

Model 3:

Model 3'de küme sayısı, 3 olarak önceden belirlenmiştir. Birinci kümeye müşterilerin %53'ü (208), ikinci kümeye %31'i (122), üçüncü kümeye %16'sı (64) atanmıştır. Birinci küme inovatif olup, ödeme gücü puanları yüksek olan müşterilerden, ikinci küme inovatif olmayan müşterilerden, üçüncü küme ise inovatif

olup, ödeme gücü puanı ve firma kurumsallık puanı düşük olan müşterilerden oluşmaktadır. Yani PSO ile aynı değişkenlere bakılarak aynı şekilde kümelenmiştir, yalnızca küme yoğunluklarında değişiklikler bulunmaktadır.

Model 4:

Model 4'de küme sayısı, 2 olarak önceden belirlenmiştir. Birinci kümeye müşterilerin %100'e yakın (393), ikinci kümeye ise yalnızca 1 müşteri atanmıştır. Diğer iki kümeleme yönteminde de buna yakın sonuçlar çıkmıştır. Bu modelde doğru şekilde bir kümeleme yapılamamaktadır. Bunun sebebi, daha önce PSO ile çözülen modelde de açıklandığı gibi modele sadece iki değişkenin alınması olabilir.

Model 5:

Model 5'de küme sayısı, 2 olarak önceden belirlenmiştir. Birinci kümeye müşterilerin %35'i (139), ikinci kümeye ise %65'i (255) atanmıştır. Bu modelde PSO kümeleme algoritması ile bulunan küme elemanlarının neredeyse tamamı aynıdır.

Model 6:

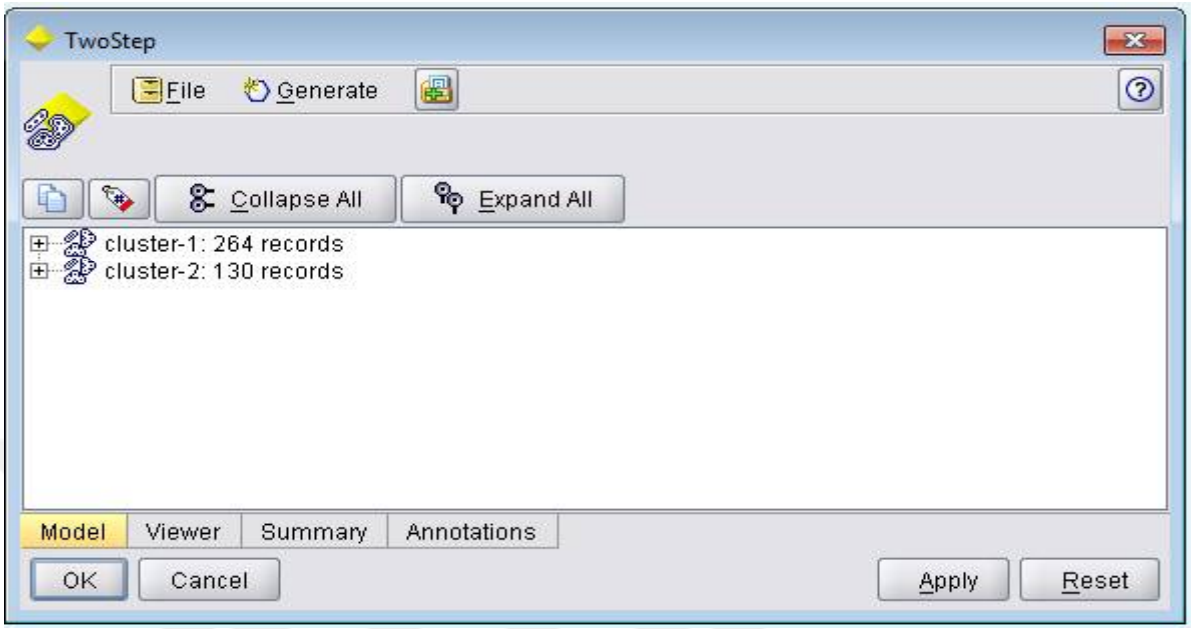
Model 6'da küme sayısı, 2 olarak önceden belirlenmiştir. Birinci kümeye müşterilerin %72'si (284), ikinci kümeye ise %28'i (110) atanmıştır. PSO kümeleme algoritması ile bulunan sonuçlarla benzerlik göstermektedir.

Model 7:

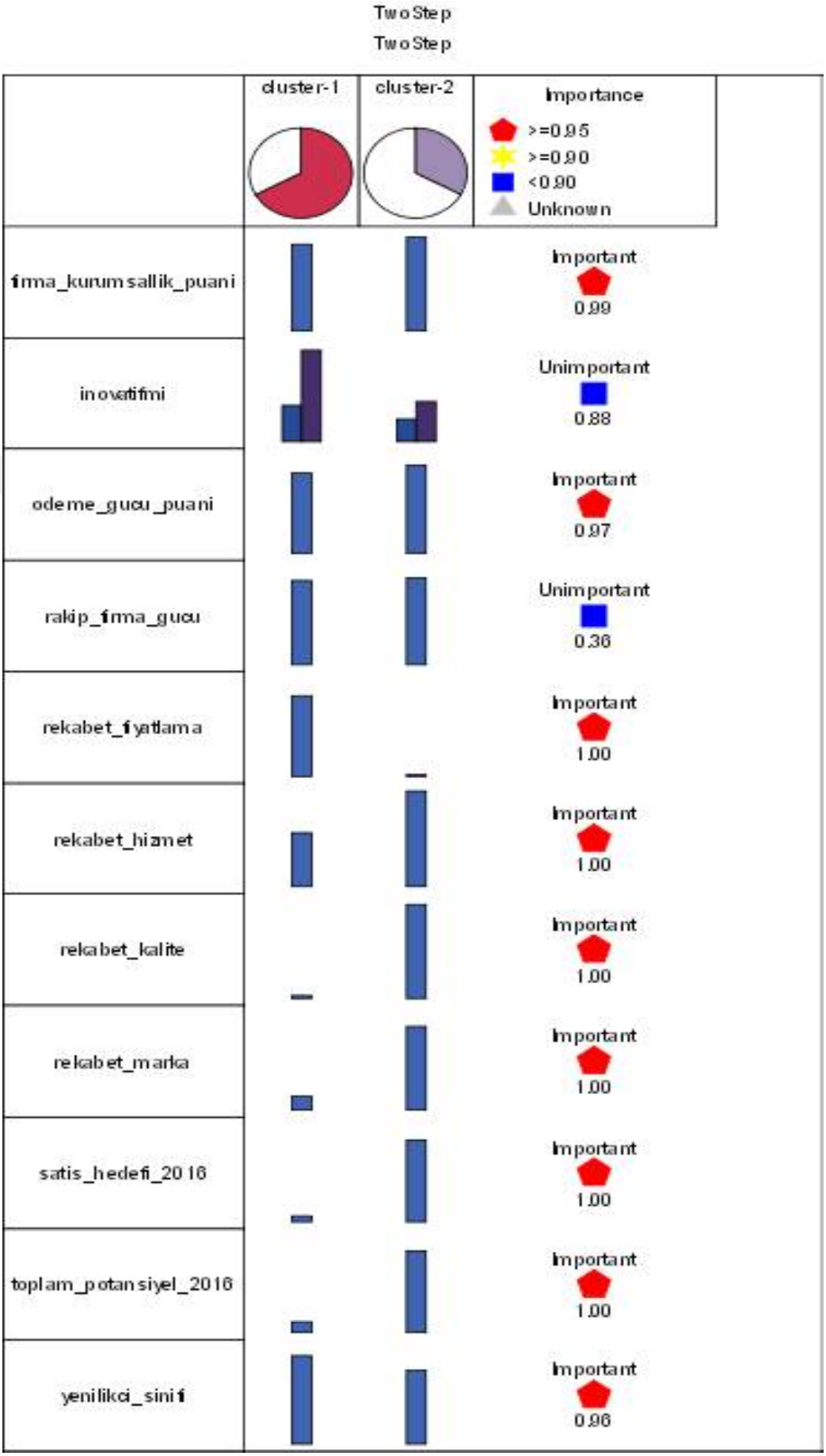
Firma yapısı ve iş potansiyeli odaklı Model 7'de küme sayısı, 3 olarak önceden belirlenmiştir. Birinci kümeye müşterilerin %31'i (122), ikinci kümeye %59'u (234), üçüncü kümeye ise %10'u (38) atanmıştır. Küme sayısı ve kümelere atanan müşteri sayıları PSO kümeleme algoritması ile aynı bulunmuştur. Bu model için müşterilere yaklaşımı belirleyecek değişken "inovatiflik" değişkenidir.

5.4.3.Modellerin İki Aşamalı Kümeleme Yöntemi ile Çözümü

Model 1:

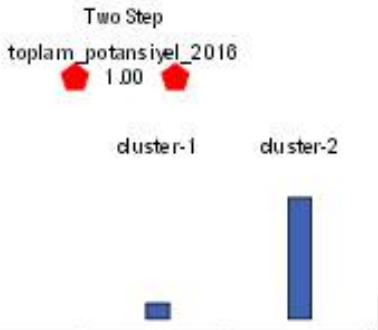


Şekil de görüldüğü üzere tüm değişkenlerin ele alındığı modelde küme sayısı 2 olarak belirlenmiştir. Birinci kümeye müşterilerin %67'si (264) atanırken, ikinci kümeye %33'ü (130) atanmıştır.


TwoStep
TwoStep
cluster-1
cluster-2
Importance
>=0.95
>=0.90
<0.90
Unknown
firma_kurumsallik_puani
Important
0.99
inovatifmi
Unimportant
0.88
odeme_gucu_puani
Important
0.97
rakip_firma_gucu
Unimportant
0.36
rekabet_fiyatlama
Important
1.00
rekabet_hizmet
Important
1.00
rekabet_kalite
Important
1.00
rekabet_marka
Important
1.00
satis_hedefi_2016
Important
1.00
toplam_potansiyel_2016
Important
1.00
yenilikci_sinifi
Important
0.96

%90 önemlilik seviyesinin altındaki değişkenler önemsiz, %90 - %95 arasında kalan değişkenler marjinal, %95 üzerinde olan değişkenler ise önemli olarak kabul edilmekte ve bu şekilde yorumlanmaktadır.

Değişkenlerin üzerine tıkladığımızda her bir küme için önem seviyeleri, ortalamaları ve standart sapmaları görülmektedir. Çıktı görüntüsü aşağıdaki gibidir:



	cluster-1	cluster-2
Importance	0.00	1.00
Mean	0.01	0.08
Standard Deviation	0.02	0.13

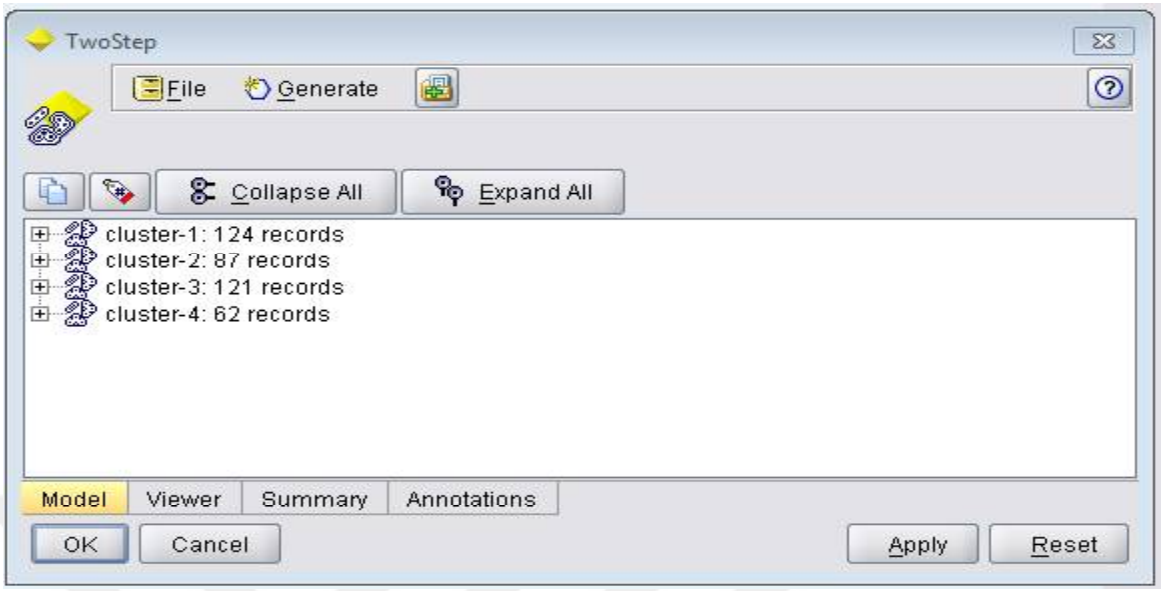
Her iki küme içerisindeki değişkenlerin tanımlayıcı istatistikleri aşağıdaki gibi elde edilmiştir.

Tablo 4: Model 1 için Seçilen Değişkenlerin Önem Seviyeleri, Ortalamaları ve Standart Sapmaları Tablosu

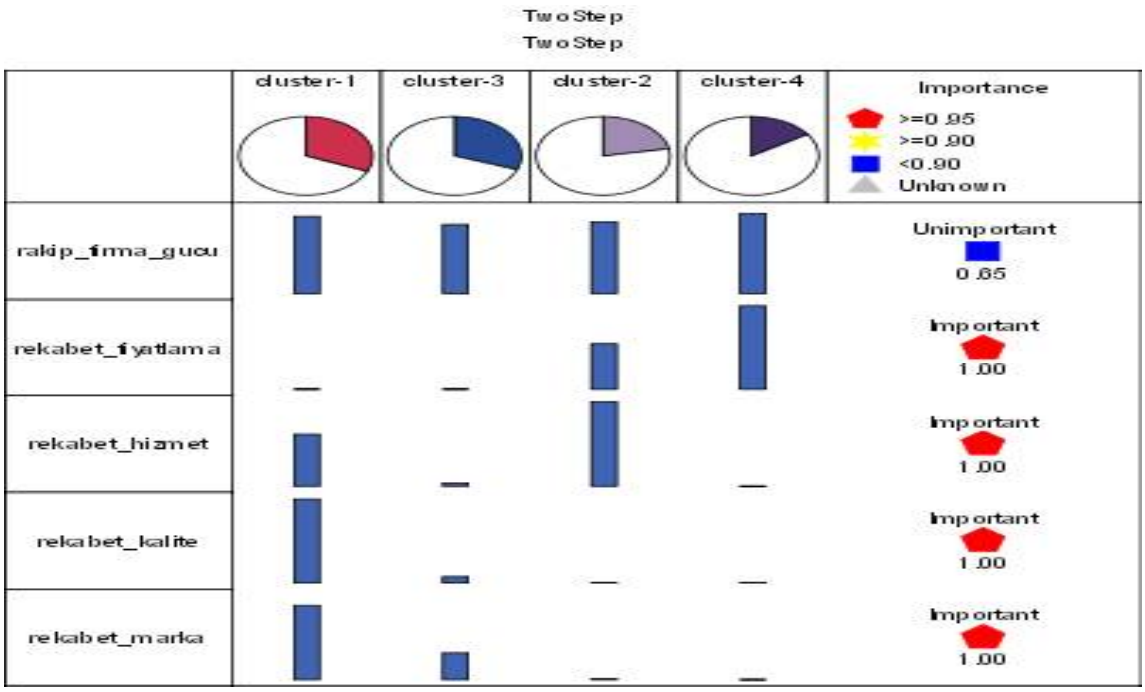
DEĞİŞKENLER	**KÜME 1**			**KÜME 2**		
	Ö.S	Ort	SS	Ö.S	Ort	SS
toplam_potansiyel_2016	0.00	0.01	0.02	1.00	0.08	0.13
satis_hedefi_2016	0.00	0.01	0.02	1.00	0.10	0.17
rakip_firma_gucu	0.39	0.60	0.38	0.65	0.62	0.38
rekabet_marka	0.00	0.14	0.23	1.00	0.86	0.23
rekabet_fiyatlama	1.00	0.41	0.49	0.00	0.01	0.09
rekabet_kalite	0.00	0.03	0.11	1.00	0.96	0.16
rekabet_hizmet	0.00	0.34	0.47	1.00	0.60	0.23
inovatifmi	0.63	-	-	0.80	-	-
yenilikci_sinifi	0.88	0.54	0.41	0.05	0.45	0.42
firma_kurumsallik_puani	0.07	0.71	0.21	0.99	0.77	0.17
odeme_gucu_puani	0.10	0.66	0.27	0.97	0.72	0.26

Elde edilen sonuçlara göre Küme 1 içerisindeki değişkenlerin ortalamaları, merkezleri göstermekte olup, standart sapmaları ise yayılımı belirtmektedir. En fazla yayılıma sahip olan değişkenin önem seviyesi oldukça yüksektir. Bu değişken, firma tarafından satış politikasında dikkate alınması gereken bir faktördür. Küme 1 içerisinde yer alan müşteriler, rakip firmayı fiyat politikası nedeniyle tercih etmektedir. Bu durumda firma bundan sonraki fiyat politikalarında iyileştirmeye gitmelidir. Fakat Küme 2 de ise yenilikçi sınıfı, inovatiflik, rekabet fiyatlama, rakip firma gücü dışındaki tüm değişkenler yüksek öneme sahiptir. Küme 2 firmanın kurumsallığı, ödeme alışkanlıkları, alım potansiyeli gibi değişkenlere göre kümelenmiştir ve bu kümeye ait müşteriler ise rakip firmayı kalite ve hizmet değişkenlerini dikkate alarak tercih etmektedir. Bu müşterilere uygulanacak satış politikasında ürünün kalitesini ispatlamaya odaklı olunması, satış öncesi ve sonrası hizmetlerin eksiksiz sunulması gerekmektedir.

Model 2:



Şekil de görüldüğü üzere rekabete dayalı modelde küme sayısı 4 olarak belirlenmiştir. Birinci kümeye müşterilerin %31'i (124), ikinci kümeye %22'si (87), üçüncü kümeye %31'i (121), dördüncü kümeye ise %16'sı (62) atanmıştır. Her bir küme içerisindeki değişkenlerin tanımlayıcı istatistikleri aşağıdaki gibi elde edilmiştir.

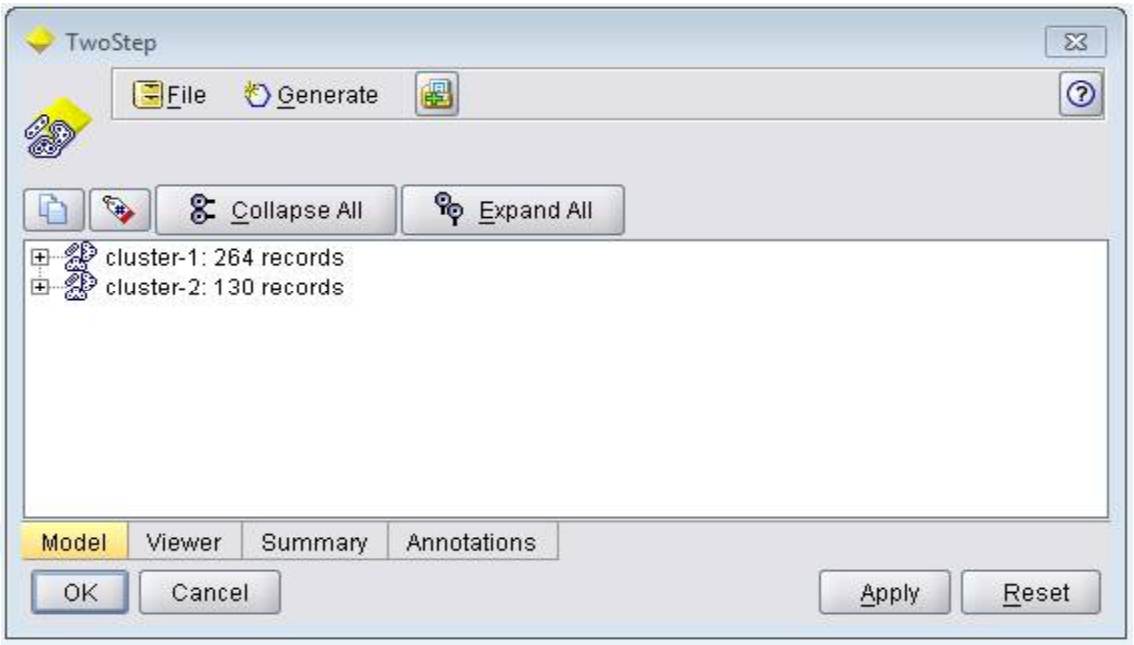


Tablo 5: Model 2 için Seçilen Değişkenlerin Önem Seviyeleri, Ortalamaları ve Standart Sapmaları Tablosu

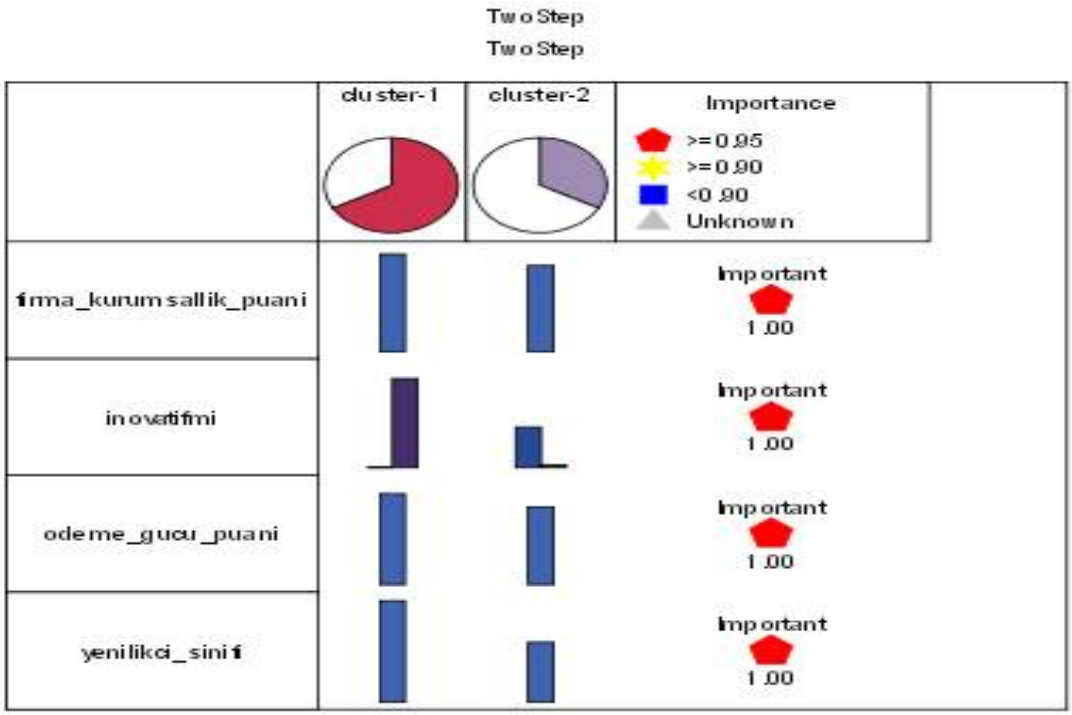
DEĞİŞKENLER	**KÜME 1**			**KÜME 2**		
	Ö.S	Ort	SS	Ö.S	Ort	SS
rakip_firma_gucu	0.78	0.64	0.38	0.33	0.59	0.39
rekabet_marka	1.00	0.88	0.21	0.00	0.01	0.11
rekabet_fiyatlama	1.00	0.00	0.00	1.00	0.55	0.50
rekabet_kalite	1.00	0.99	0.06	1.00	0.00	0.00
rekabet_hizmet	1.00	0.61	0.21	1.00	1.00	0.00
DEĞİŞKENLER	**KÜME 3**			**KÜME 4**		
	Ö.S	Ort	SS	Ö.S	Ort	SS
rakip_firma_gucu	0.13	0.57	0.38	0.87	0.66	0.36
rekabet_marka	0.00	0.32	0.24	1.00	0.00	0.00
rekabet_fiyatlama	1.00	0.00	0.00	1.00	1.00	0.00
rekabet_kalite	0.00	0.07	0.18	1.00	0.00	0.00
rekabet_hizmet	0.00	0.03	0.12	1.00	0.00	0.00

Diğer iki yöntemde 5 olarak bulunan küme sayısı, bu yöntemde 4 olarak bulunmuştur. Yukarıdaki sonuçlara bakıldığında bu modelde, PSO' da iki ayrı küme olan birinci ve üçüncü küme, tek bir küme olarak(ikinci küme) ele alınmıştır. PSO' da birinci kümede yer almayıp üçüncü kümeyle birleştirilince dikkate alınan değişken fiyatlama değişkeni olarak görülmektedir.

Model 3:



Şekil de görüldüğü üzere firma yapısına dayalı modelde küme sayısı 2 olarak belirlenmiştir. Birinci kümeye müşterilerin %67'si (264) müşteri atanırken, ikinci kümeye %33'ü (130) müşteri atanmıştır. Tüm değişkenlerin ele alındığı birinci modelle aynı sayıda atama yapılmıştır. Ancak hücrelerin atandığı kümelerde farklılık bulunmaktadır.

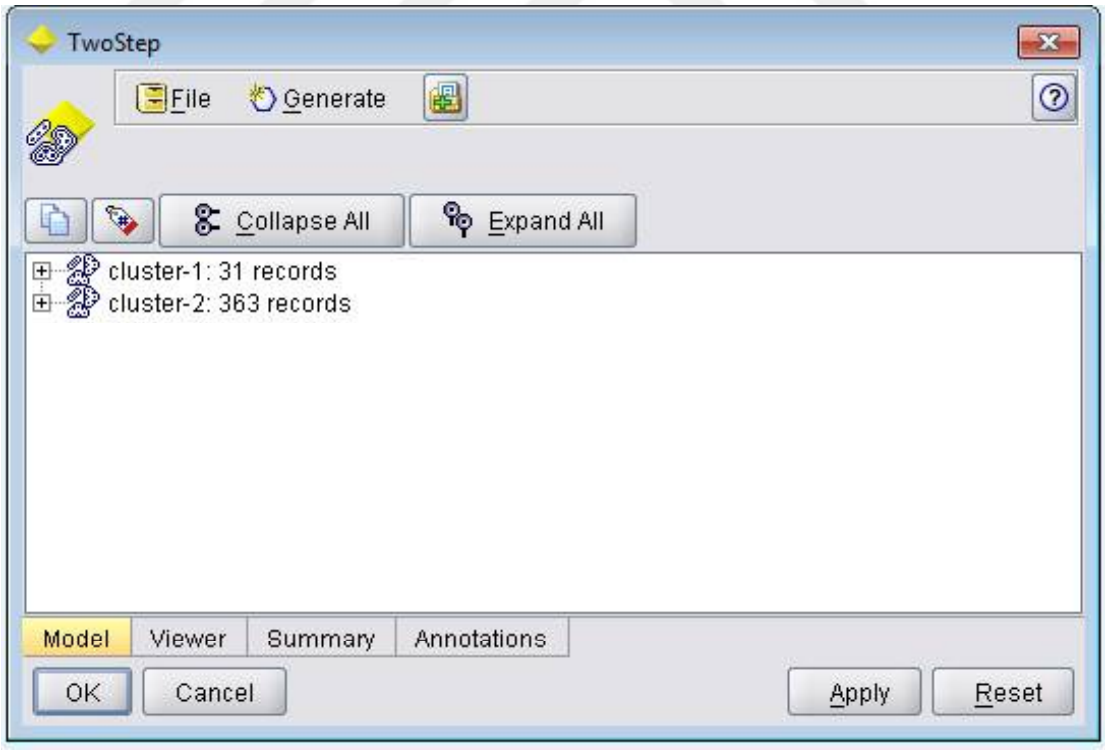


Tablo 6: Model 3 için Seçilen Değişkenlerin Önem Seviyeleri, Ortalamaları ve Standart Sapmaları Tablosu

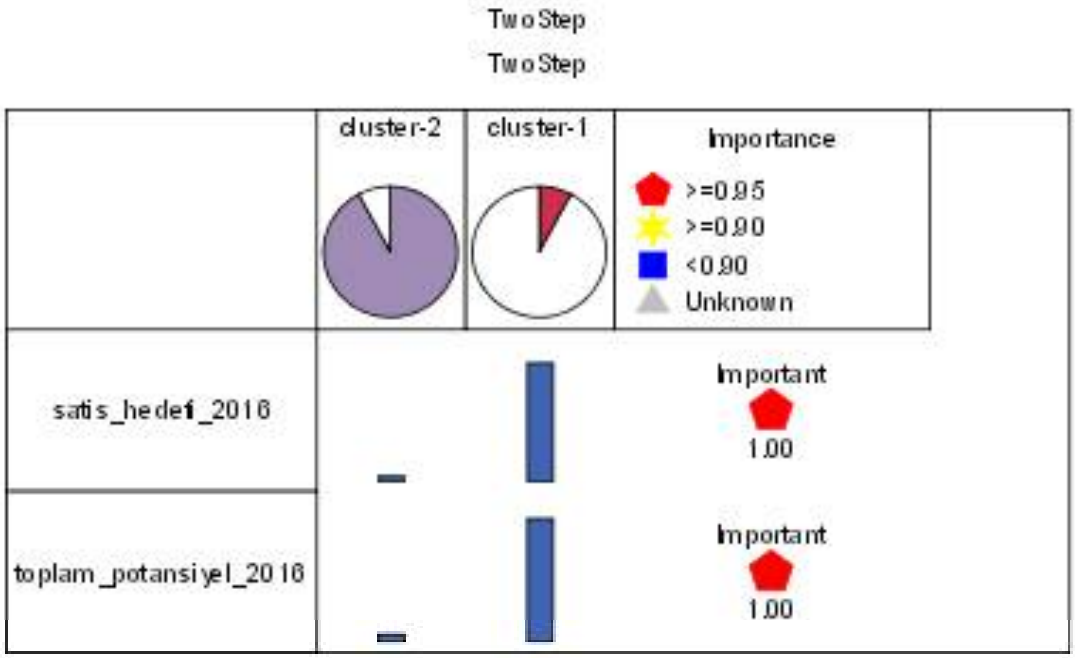
DEĞİŞKENLER	**KÜME 1**			**KÜME 2**		
	Ö.S	Ort	SS	Ö.S	Ort	SS
inovatifmi	1.00	-	-	1.00	-	-
yenilikci_sinifi	1.00	0.59	0.36	0.00	0.35	0.46
firma_kurumsallik_puani	1.00	0.76	0.16	0.00	0.67	0.25
odeme_gucu_puani	0.99	0.71	0.23	0.01	0.61	0.32

Küme 1 için tüm değişkenler, Küme 2 için ise yalnızca inovatiflik değişkeni önem arz etmektedir. PSO' da 3 olarak belirlenen küme sayısının burada 2 olmasının sebebi firma kurumsallık puanı değişkeni de eklenerek PSO' daki birinci ve üçüncü kümelerin birleştirilerek bir kümeye indirgenmesidir.

Model 4:



Şekil de görüldüğü üzere iş potansiyeline dayalı modelde küme sayısı 2 olarak belirlenmiştir. Birinci kümeye müşterilerin %8' i (31) müşteri atanırken, ikinci kümeye %92'si (363) atanmıştır.



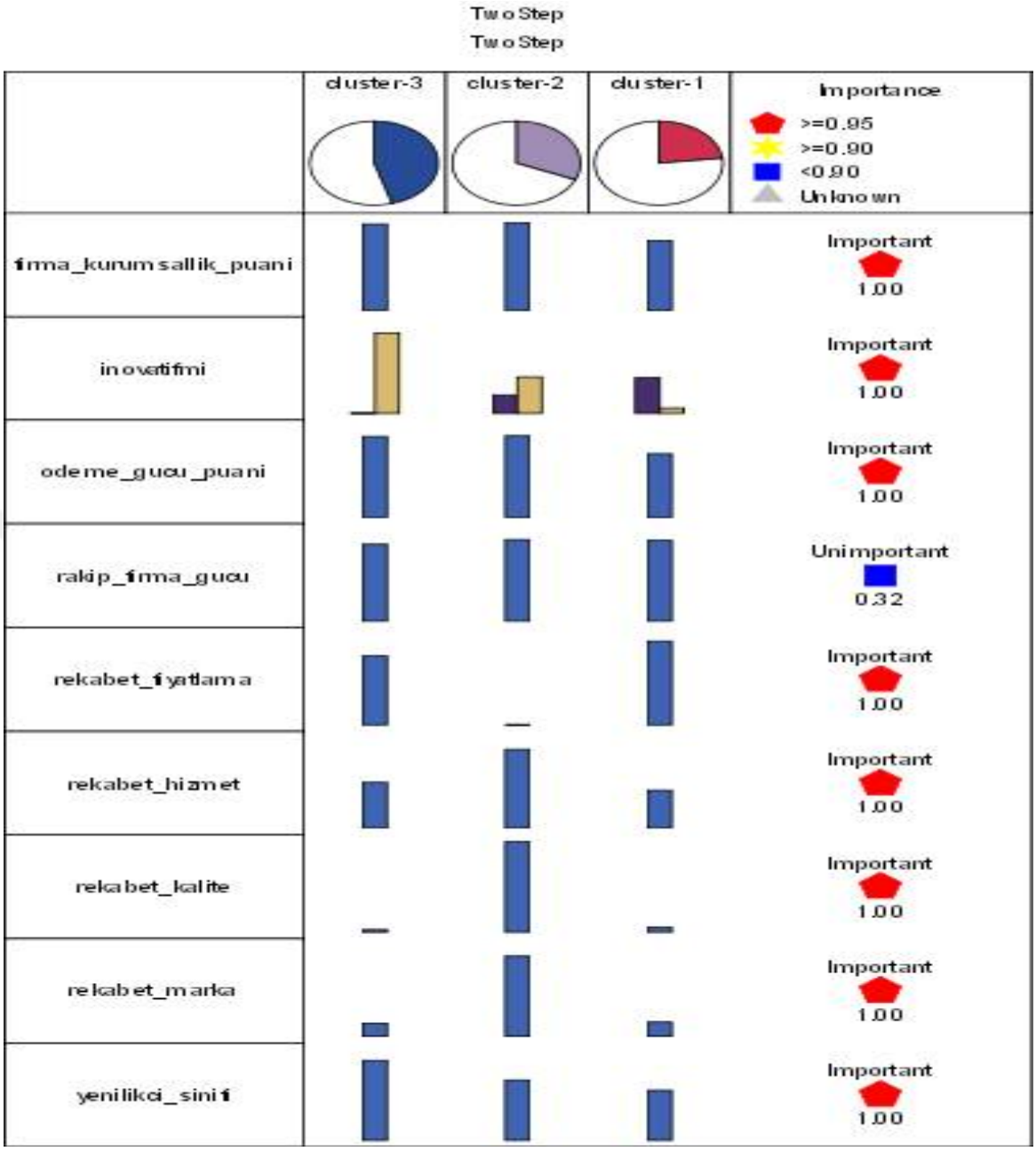
Tablo 7: Model 4 için Seçilen Değişkenlerin Önem Seviyeleri, Ortalamaları ve Standart Sapmaları Tablosu

DEĞİŞKENLER	**KÜME 1**			**KÜME 2**		
	Ö.S	Ort	SS	Ö.S	Ort	SS
toplam_potansiyel_2016	0.00	0.01	0.02	1.00	0.08	0.13
satis_hedefi_2016	0.00	0.01	0.02	1.00	0.10	0.17

Yalnızca bu iki değişkenin alındığı modelde kümeleme işlemi uygun sonuç vermemektedir. Küme 1 de değişkenlerin her ikisi de önemsiz, Küme 2 de ise önemli olarak görülmektedir. Bu durum optimum kümeleme işlemini gerçekleştirmemize engel teşkil etmektedir. Farklı değişkenler eklenerek yeni bir model kurulmalıdır. Aksi taktirde bu model firma için yararlı olmayacaktır.

Model 5:

Şekil de görüldüğü üzere rekabet ve firma yapısına dayalı modelde küme sayısı 3 olarak belirlenmiştir. Birinci kümeye müşterilerin %23'ü (91), ikinci kümeye %32'si (124), üçüncü kümeye ise %45'i (179) atanmıştır.


Two Step
Two Step
cluster-3
cluster-2
cluster-1
Importance
>=0.95
>=0.90
<0.90
Unknown
firma_kurumsallik_puani
inovatifmi
odeme_gucu_puani
rakip_firma_gucu
rekabet_fiyatlama
rekabet_hizmet
rekabet_kalite
rekabet_marka
yenilikci_sinif
Important
1.00
Important
1.00
Important
1.00
Unimportant
0.32
Important
1.00
Important
1.00
Important
1.00
Important
1.00
Important
1.00

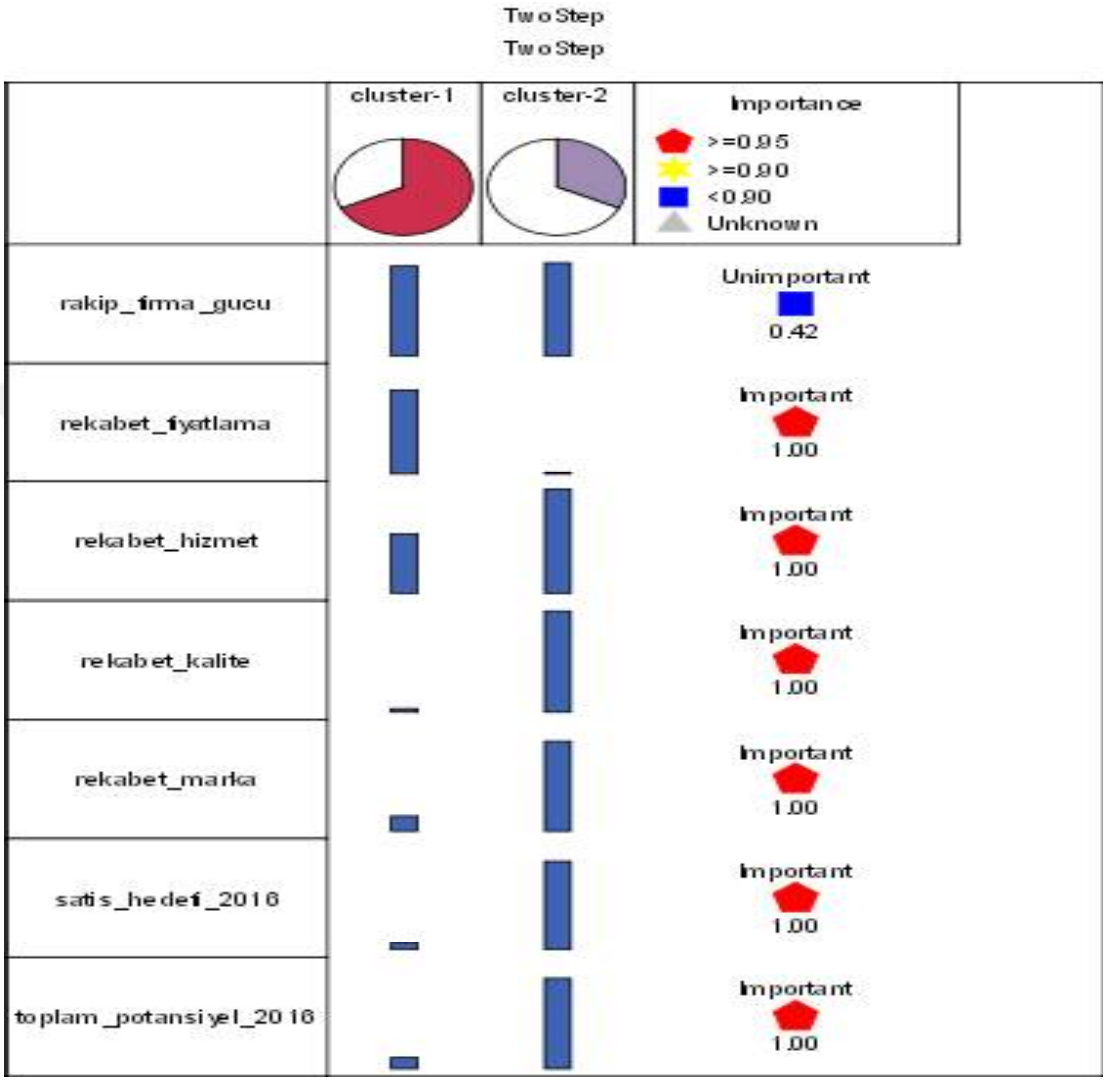
Tablo 8: Model 5 için Seçilen Değişkenlerin Önem Seviyeleri, Ortalamaları ve Standart Sapmaları Tablosu

DEĞİŞKENLER	KÜME 1			KÜME 2			KÜME 3		
	Ö.S	Ort	SS	Ö.S	Ort	SS	Ö.S	Ort	SS
rakip_firma_gucu	0.60	0.62	0.38	0.71	0.63	0.38	0.26	0.59	0.38
rekabet_marka	0.00	0.15	0.23	1.00	0.88	0.21	0.00	0.14	0.23
rekabet_fiyatlama	1.00	0.46	0.50	1.00	0.00	0.00	1.00	0.38	0.49
rekabet_kalite	0.00	0.05	0.15	1.00	0.99	0.06	0.00	0.03	0.11
rekabet_hizmet	0.00	0.30	0.45	1.00	0.61	0.21	0.03	0.36	0.47
inovatifmi	1.00	-	-	0.52	-	-	1.00	-	-
yenilikci_sinifi	0.00	0.38	0.46	0.09	0.46	0.41	1.00	0.61	0.35
firma_kurumsallik_puani	0.00	0.62	0.27	1.00	0.77	0.17	1.00	0.76	0.14
odeme_gucu_puani	0.00	0.56	0.32	0.95	0.72	0.27	0.98	0.71	0.22

Burada küme sayısı PSO' da bulunan küme sayısına göre artmıştır. PSO' da bulunan ikinci küme, İki Aşamalı Algoritma yöntemi ile iki kümeye ayrılmıştır. Yenilikçi sınıfı, firma kurumsallık puanı, ödeme gücü puanı gibi değişkenler ayrıca dikkate alınarak değerlendirilecek, böylece müşterilere olan yaklaşım optimum hale gelecektir.

Model 6:

Şekil de görüldüğü üzere rekabet ve iş potansiyeline dayalı modelde küme sayısı 2 olarak belirlenmiştir. Birinci kümeye müşterilerin %68'i (267), ikinci kümeye ise %32'si (127) atanmıştır.

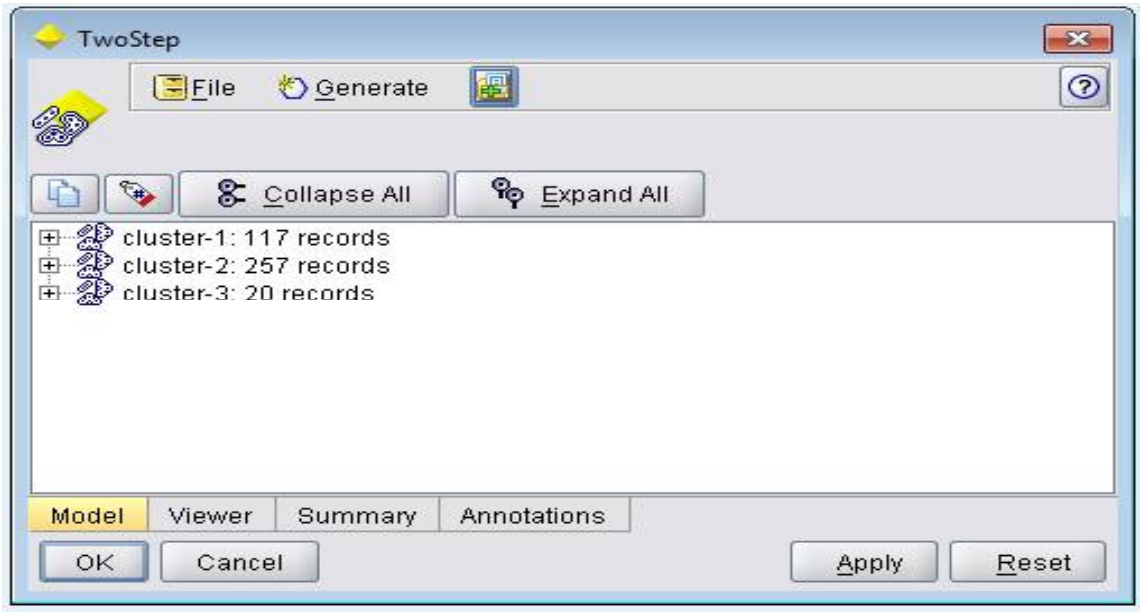


Tablo 9: Model 6 için Seçilen Değişkenlerin Önem Seviyeleri, Ortalamaları ve Standart Sapmaları Tablosu

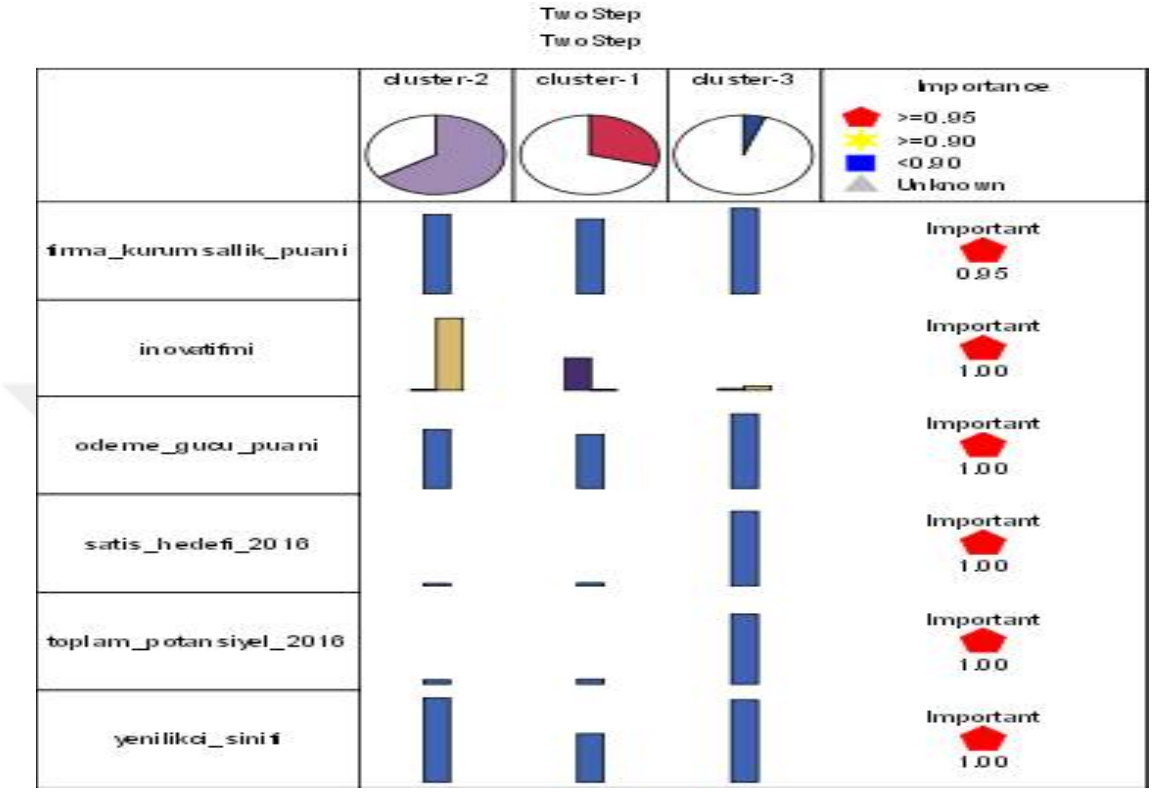
DEĞİŞKENLER	KÜME 1			KÜME 2		
	Ö.S	Ort	SS	Ö.S	Ort	SS
toplam_potansiyel_2016	0.00	0.01	0.02	1.00	0.09	0.14
satis_hedefi_2016	0.00	0.01	0.02	1.00	0.10	0.17
rakip_firma_gucu	0.38	0.60	0.38	0.68	0.63	0.38
rekabet_marka	0.00	0.14	0.23	1.00	0.87	0.23
rekabet_fiyatlama	1.00	0.41	0.49	0.00	0.01	0.09
rekabet_kalite	0.00	0.03	0.12	1.00	0.97	0.15
rekabet_hizmet	0.00	0.34	0.47	1.00	0.60	0.23

Küme 1 için fiyatlama değişkeni, Küme 2 için ise fiyatlama ve rakip_firma_gucu dışındaki değişkenler oldukça önemlidir. Burada sonuçlar Model 1 ile neredeyse aynıdır. Model 1'in farkı firma yapısı değişkenlerinin de eklenmiş olmasıdır. Ancak bu değişkenler çıkartıldığında büyük farklılıklar görülmediği için müşterileri rekabet odaklı kümelemenin daha iyi sonuçlar vereceği gözlemlenmektedir. PSO ile karşılaştırıldığında da neredeyse aynı sonuçlar çıkmaktadır.

Model 7:



Şekil de görüldüğü üzere firma yapısı ve iş potansiyeline dayalı modelde küme sayısı 3 olarak belirlenmiştir. Birinci kümeye müşterilerin %30'u (117), ikinci kümeye %65'i (257), üçüncü kümeye ise %5'i (20) atanmıştır.



Tablo 10: Model 7 için Seçilen Değişkenlerin Önem Seviyeleri, Ortalamaları ve Standart Sapmaları Tablosu

DEĞİŞKENLER	**KÜME 1**			**KÜME 2**			**KÜME 3**		
	Ö.S	Ort	SS	Ö.S	Ort	SS	Ö.S	Ort	SS
toplam_potansiyel_2016	0.00	0.02	0.04	0.00	0.02	0.03	1.00	0.33	0.19
satis_hedefi_2016	0.00	0.02	0.04	0.00	0.02	0.04	1.00	0.44	0.17
inovatifmi	1.00	-	-	1.00	-	-	0.44	-	-
yenilikci_sinifi	0.00	0.34	0.45	1.00	0.59	0.36	0.75	0.58	0.41
firma_kurumsallik_puani	0.06	0.70	0.23	0.80	0.74	0.19	0.99	0.80	0.11
odeme_gucu_puani	0.03	0.63	0.29	0.71	0.69	0.25	1.00	0.87	0.21

Küme 1 için inovatiflik değişkeni, Küme 2 için hem inovatiflik hem de hangi sınıfa girdiğiyle alakalı yenilikçi sınıfı değişkeni önemlidir. Küme 3 için ise bu iki değişken dışındakiler önem arz etmektedir. PSO çözümüyle neredeyse aynı çözümleri vermektedir. Satış politikası yenilik tabanlı oluşturulmalıdır.

Aşağıdaki tabloda 3 yönteme göre elde edilen 21 adet analiz sonucunda ortaya çıkan küme sayıları ve kümelere atanan müşteri sayıları yer almaktadır.

Tablo 11: Modellere İlişkin Analiz Sonuçları

		PSO	k-ortalamalar	İki Aşamalı
Model 1	Küme Sayısı	3	3	2
	Müşteri Sayıları	79-224-91	126-110-158	264-130
Model 2	Küme Sayısı	5	5	4
	Müşteri Sayıları	41-121-48-62-122	128-77-57-62-70	124-87-121-62
Model 3	Küme Sayısı	3	3	2
	Müşteri Sayıları	107-256-31	208-12-64	264-130
Model 4	Küme Sayısı	2	2	2
	Müşteri Sayıları	383-11	393-1	31-363
Model 5	Küme Sayısı	2	2	3
	Müşteri Sayıları	138-256	139-255	91-124-179
Model 6	Küme Sayısı	2	2	2
	Müşteri Sayıları	272-122	284-110	267-127
Model 7	Küme Sayısı	3	3	3
	Müşteri Sayıları	122-234-38	122-234-38	117-257-20

SONUÇ

Yapılan çalışmada, parçacık sürü optimizasyonu ile kümeleme algoritması, veri madenciliği yöntemlerinden k-ortalamalar ve iki aşamalı kümeleme yöntemleri kullanılarak bir firmanın en uygun küme sayısını belirlemek ve belirlenen kümeler doğrultusunda anlamlı sonuçlar çıkarmak amaçlanmıştır. Firmada kurumsallığın sağlanması için optimum satış politikaları oluşturmak amacıyla bu kümeleme sonuçlarından yararlanılabildiği görülmüştür.

Firma ile çalışan müşteriler hakkında 11 değişken belirlenmiş, bu değişkenler 394 müşteri için firmanın satış mühendisi tarafından oluşturulmuştur. Elde edilen veriler Excel programında Min-Maks normalizasyonu yöntemi ile normalize edilerek kullanılabilir hale getirilmiş, daha sonra SPSS Clementine 10. 1 ve MATLAB 2016a programlarında 7 farklı model için çalıştırılmış ve küme sayıları, kümelere ayrılan müşteri sayıları ve hangi müşterinin hangi kümeye ait olduğu gibi bilgiler elde edilmiştir. Böylece elde ettiği bilgiler doğrultusunda firma, hangi müşteri kümesine nasıl bir yaklaşım uygulayacağına kolayca karar verebilmektedir

Aynı zamanda kümelemede kullanılan yöntemlerin hangi açıdan daha iyi sonuç verdiği bilgisine de ulaşılabilmektedir. PSO kümeleme algoritması, genel olarak diğer iki yönteme göre kümeleri daha doğru ayırmaktadır. Bazı modellerde İki Aşamalı Kümeleme Yöntemi, PSO'da bulunan ayrı özelliklere sahip iki kümeyi birleştirerek, bir küme oluşturmaktadır. Bu durum müşteri özelliklerini tam olarak ayrıştıramaması nedeniyle daha iyi sonuç çıkmasını engellemektedir. İki aşamalı ve k-ortalamalar algoritmalarının avantajı ise her değişkenin önem seviyesini vermesidir. Çıktıları yorumlamak daha az zaman aldığı için daha kolay hale gelmektedir.

Yapılan analizlere genel olarak bakıldığında firmaları rekabet odaklı kümelemek, rekabetin hangi değişkenden kaynaklandığının bilinmesi ve satış politikalarının buna göre düzenlenmesi durumunda müşteri memnuniyetinin ve satışların artacağı sonucu çıkmaktadır. Kümeleme işlemi firma satış mühendisi tarafından oldukça yararlı bulunmuş ve hayata geçirme kararı alınmıştır.

Bu konuda farklı değişkenler seçilerek, farklı modeller geliştirilebilir. Kümeleme sonucunda çıkan değerlendirmelere uyularak uygulanan satış politikaları

sonrasında kümelemenin firma açısından faydalı olup olmadığını gösteren, karşılaştırmanın yapıldığı yeni bir çalışma ortaya atılabilir.

KAYNAKÇA

Aher, G.J. ve Metre, V.A. (2014). Clustering Multidimensional Data with PSO based Algorithm. *Third Post Graduate Symposium on Computer Engineering,* 28-29 March, University of Pune, India.

Ahi, L. (2015). *Veri Madenciliği Yöntemleri İle Ana Harcama Gruplarının Paylarının Tahmini* (Yayınlanmamış Yüksek Lisans Tezi). Ankara: Hacettepe Üniversitesi İstatistik Anabilim Dalı.

Akçay, A. (2014). *Bilgi ve Belge Yönetiminde Veri Madenciliği.* (Yayınlanmamış Yüksek Lisans Tezi). İstanbul: İstanbul Üniversitesi Sosyal Bilimler Enstitüsü.

Akman, M. (2010). *Veri Madenciliğine Genel Bakış ve Random Forests Yönteminin İncelenmesi: Sağlık Alanında Bir Uygulama.* (Yayınlanmamış Yüksek Lisans Tezi). Ankara: Ankara Üniversitesi Sağlık Bilimleri Enstitüsü.

Akpınar, H. (2000). Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği. *İ.Ü. İşletme Fakültesi Dergisi,* 29(1): 1-22.

Aksaraylı, M. ve Pala, O. (2017). Uzaktan Eğitimi Tercih Etme Nedenleri Ve Başarı Arasındaki İlişkinin Kümeleme Analizi İle İncelenmesi, *Batı Anadolu Eğitim Bilimleri Dergisi*, 8(2), 37-48.

Alagöz, A. ve Kutlu, M. (2012). Parçacık Sürü Optimizasyonu Yaklaşımı İle Emtia Piyasasında Portföy Optimizasyonu. *Selçuk Üniversitesi Sosyal ve Ekonomik Araştırmalar Dergisi,* 23: 35-50.

Alam, S., Dobbie, G. ve Rehman, S.U. (2015). Analysis of Particle Swarm Optimization Based Hierarchical Data Clustering Approaches. *Swarm and Evolutionary Computation*, 25: 36–51.

Albayrak, S. (2009). *Aykırı Değer Tespitinde Yoğunluk Tabanlı Kümeleme Yöntemleri.* (Yayınlanmamış Yüksek Lisans Tezi). İstanbul: Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü.

Albayrak, S. ve Yılmaz, Ş.K. (2009). Veri Madenciliği: Karar Ağacı Algoritmaları ve İMKB Verileri Üzerine Bir Uygulama. *Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi,* 14(1): 31-52.

Alzand, H. ve Karacan, H. (2014). Bölümleyici Kümeleme Algoritmalarının Farklı Veri Yoğunluklarında Karşılaştırılması. *Erciyes Üniversitesi Fen Bilimleri Enstitüsü Dergisi,* 30(1): 56-62.

Alswaitti, M., Albughdadi, M. ve Mat Isa, N. A. (2018). Density-Based Particle Swarm Optimization Algorithm For Data Clustering. *Expert Systems With Applications,* 91: 170-186.

Arıkan, (2017). *Farklı Kümeleme Teknikleri İle Ülkelerin İnsani Gelişmişlik Endeksi Verilerinin Karşılaştırılmalı Analizi.* (Yayınlanmamış Yüksek Lisans Tezi). Marmara Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.

Altinoz, O. T., Yilmaz, A. E., & Weber, G. W. (2010). Chaos particle swarm optimized PID controller for the inverted pendulum system. *In 2nd international conference on engineering optimization.*

Armano, G. ve Framani, M. R. (2016), Multiobjective Clustering Analysis Using Particle Swarm Optimization. *Expert Systems With Applications*, 55, 184–193.

Aydoğan, E. ve Gencer, C. (2007). Kaba Küme Yaklaşımı Kullanılarak Veri Madenciliği Problemlerinde Sınıflandırma Amaçlı Yapılmış Olan Çalışmalar. *Savunma Bilimleri Dergisi,* 6(2): 17-32.

Bacher, J., Wenzig, K ve Vogler, M. (2004). *SPSS TwoStep Cluster - A First Evaluation*. Arbeits- und Diskussionspapiere, Nürnberg.

Badem, M. (2007). *Genetik Algoritmaların Yaratıcı Mimari Tasarımda Kullanımı*. (Yayınlanmamış Yüksek Lisans Tezi). İstanbul: İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü.

Bayer, H. (2011). *Veri Madenciliğinde Bir Metin Madenciliği Uygulaması*. (Yayınlanmamış Yüksek Lisans Tezi). İstanbul: Beykent Üniversitesi Fen Bilimleri Enstitüsü.

Binay, H.S. (2002). *Yatırım Kararlarında Kaba Küme Yaklaşımı*. (Yayınlanmamış Doktora Tezi). Ankara: Ankara Üniversitesi Sosyal Bilimler Enstitüsü.

Bircan H. ve Çam, S. (2016). Veri Madenciliğinde Kümeleme Analizi ve Sağlık Sektöründe Bir Uygulaması. *C.Ü. İktisadi ve İdari Bilimler Dergisi*, 17(2): 85-96.

Blondin, J. (2009). *Particle Swarm Optimization,* A tutorial.

Bora, D.J., Gupta, A.K. ve Khan, F.A. (2015). Comparing the Performance of L*A*B* and HSV Color Spaces with Respect to Color Image Segmentation. *International Journal of Emerging Technology and Advanced Engineering*, 5(2): 192-203.

Ceylan, Z., Gürsev, S. ve Bulkan, S. (2017). İki Aşamalı Kümeleme Analizi ile Bireysel Emeklilik Sektöründe Müşteri Profilinin Değerlendirilmesi. *Bilişim Teknolojileri Dergisi,* 10(4): 475-485.

Cura, T. (2012). A particle swarm optimization approach to clustering. *Expert Systems with Applications*, *39*(1), 1582-1588.

Çalışkan, F., Yüksel, H. ve Dayık, M. (2016). Genetik Algoritmaların Tasarım Sürecinde Kullanılması. *Teknik Bilimler Dergisi*, 6(2): 21-27.

Çalışkan, S.K. ve Soğukpınar, İ. (2008). KxKNN: K-Means ve K En Yakın Komşu Yöntemleri İle Ağlarda Nüfuz Tespiti. *2. Ağ ve Bilgi Güvenliği Sempozyumu*, 16-18 Mayıs, Girne, 120-124.

Çavuşlu, M., Karakuzu, C. ve Şahin, S. (2010). Parçacık Sürü Optimizasyonu Algoritması ile Yapay Sinir Ağı Eğitiminin FPGA Üzerinde Donanımsal Gerçeklenmesi. *Politeknik Dergisi*, 13(2): 83-92.

Çevik, K. ve Koçer, H. (2013). Parçacık Sürü Optimizasyonu ile Yapay Sinir Ağları Eğitimine Dayalı Bir Esnek Hesaplama Uygulaması. *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi,* 17(2): 39-45.

Demiralay, M. ve Çamurcu, A.Y. (2005). Cure, Agnes ve K-Means Algoritmalarındaki Kümeleme Yeteneklerinin Karşılaştırılması. *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi,* 4(8), 1-18.

Doğan, H., Ayas, S., Gedikli, E. ve Ekinci, M. (2018). Parçacık Sürü Zekâsı Optimizasyonu ile Mikroskobik Görüntülerin Segmentasyonunda Farklı Entropi Ölçülerinin Etkisi. *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, DOI: 10.19113/sdufbed.64012.

Dunham, M.H. (2003). *Data Mining Introductory and Advanced Topics*. New Jersey: Pearson Education Inc.

Edelstein, H. A. (1999). *Introduction to Data Mining and Knowledge Discovery.* Maryland: Two Crows Corporation.

Erdoğmuş, P. ve Yalcin, E. (2015). Parçacık Sürü Optimizasyonu İle Kısıtsız Optimizasyon Test Problemlerinin Çözümü. *İleri Teknoloji Bilimleri Dergisi*, *4*(1).

Ergüden, Y. ve Erşahin, B. (2008). *Veri Madenciliği.* İstanbul: Arge Danışmanlık.

Everitt, B., Landau, S. ve Leese, M. (2011). *Cluster Analysis*. London: John Wiley & Sons.

Fırat, M., Dikbaş, F., Koç, A.C. ve Güngör, M. (2012). Classification of Annual Precipitations and Identification of Homogeneous Regions using K-Means Method. *Digest,* 1609-1622.

Getoor, L. (2003). *Link Mining: A New Data Mining Challenge*, ACM SIGKDD Explorations Newsletter, No:1.

Giray, S. (2016). İki Aşamalı Kümeleme Analizi İle Hükümlü Verilerinin İncelenmesi. *Ekonometri ve İstatistik,* 25: 1-31.

Han, J., Kamber, M. ve Pei, J. (2012), *Data Mining: Concepts and Techniques*. New York: Morgan Kaufmann Publisher.

Hatamlou, A. (2013). Black hole: A New Heuristic Optimization Approach For Data Clustering. *Information Sciences*, 222: 175–184.

Hosmer, D. W. ve Lemeshow, S. (2000). *Applied Logistic Regression* (Second Edition). Canada: John Wiley & Sons.

Izakian, Z., Mesgari, M.S. ve Abraham, A. (2016). Automated Clustering of Trajectory Data Using A Particle Swarm Optimization. *Computers, Environment and Urban Systems,* 55: 55-65.

İşler, Y. ve Narin, A. (2012). WEKA Yazılımında k-Ortalama Algoritması Kullanılarak Konjestif Kalp Yetmezliği Hastalarının Teşhisi. *Süleyman Demirel Üniversitesi Teknik Bilimler Dergisi*, 2(4): 21-29.

Jiang, M.F., Tseng, S.S. ve Su, C.M. (2001). Two-phase clustering process for outliers detection. *Pattern Recognition Letters*, 22: 691-700.

Kennedy, J. ve Eberhart, R. (1995). Particle swarm optimization. *Proceedings of the IEEE International Conference on Neural Networks*, 4: 1942-1948.

Keskender, M. ve Keskender, E. (2017). Geçmişten Günümüze Yapay Sinir Ağları ve Tarihçesi. *Takvim-i Vakayi,* 5(2): 8-18.

Khan, A., Bawane, N. G. ve Bodkhe, S. (2010). An Analysis of Particle Swarm Optimization with Data Clustering-Technique for Optimization in Data Mining. *International Journal on Computer Science and Engineering*, 2(7): 2223-2226.

Koyuncugil, A. S. (2007). *Veri Madenciliği ve Sermaye Piyasalarına Uygulaması.* Sermaye Piyasası Kurulu Araştırma Raporu, Ankara.

Larose, D.T. (2005). *Discovering Knowledge in Data: An Introduction to Data Mining.* Canada: John Wiley & Sons Inc.

Li, G. ve Sun, L. (2018). Characterizing Heterogeneity in Drivers' Merging Maneuvers Using Two-Step Cluster Analysis. *Hindawi Journal of Advanced Transportation*, Article ID 5604375, https://doi.org/10.1155/2018/5604375.

Norusis, M.J. (2011). Cluster Analysis (Chapter 17). *IBM SPSS Statistics 19 Statistical Procedures Companion.* Prentice Hall. pp.361-391.

Omran, M. G., Salman, A., & Engelbrecht, A. P. (2006). Dynamic clustering using particle swarm optimization with application in image segmentation. *Pattern Analysis and Applications*, *8*(4), 332.

Oğuzlar, A. (2003). Veri Ön İşleme. *Erciyes Üniversitesi İ.İ.B.F. Dergisi,* 21: 67-76.

Ortakçı, Y. ve Göloğlu, C. (2012). Parçacık Sürü Optimizasyonu İle Küme Sayısının Belirlenmesi. *Akademik Bilişim,* 335–341.

Özdemir, A. ve Orçanlı, K. (2012). İki Aşamalı Kümeleme Algoritması ile Pazar Bölümlemesi, Müşteri Profillerinin Belirlenmesi ve Niş Pazarların Tespiti. *Uşak Üniversitesi Sosyal Bilimler Dergisi,* 5(3): 1-27.

Özdemir, A., Aslay, F.Y. ve Çam, H. (2010). Veri Tabanında Bilgi Keşfi Süreci: Gümüşhane Devlet Hastanesi Uygulaması. *SÜ İİBF Sosyal ve Ekonomik Araştırmalar Dergisi,* 20: 347-365.

Özkan, H. (2013). *K-means kümeleme ve K-NN Sınıflandırma Algoritmalarının Öğrenci Notları ve Hastalık Verilerine Uygulanması.* İstanbul Teknik Üniversitesi Fen Edebiyat Fakültesi, İstanbul.

Özmen, Ş. (2001). İş Hayatı Veri Madenciliği İle İstatistik Uygulamalarını Yeniden Keşfediyor, *V. Ulusal Ekonometri ve İstatistik Sempozyumu*, Çukurova Üniversitesi, Adana.

Özmen, Ş. (2003). Veri Madenciliği Süreci. *Veri Madenciliği ve Uygulama Alanları Konferansı,* İstanbul Ticaret Üniversitesi, İstanbul.

Özmen, M., Delice, Y., & Aydoğan, E. K. (2018). Telekomünikasyon Sektöründe PSO ile Müşteri Bölümlenmesi. *International Journal of Informatics Technologies*, 11(2).

Özsağlam, M. ve Çunkaş, M. (2008). Optimizasyon Problemlerinin Çözümü için Parçaçık Sürü Optimizasyonu Algoritması. *Politeknik Dergisi,* 11(4): 299-305.

Pala, O., ve Aksaraylı, M. (2017) Parçacık Sürü Optimizasyonu Ve Yüksek Dereceden Moment Tabanlı Sayısal Kısıtlı Portföy Seçimi. *II. Uluslararası Stratejik Araştırmalar Kongresi, Bildiri Kitabı* (ss.614-620), Düzenleyen Celal Bayar Üniversitesi. Antalya. 28 Eylül – 1 Ekim 2017.

Pawlak, Z. ve Skowron, A. (1994). Rough Set Rudiments. The International Workshop on Rough Sets and Soft Computing, San Jose. California.

Richards, M. ve Ventura, D. (2004). Choosing a Starting Configuration for Particle Swarm Optimization. *IEEE International Joint Conference on Neural Networks*, 3: 2309-2312.

Shelokar, P. S., Jayaraman, V. K., & Kulkarni, B. D. (2004). An ant colony approach for clustering. *Analytica Chimica Acta*, 509, 187–195.

Stepaniuk, J. R. (2008). *Granular Computing in Knowledge Discovery and Data Mining*. New York: Springer.

Sumathi, S. ve Sivanandam, S. (2006). *Introduction to Data Mining and its Applications*. New York: Springer.

Sun, J. ve Li, H. (2008). Data Mining Method for Listed Companies, Financial Distress Prediction. *Knowledge-Based Systems*, 21(1): 1-5.

Şchiopu, D. (2010). Applying TwoStep Cluster Analysis for Identifying Bank Customers' Profile. *Buletinul,* 62(3): 66-75.

Şekeroğlu, S. (2010). *Hizmet Sektöründe Bir Veri Madenciliği Uygulaması*. (Yayınlanmamış Yüksek Lisans Tezi). İstanbul: İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü.

Şen, F. (2008). *Veri Madenciliği İle Birliktelik Kurallarının Bulunması.* (Yayınlanmamış Yüksek Lisans Tezi). Sakarya: Sakarya Üniversitesi Fen Bilimleri Enstitüsü.

Şentürk, A. (2006). *Veri Madenciliği Kavram ve Teknikler*. Bursa: Ekin Kitabevi.

Şık, M. (2014). *Veri Madenciliği ve Kanser Erken Teşhisinde Kullanımı.* (Yayınlanmamış Yüksek Lisans Tezi). Malatya: İnönü Üniversitesi Sosyal Bilimler Enstitüsü.

Tripathi, P. K., Bandyopadhyay, S., & Pal, S. K. (2007). Multi-objective particle swarm optimization with time variant inertia and acceleration coefficients. *Information Sciences*, 177(22), 5033-5049.

Trpkova, M. ve Tevdovski, D. (2009). Twostep cluster analysis: Segmentation of largest companies in Macedonia. *Challenges for Analysis of the Economy, the Businesses, and Social Progress International Scientific Conference*, November 19-21, Szeged.

Tüzüntürk, S. (2010). Veri Madenciliği ve İstatistik. *Uludağ Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi,* 29(1): 73.

Tzortzis, G. ve Likas, A. (2014). The MinMax k-means clustering algorithm. *Pattern Recognition*, 47: 2505–2516.

Wagstaff, K., Cardie, C., Rogers, S. ve Schroedl, S. (2001). Constrained K-means Clustering with Background Knowledge. *Proceedings of the Eighteenth International Conference on Machine Learning*, 577-584.

Wan, S. (2013). Entropy-Based Particle Swarm Optimization With Clustering Analysis On Landslide Susceptibility Mapping. *Environ Earth Sci*, 68: 1349-1366.

Wang, W., Yang, J. ve Muntz, R. (1997). STING: A statistical information grid approach to spatial data mining, *Very Large Data Bases*, 186-195.

Zhang, C., Ouyang, D., & Ning, J. (2010). An artificial bee colony approach for clustering. *Expert Systems with Applications*, 37, 4761–4767.