KOCAELİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

ENDÜSTRİ MÜHENDİSLİĞİ ANABİLİM DALI

YÜKSEK LİSANS TEZİ

HİYERARŞİK KÜMELEME TEKNİKLERİNDE KÜME ELEMAN SAYISININ EŞİTLENMESİNE YÖNELİK BİR YAKLAŞIM ÖNERİSİ VE GERÇEK KARAYOLU UZAKLIK VERİLERİNE DAYALI KÜMELEME ANALİZİ

AKİF TAŞATAN

KOCAELİ 2018

KOCAELİ ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

ENDÜSTRİ MÜHENDİSLİĞİ ANABİLİM DALI

YÜKSEK LİSANS TEZİ

HİYERARŞİK KÜMELEME TEKNİKLERİNDE KÜME ELEMAN SAYISININ EŞİTLENMESİNE YÖNELİK BİR YAKLAŞIM ÖNERİSİ VE GERÇEK KARAYOLU UZAKLIK VERİLERİNE DAYALI KÜMELEME ANALİZİ

AKİF TAŞATAN

Doç.Dr. Kasım BAYNAL
Danışman, Kocaeli Üniv. .....

Doç.Dr. Gülşen AYDIN KESKİN
Jüri Üyesi, Kocaeli Üniv. .....

Doç.Dr. Semra BORAN
Jüri Üyesi, Sakarya Üniv. .....

Tezin Savunulduğu Tarih: 03.07.2018

ÖNSÖZ VE TEŞEKKÜR

Geleceğe yön verenler bilgiyi yöneten toplumlardır. Teknolojik gelişmeler hayatımıza her gün yeni imkanlar sunmaktadır. Aynı zamanda küresel rekabet işleri en etkin zaman planı ile yapmayı zorunlu kılmaktadır. Doğru planlama ve büyük veri yönetimi için bilgi teknolojilerinden mümkün olduğunca faydalanılması gerekmektedir. Bu çalışma da lokasyon bazlı büyük projelerde sıkça karşılaşılan ve yanlış yapılması durumunda ciddi zararlara neden olan gruplandırma/kümeleme konusunu ele almakta ve var olan tekniklerden ikisinin geliştirilerek lokasyon bazlı kümeleme problemlerini çözebilecek bilgisayar destekli yeni bir yaklaşım ortaya koymaktadır.

Bu tez çalışmasında, desteğini, rehberliğini ve zamanını esirgemeden beni sabırla dinleyen, bana yol gösteren ve engin bilgi ve tecrübelerinden yararlandığım danışman hocam Doç. Dr. Kasım BAYNAL'a, desteğini hep yanımda hissettiğim, beni her konuda destekleyen ve tecrübeleri ile yönlendiren sayın Dr. R. Reha ÇETİN'e, programlama konusunda her türlü isteğime cevap veren, vaktinden fedakarlık ederek emeklerini esirgemeyen değerli arkadaşım Yavuz SEVGİ'ye teşekkür ederim. Hayatım boyunca varlıklarını hep yanımda hissettiğim, çalışmalarım sırasında beni hep destekleyen sevgili aileme de sonsuz teşekkür ederim.

Haziran – 2018 Akif TAŞATAN

İÇİNDEKİLER

ŞEKİLLER DİZİNİ

TABLOLAR DİZİNİ

SİMGELER VE KISALTMALAR DİZİNİ

σ : Standart Sapma
R : Değişim Genişliği

Kısaltmalar

VM : Veri Madenciliği
CL : Clustering Limited Programı
CLC : Complete Link Clustering (Tam Bağlantı Tekniği)
CLC-Kn : Complete Link Clustering-Kn Algorithm (Tam Bağlantı-Kn Tekniği)
DAM : Data Matrix (Veri Matrisi)
DD : Decimal-Degrees (Derece-Ondalık)
DIM : Dissimilarity Matrix (Uzaklık Matrisi)
DM : Degrees:Minute-Decimal (Derece:Dakika-Ondalık)
DMS : Degrees:Minute:Seconds (Derece:Dakika:Saniye)
GCS : Geographical Coordinate System (Coğrafi Koordinat Sistemi)
GIS : Geographical Information System (Coğrafi Bilgi Sistemi)
GPS : Global Positioning System (Global Konumlama Sistemi)
HKT : Hata Kareler Toplamı
LAT : Latitude (Enlem)
LONG : Longitude (Boylam)
SLC : Single Link Clustering (Tek Bağlantı Tekniği)
SLC-Kn : Single Link Clustering-Kn Algorithm (Tek Bağlantı-Kn Tekniği)
TSP : Traveling Salesman Problem (Gezgin Satıcı Problemi)
VM : Veri Madenciliği
VTBK : Veri Tabanlarında Bilgi Keşfi
YSA : Yapay Sinir Ağları

HİYERARŞİK KÜMELEME TEKNİKLERİNDE KÜME ELEMAN SAYISININ EŞİTLENMESİNE YÖNELİK BİR YAKLAŞIM ÖNERİSİ VE GERÇEK KARAYOLU UZAKLIK VERİLERİNE DAYALI KÜMELEME ANALİZİ

ÖZET

Endüstri alanında son yıllarda kaydedilen ilerlemeler hız, verimlilik ve kazanç kavramlarını daha da önemli hale getirmiştir. Gerek başarılı iş sonuçları gerekse ekonomik istikrar açısından işgücü, malzeme ve zaman gibi kaynakların itinalı kullanımı bir zaruret halini almıştır. Doğru yapılmayan iş planları, projelerin gecikmesine ve kaynak israfına neden olmaktadır. Bu noktada özellikle sahada ekipler tarafından günlük iş planları doğrultusunda yapılması gereken işlerde ciddi bir optimizasyon gereksinimi ortaya çıkmaktadır. Belirli bir zaman diliminde tamamlanması gereken ve geniş bir bölgeye yayılmış proje bazlı işlerde plansız hareket edildiğinde veya doğru planlama yapılmadığında projelerin gecikmesine neden olan en büyük unsurun ekiplerin uygulama noktaları arasında karayolu ulaşımında kaybettikleri zaman olduğu görülmektedir.

Kümeleme analizi, veri madenciliğinde önemli bir yere sahip olup birçok alanda kullanılmakta ve çeşitli tekniklerle yürütülmektedir. Bu tekniklerden yaygın olarak kullanılan hiyerarşik kümeleme teknikleri nesnelerin kümelenmesi konusunda sonuçlar üretmekte ancak eşit sayıda eleman içeren kümeler oluşturulmasına olanak sağlamamaktadır. Bu tezde hiyerarşik kümeleme tekniklerinden Tek Bağlantı Tekniği ile Tam Bağlantı Tekniği üzerinde geliştirmeler yapılarak yeni teknikler türetilmesi ve bunun gerçek karayolu uzaklık verilerine dayalı kümeleme yapılması istenen bir saha projesinde uygulaması ele alınmıştır. Geliştirilen tekniklerin başarılı bir şekilde amaca hizmet ettiği gözlenmiştir. Geliştirilen bu yaklaşımın, karayolu ulaşım mesafelerinin minimize edilmesi istenen benzer kümeleme projelerinde uygulanabilir olduğu değerlendirilmiştir.

Anahtar Kelimeler: Hiyerarşik Kümeleme, Karayolu Uzaklık, Metrik Uzaklık, Optimizasyon, Saha Projeleri.

AN APPROACH PROPOSAL FOR EQUALIZATION OF THE NUMBER OF CLUSTER ELEMENTS AT HIERARCHICAL CLUSTERING TECHNIQUES AND A CLUSTER ANALYSIS BASED ON REAL HIGHWAY DISTANCE DATA

ABSTRACT

The progress experienced recently in the industry has made the concepts of speed, productivity and profit even more important. Careful use of sources such as labor force, materials and time has become a necessity in terms of both successful business outputs and economic stability. Incorrect business plans cause delays in the projects and waste of resources. At this point a serious optimization need arises, especially for the tasks which should be carried out through daily plans by teams on site. It is seen that the major factor causing projects to be delayed when acting without any plans or making incorrect planning for the project-based works that should be completed in a certain period of time on a wide area is the time lost during commuting of the teams in the highway between the implementation points.

Clustering analysis has an important place in data mining, used in many areas and conducted with various techniques. Hierarchical clustering techniques commonly used among these techniques generate results on clustering of objects but do not allow the creation of clusters with an equal number of elements. In this study, new algorithms are derived from hierarchical clustering techniques by making improvements on Single Link Clustering and Complete Link Clustering and its application in a field project where clustering based on real highway distance data is required. It has been observed that the developed algorithms serve successfully to the purpose. It has been assessed that this developed approach is feasible in similar clustering projects which require minimization of highway transport distances.

Keywords: Hierarchical Clustering, Highway Distance, Metric Distance, Optimization, Field Projects.

GİRİŞ

Çevredeki nesneleri bazı özelliklerine göre çeşitli gruplara ayırma eğilimi insanoğlunun varoluşundan beri süregelmektedir. En genel tanımıyla bu ayırma işlemi benzer olanı benzemezden ayrı tutmaktır. Zamana bağlı olarak artan birim sayısı neticesinde gruplara ayırma işi daha zorlu bir iş haline gelmiş ve bu durum olayın amacını ve kapsamını değiştirmiştir. Bu durum gruplara ayırma işlemi için yeni bazı tekniklerin geliştirilmesi ihtiyacını doğurmuştur. Bunun neticesinde de kümeleme analizi kavramı ortaya çıkmıştır. Kavram 1960 yıllar ve sonrasında veri madenciliğinde yaşanan hızlı gelişmelerle çok farklı alanlardaki uygulamalarda kullanılmaya başlanmıştır. Özellikle bilgi teknolojilerindeki gelişimle büyük verinin yönetimi mümkün hale gelmiş, kümeleme analizi de bu alanda önemli bir yer almıştır. Kümeleme analizinin amacı bir grup nesnenin benzer niteliklerde olanların bir arada bulunacağı şekilde alt gruplara ayrılmasıdır. Analizde hakkında kesin bilgi bulunmayan nesnelerin kümelere ayrılması temel işlev olmasına karşın sonuçta kaç küme elde edileceği, kümelerde ne kadar nesne bulunacağı da üzerinde durulması gereken araştırma konularıdır.

Çalışmanın birinci bölümünde, Veri Madenciliği ve Veri Ambarı kavramları, bunların farkları, Veri Madenciliğinde Bilgi Keşfi süreci, Veri Madenciliğinde kullanılan yöntemler ve Veri Madenciliğinin uygulama alanları üzerinde durulmaktadır.

İkinci bölümde, Kümeleme Analizi kavramı, Kümeleme Analizinde anahtar olan uzaklık ölçüleri ve kümeleme teknikleri detaylı olarak anlatılmaktadır.

Üçüncü bölümde Coğrafi Koordinat Sistemi ve Coğrafi Bilgi Sistemi ile ilgili tanımlar yapılarak geniş bir alanda uygulanan bir saha projesindeki kümeleme problemi üzerinde Hiyerarşik Kümeleme Teknikleri ile yapılan bir uygulama üzerinde durulacaktır.

Son bölümde ise elde edilen bulgular ışığında çalışmanın genel bir değerlendirmesi yapılarak ve çeşitli önerilerde bulunulmaktadır.

Bu çalışmanın amacı saha projelerinde lokasyon bazlı gruplama gereksinimi olan durumlarda projenin en kısa sürede tamamlanabilmesi için birbirlerine en yakın olan noktaların bir arada olacak şekilde eşit elemanlı kümelenmesini sağlamak üzere Hiyerarşik Kümeleme Teknikleri üzerinde geliştirme yapılarak yeni bir yaklaşım önerisi ortaya konulması ve bunun gerçek hayatta uygulanabilirliğinin gösterilmesidir.

1. VERİ MADENCİLİĞİ

1.1. Veri Madenciliğine Genel Bakış

Bilgisayarların insanoğlunun hayatına girmesiyle teknolojik gelişmeler bu alanda hızla ilerlemiş, her alanda oluşan değerli verilerin depolanması ihtiyacı bilgisayarlar aracılığıyla karşılanmaya başlamıştır. Depolanan veriler de "Veritabanı" kavramını ortaya çıkarmıştır. Günümüzde neredeyse tüm organizasyonlar gelişen bilgisayar sistemleri ve internet erişiminin yaygınlaşmasıyla kayıtlarını elektronik ortama yani veritabanlarına aktarmışlardır. Artık veritabanlarının boyutları 'terabayt'lar ile ifade edilmektedir. Zamanla veritabanlarında bulunan verilerin işe yarar örüntüler ve şablonlar elde etmede kullanılabileceği düşüncesi yaygınlaşmıştır [1].

Veri Madenciliği (VM) kavramı da 60'lı yıllarda bilgisayarların veri analizi amacıyla kullanılmaya başlanmasıyla birlikte ortaya çıkmıştır. İlk zamanlarda veritabanlarında yeterli taramalar yapılarak istenen bilgiye ulaşılmasının mümkün olacağı düşünülmüştür. Bu işleme veri taraması (data dredging), veri yakalaması (data fishing) gibi isimler verilmiştir. 90'lı yıllarda ise VM kavramı bilgisayar mühendisleri tarafından ortaya atılmıştır. Bilgisayar mühendisleri veri analizlerinin klasik istatistiksel modeller yerine, çeşitli algoritmalara dayanan bilgisayar modelleri ile yapılabileceğini öne sürmüşlerdir. Sonrasında VM'nde çeşitli yaklaşımlar ortaya konulmaya başlanmıştır. Bu yaklaşımların temelinde makine öğrenimi (machine learning), istatistik, otomasyon, veritabanları, pazarlama, araştırma gibi disiplinler bulunmaktadır [2].

VM basit bir ifadeyle büyük veri yığınları arasından gereken bilgiye ulaşma işidir. Başka bir ifadeyle bilgisayar yardımıyla büyük veritabanları içerisinden gelecekle ilgili tahminler yürütülmesini sağlayacak ilişkilerin araştırılmasıdır.

Yukarıda belirtilen ifadeye ilave olarak VM kavramı için yapılmış tanımlardan bazıları şu şekildedir;

VM, bilgi keşfi sürecinde bir adımdır ve kapsamında örüntüleri ortaya çıkarmak için kullanılan bazı algoritmalar bulunur. Ortaya çıkarılan bilgi sonrasında bir öngörü veya sınıflandırma modeli kurmak, eğilimleri ve ilişkileri belirlemek, mevcut bir modeli yenilemek veya üzerinde önceden madencilik yapılmış bir veritabanının özetini çıkarmak için kullanılabilir [3].

VM, önceden bilinmeyen, geçerli ve etkin bilginin veritabanlarından çekilmesi ve sonrasında bu bilginin iş kararlarını almak için kullanılmasını çevreleyen bir süreçtir [4].

VM'nin amacı, mevcut veri içerisindeki geçerli, sıradışı, kullanışlı ve anlaşılabilir korelasyonları ve örüntüleri tespit etmektir [5].

VM, çok büyük boyutlardaki veri kümesinden şirketlerin daha iyi kararlar almalarına yardımcı olan ve piyasada rekabetçi yapılarını sürdürmelerini sağlayabilecek ilginç bilgilere ulaşma sürecidir [6].

VM, anlamlı örüntüleri ve ilişkileri tespit etmek için büyük miktardaki veriyi, otomatik ya da yarı otomatik tekniklerle araştırma ve inceleme sürecidir [7].

VM, eldeki veriler içerisinden belirgin olmayan, önceden bilinmeyen fakat kullanılma potansiyeli olan bilginin elde edilmesi sürecidir. Bu da; kümeleme, veri özetleme, değişiklik analizi, sapma tespiti gibi belirli teknikleri kapsar [8].

VM büyük miktardaki verinin analizi sonucu anlamlı bilgiler ve kurallar keşfi maksadıyla geliştirilen bir yaklaşımdır [43].

1.2. Veri Ambarı Kavramı

Veri Ambarı terimi ilk kez William H. Inmon tarafından 1991 yılında ortaya atılmıştır. Temel amacı yönetimin kararlarının desteklenmesi amacıyla çeşitli kaynaklardan elde edilen bilgilerin zaman değişkeni kullanılarak veri toplama işidir. Çok özet ifadeyle veri ambarları, birçok veritabanından alınarak bir araya getirilen verilerin depolandığı alanlardır. En önemli özelliği, kullanıcılara veriler hakkında farklı detay katmanları sunabilmesidir. Alt düzeyler saklanan kayıtlar ile ilgili iken üst düzeyler zaman gibi bilginin elde edilmesi ile ilgili olan detayları içerir.

Veritabanlarına göre çok daha kapsamlı olan veri ambarlarının oluşturulması ve uygulamaya konulması bir yılı aşkın bir zaman dilimi gerektirmekte ve tesis edilmesi için büyük yatırım kaynağına ihtiyaç duymaktadır [9].

Bill Inmon ise veri ambarını yönetimin karar sürecine destek olarak kullanılan konuya özgü, entegre, zamana bağlı ve kalıcı veri topluluğu olarak tanımlamıştır [10].

Veri ambarlarının üç çeşidi bulunmaktadır:

1. Tüm kuruma hizmet eden kurumsal (geleneksel) veri ambarı,
2. İşletmedeki belirli bir iş birimini veya bölümü desteklemek üzere tasarlanmış küçük bir veri ambarı olan veri pazarı (data mart),
3. Veri ambarı tekniklerinin hareket sistemlerine uyarlandığı operasyonel veri deposu [11].

Bir veri ambarı birbiri ile ilişkili bazı alt bileşenlerden meydana gelir. Bunlar [12];

- Operasyonel Veritabanı / Harici Veritabanı Katmanı
- Enformasyon Ulaşım Katmanı
- Veri Ulaşım Katmanı
- Metaveri
- İşlem Yönetim Katmanı
- Uygulama Haberleşmesi Katmanı
- Veri Ambarı Katmanı
- Veri Sunum Katmanı

şeklindedir.

Veri ambarı kavramı, karar vermede kullanılabilecek yapısal kaliteli bilgiye kolay erişimi sağlama ihtiyacından ortaya çıkmıştır. Veri ambarları, karar verme ve çözümleme amacıyla kullanılacak olan kaliteli veriye kolayca erişmek için kurulmaktadır.

Rekabetin her geçen gün arttığı iş dünyasında bilginin organizasyonlara önemli avantajlar sağladığı görülmektedir. Birçok organizasyon büyük miktarda veriye sahip olmasına rağmen, sürekli artan veri hacmi nedeniyle bu verilere erişim ve onların kullanılması gitgide zorlaşmaktadır.

Veri ambarları farklı düzlemlerdeki veri kaynaklarına erişerek veriyi temizleyip, süzüp değiştirdikten sonra, anlaşılabilir ve kolay erişilebilir bir yapıda muhafaza eder. Bu veri, daha sonra sorgulama, raporlama ve veri çözümlemede kullanılır.

1.3. Veritabanı ile Veri Ambarı Kavramlarının Farkları

Veri ambarları ile günlük hayatta kullanılan veritabanları karşılaştırıldığında aşağıda belirtilen farklılıklar dikkati çekmektedir;

- Veritabanında yer alan veri, bir süzme işleminin ardından veri ambarına aktarılır.
- Veritabanı üzerinde bulunan veri çok taze, veri ambarındaki ise eskidir.
- Veri ambarı özet bilgileri içerebilir; veritabanındaki veri ise içermez.
- Bütünlüğü sağlamak adına verinin önemli bir kısmı belirli bir işlemden sonra veri ambarına aktarılır.

Veritabanlarının genellikle günlük işlemlerde sıklıkla kullanılması dikkat çekmektedir. Veritabanlarında, günlük veri giriş-çıkışı çok fazla olmaktadır. Veritabanlarının genel kullanıcıları şirketlerde çalışan personellerdir (Mali İşler, Üretim, İnsan Kaynakları vb.). Veri ambarlarının kullanıcıları ise sıklıkla analistler ve şirket yönetiminde karar alabilme yetkisine sahip kişilerdir. Veri ambarı kullanılarak yapılan bir çalışmaya örnek şu olabilir; müşteri ilişkileri yönetimi ile ilgili genellemeler yapmak suretiyle bir müşteri profili ortaya çıkarılarak üretilen mal veya hizmetin bu profile uygun olacak şekilde değiştirilmesi; ya da satın alma tercihi, satın alma zamanı, harcama arzuları gibi müşterinin satın alma biçimlerini inceleyerek müşteri ürüne veya hizmete olan ilgisini arttırmak. Dolayısıyla buradan yapılan çıkarımla veri ambarlarının kullanıcı sayısı, veritabanlarının kullanıcı sayısına göre çok daha azdır [13].

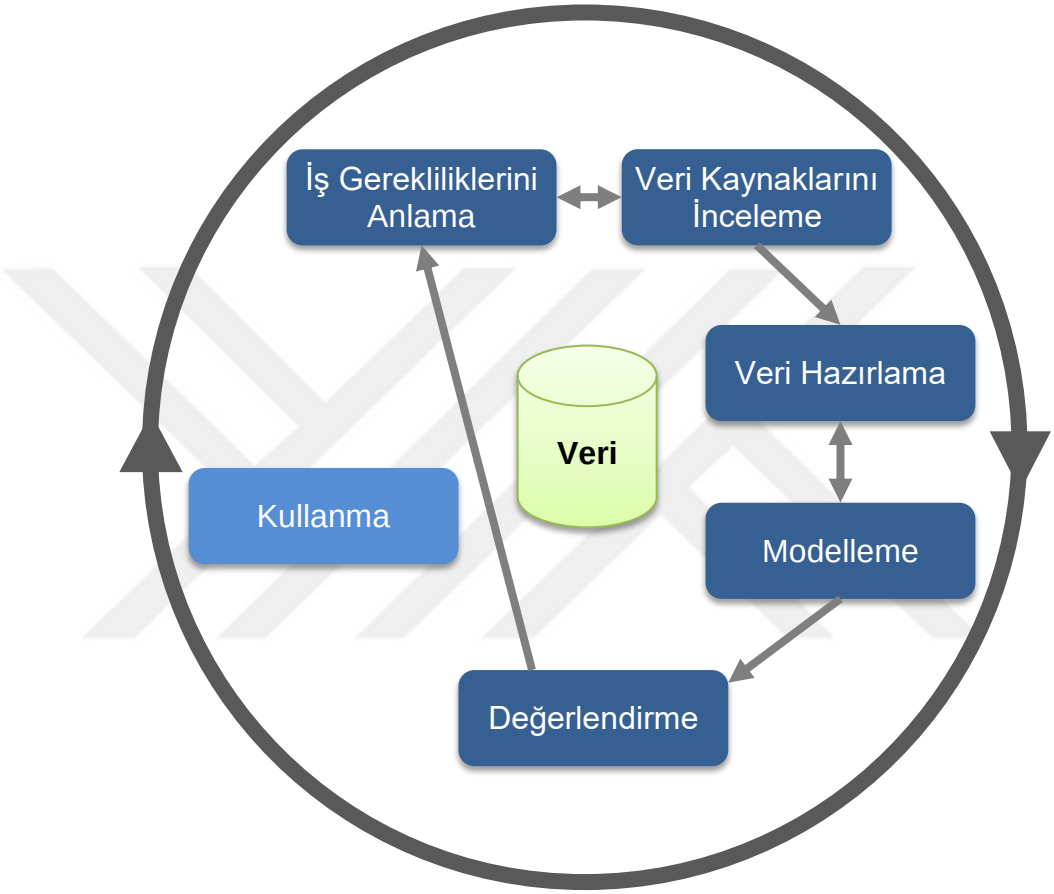
1.4. Veri Madenciliğinde Bilgi Keşfi Süreci

Bir VM'nin fayda sağlayabilmesi ancak üzerinde araştırma yapılan işin ve verilerin özelliklerinin bilinmesi durumunda mümkün olmaktadır. Bu nedenden dolayı öncelikle iş ve veri özelliklerinin tanımlanması gerekmektedir. Başarılı bir VM projesi için şu adımlar izlenmelidir [14];

1. Problemin tanımlanması
2. Verilerin hazırlanması

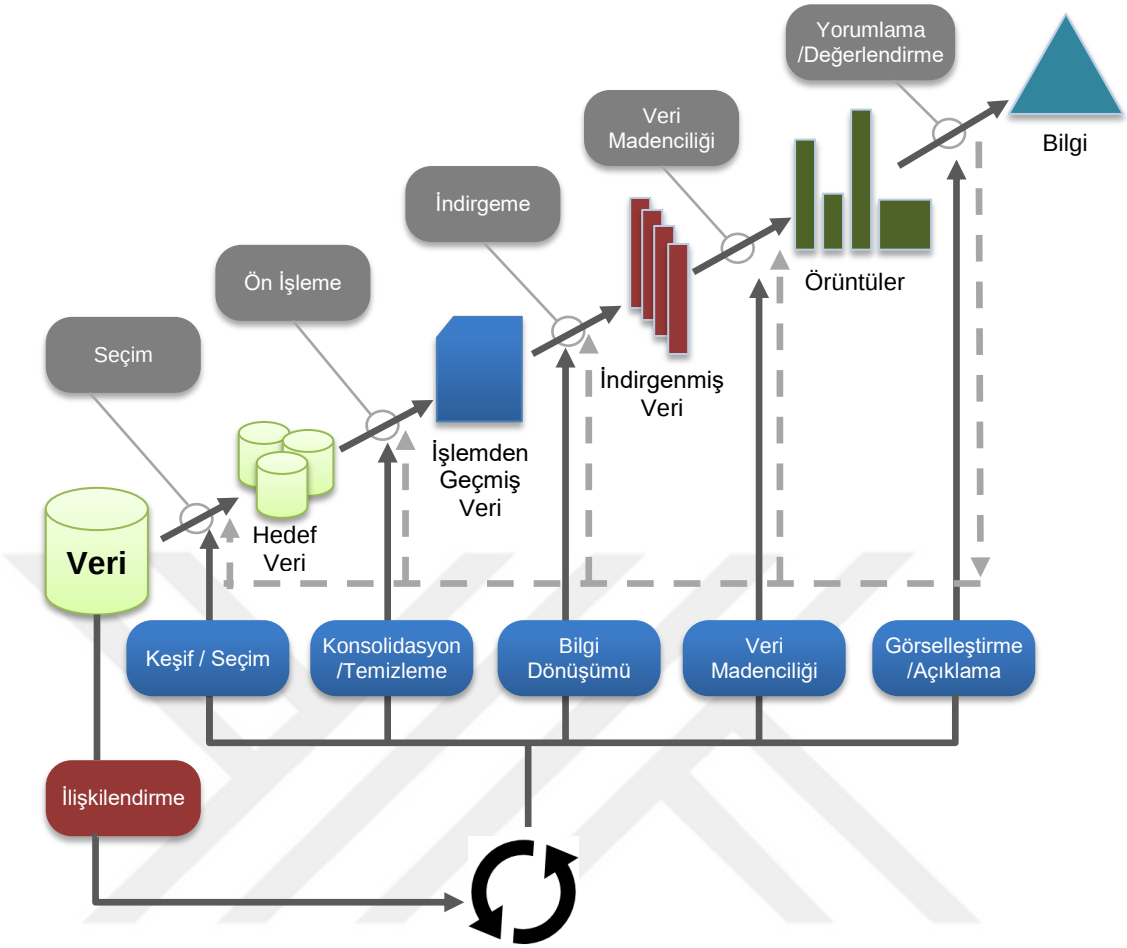
3. Modelin kurulması ve değerlendirilmesi
4. Modelin kullanılması
5. Modelin izlenmesi

VM referans modeli Şekil 1.1'de gösterilmiştir.



Şekil 1.1. Veri madenciliği referans modeli [15]

Veritabanı sistemlerinin yaygınlaşması ve toplanan verilerin büyük boyutlara ulaşması, organizasyonlarda toplanan bu verilerin nasıl değerlendirileceği konusunu gündeme getirmiştir. Klasik sorgulama veya raporlama araçlarının veri yığınları karşısında yetersiz kalması bir takım yeni arayışlara neden olmaktadır. Buna Veri Tabanlarında Bilgi Keşfi (VTBK) süreci denilmektedir. Bu süreç içerisinde, modelleme ve değerlendirme aşamalarından meydana gelen VM çalışmaları en önemli kesimi oluşturmaktadır. VM'ne verilen bu önem neticesinde VTBK ile VM kavramları bazı araştırmacılar tarafından eş anlamlı olarak kullanılmaktadır. Şekil 1.2'de Bilgi keşfi sürecinde VM'nin yeri grafiksel olarak ifade edilmiştir.



Şekil 1.2. Bilgi keşfi sürecinde VM'nin yeri [16]

1.5. Veri Madenciliğinde Kullanılan Teknikler

Veri madenciliğinde kullanılan teknikler dört grupta toplanabilir. Bunlar [1];

1. İstatistiksel Yöntemler: VM esası itibariyle bir istatistik uygulamasıdır. Verilen örnek veri kümesine bir tahminleme oturtmayı hedefler. İstatistik literatüründe son elli yılda bu amaç için değişik teknikler önerilmiştir. Bunlar istatistik literatüründe çok boyutlu analiz (multivariate analysis) kategorisi altında yer alır ve genelde verinin parametrik bir modelden (sıklıkla da çok boyutlu bir Gauss dağılımından) geldiğini varsayar. Bu varsayım altında istatistikte uzun yıllardır kullanılan teknikler şunlardır; sınıflandırma, kümeleme, regresyon, hipotez testi, varyans analizi, boyut azaltma, bağıntı kurma.

2. Bellek Tabanlı Yöntemler: Bellek tabanlı veya örnek tabanlı yöntemler istatistikte 1950'li yıllarda önerilmiş olmasına karşın o tarihlerde bilgisayar gibi büyük

hesaplamalar yapabilecek cihazlar bulunmamasından ötürü kullanılamamıştır. Ancak günümüzde bilgisayarların oldukça yaygınlaşması ve karmaşık işlemleri yapabilir hale gelmesiyle bu teknikler kullanılabilir duruma gelmiştir. K En Yakın Komşu Algoritması bu tür yöntemlere verilecek en iyi örneklerden birisidir.

3. Yapay Sinir Ağları: 1980'lerden sonra yaygınlaşan Yapay Sinir Ağları (YSA) tekniğinde kullanılan öğrenme algoritmaları veriden üniteler arasındaki bağlantı ağırlıklarını hesaplar. Diğer bellek tabanlı teknikler gibi yüksek işlem ve bilgisayar belleği gerektirmeyen YSA, veri hakkında parametrik bir model varsaymaz. Daha açık bir ifadeyle YSA örneklerle ilgili bilgiler toplamakta, bunlarla ilgili genellemeler yapmakta ve sonrasında hiç karşılaşmadığı örnekler ile karşılaşınca öğrendiği bilgileri kullanarak o örnekler hakkında karar verebilmektedir.

4. Karar Ağaçları: İstatistiksel yöntemlerde veya YSA'nda veriden bir fonksiyon elde edildikten sonra bu fonksiyonun insanlar tarafından anlaşılabilecek bir kural olarak yorumlanması zordur. Karar ağaçlarında bundan biraz daha farklı olarak veri oluşturulduktan sonra kökten uca doğru gidilerek bazı kurallar yazılabilir. Bu kural çıkarsama işi, VM çalışmasının sonucunun geçerli olmasını sağlar. Ortaya konan kurallar uygulama konusunda uzman bir kişiye sunularak çıkan sonucun anlamlı olup olmadığını tespit etmek için test edilebilir. Sonradan başka bir teknik uygulanacak olsa bile ilk olarak karar ağacı ile kısa bir ön çalışma yapmak, önemli değişkenler ve yaklaşık kurallar konusunda fikir verebilir.

1.6. Veri Madenciliğinin Uygulama Alanları

VM gelişmiş ülkelerde yoğun olarak kullanılmaktadır. Ülkemizde de VM'nin uygulama alanları gün geçtikçe artmaktadır. Özellikle büyük ölçekli firmalarda organizasyon etkinliğinin ve iş sonuçları başarı oranının artırılmasında VM'ne yönelik bazı yazılımlar kullanılmaktadır. Aşağıdaki alt başlıklarda VM'nin sektörel bazda kullanıldığı alanlar ve kullanılma amaçlarına yönelik örneklere yer verilmiştir [1, 20].

Bankacılık;

- Farklı finansal göstergeler arasında gizli korelasyonların bulunması
- Kredi kartı dolandırıcılıklarının tespiti

- Kredi kartı harcamalarına göre müşteri gruplarının belirlenmesi
- Kredi taleplerinin değerlendirilmesi
- Risk yönetimi
- Firma derecelendirme
- Faiz oranlarının tahmini
- Borçlanma ve iflas tahminleri
- Müşterilerin bölümlenmesi
- Müşteri kârlılığının analizi
- Döviz kurlarının tahmini.

Pazarlama;

- Müşterilerin satın alma örüntülerinin belirlenmesi
- Müşterilerin demografik özellikleri arasındaki bağlantıların bulunması
- Posta kampanyalarında cevap verme oranının artırılması
- Mevcut müşterilerin elde tutulması, yeni müşterilerin kazanılması
- Pazar sepeti analizi
- Müşteri ilişkileri yönetimi
- Müşteri değerlendirme
- Satış tahmini
- Kampanya ürünlerini belirleme
- Müşteri değerleme
- Müşteri ilişkileri yönetimi
- Satış tahminleri
- Ürünlerin rafa yerleştirilmesi
- Satış ve mevsimsel farklar arasındaki örüntülerin tespit edilmesi.

Sigortacılık;

- Yeni poliçe talep edecek müşterilerin tahmin edilmesi
- Sigorta dolandırıcılıklarının tespiti
- Riskli müşteri örüntülerinin belirlenmesi
- Müşteri kaybı sebeplerinin belirlenmesi
- Usulsüzlüklerin önlenmesi
- Ana giderlerin azaltılması
- Poliçe fiyatlarının belirlenmesi.

Borsa;

- Hisse senedi fiyat tahmini
- Genel piyasa analizleri
- Alım-satım stratejilerinin uygunluğu
- Portföy yönetimi.

Telekomünikasyon;

- Kalite ve iyileştirme analizleri
- Hile tespitleri
- Hatların yoğunluk tahminleri
- Müşteri kazanma ve elde tutma analizleri.

İnternet;

- Metin madenciliği
- Web pazarlama
- Arama motorları.

Üretim;

- Envanter kontrolü
- Donanım arızası analizi
- Kaynak yönetimi
- Süreç / kalite kontrol
- Kapasite yönetimi
- Lojistik uygulamalar
- Üretim süreçlerinin uygunluğu.

Sağlık ve İlaç;

- Test sonuçlarının tahmini
- Ürün geliştirme
- Tıbbi teşhis
- Tedavi sürecinin belirlenmesi
- Yeni ilaç türlerinin keşfi ve sınıflandırılması.

Bilim ve Mühendislik;

- Ampirik veriler üzerinde modeller kurarak bilimsel ve teknik problemlerin çözümlenmesi
- Yeni virüs türlerinin keşfi ve sınıflandırılması
- Gen haritasının analizi ve genetik hastalıkların tespiti,
- Kanserli hücrelerin tespiti
- Gezegen yüzey şekillerinin, gezegen yerleşimlerinin ve yeni galaksilerin keşfi.

1.7. Veri Madenciliği Modelleri

VM'nde kullanılan modeller, tahmin edici (Predictive) ve tanımlayıcı (Descriptive) olmak üzere iki kategoride ele alınmaktadır. Tahmin edici modeller, sonuçları bilinen verilerden hareket edilerek geliştirilen modelin daha sonra sonuçları bilinmeyen veri gruplarına uygulanarak sonuçlarının tahmin edilmesi üzerine kurgulanmıştır. Bu modellere örnek olarak banka kredi uygulamaları verilebilir. Banka önceki dönemlerde müşterilerine vermiş olduğu kredilere ait verilere sahiptir. Bu veriler arasında bağımsız değişkenler müşteri özellikleri, bağımlı değişken ise kredilerin zamanında ödenip ödenmediği bilgisidir. Bu veri grubuna uygun olarak kurulan model, bankaya daha sonra gelen kredi taleplerinde müşteri özelliklerine göre tahsis edilecek kredinin geri ödenip ödenmeyeceğinin tahmininde kullanılmaktadır.

Tanımlayıcı modeller ise karar vermeye rehberlik etmede kullanılabilecek bazı yardımcı örüntülerin tanımlanmasıdır. Yine banka örneğinden hareketle 25 yaş altı bekâr kişiler ile 25 yaş üstü evli kişiler üzerinde yapılan ve ödeme performanslarını gösteren bir analiz, tanımlayıcı modellere örnek olarak gösterilebilir [1].

VM modelleri, gördükleri işlevlere göre üç ana başlık altında incelenmektedir. Bunlar [1];

- Sınıflama (Classification) ve Regresyon Analizi (Regression Analysis)
- Kümeleme Analizi (Clustering Analysis)
- Birliktelik Kuralları (Association Rules) ve Ardışık Zamanlı Örüntüler (Sequential Patterns)

Sınıflama ve regresyon analizleri tahmin edici, kümeleme analizi ve birliktelik kuralları ile ardışık zamanlı örüntü modelleri ise tanımlayıcı modellerdir.

1.7.1. Sınıflama ve regresyon analizi

Sınıflama yaygın olarak kullanılan VM yöntemlerinden biridir. Resim tanıma, örüntü tanıma, hastalık tanıları, kalite kontrol, pazarlama ve dolandırıcılık tespiti sınıflama konularının sıklıkla kullanıldığı alanlardır. Sınıflama, tahminleyici model başlığı altında yer almaktadır [17]. Satışlarını arttırmak için yeni bir kampanya yapmak isteyen firmanın önceden satış yapmış olduğu müşterilerinin verilerini kullanarak yeni kampanyaya katılma olasılığı bulunan potansiyel müşterilerini tahmin etmeye çalışması bir sınıflama yöntemidir.

Sınıflama, her bir yeni gözlemin daha önceden tanımlanmış olan veri gruplarından hangisinin altında yer alacağının kestiriminde kullanılır. Daha açık bir ifadeyle verileri, içerdikleri ortak özelliklere göre ayrıştırır. Regresyon analizinde ise gözlenen olay değerlendirilirken bunun diğer hangi olaylardan etkilenildiğini belirlemek esastır. Regresyonda amaç girdiler ile çıktılar arasındaki ilişkiyi kurmada en doğru tahmini yapabilecek modeli oluşturmaktır [18].

Sınıflama analizinde kullanılan başlıca teknikler;

- Karar Ağaçları
- Yapay Sinir Ağları
- Genetik Algoritmalar
- K-En Yakın Komşu Algoritması
- Regresyon
- Bayes Sınıflandırması.

Sınıflama analizinin uygulandığı bazı alanlar [17];

- İmzaya duyarlı belge işlemlerinde ve bankalarda (Uyuyor, uymuyor)
- Parmak izi uygulamalarında (Uyuyor, uymuyor)
- Kredi başvuru değerlendirmelerinde (İyi, kötü)
- Bankaların müşteri notu değerlendirmelerinde (İyi, vasat, zayıf)
- Görüntü kümesinden kimlik tespit etme (Dost, düşman)
- İlaçların tedavideki etkinliği (İyi, orta, zayıf).

İki adımlı bir süreç olan Sınıflamanın ilk adımı veritabanında bulunan değişkenlere uygun bir modelin belirlenmesidir. Modelin kurgulanması için verilerin bir kısmı eğitim verileri olarak kullanılır. Rassal olarak seçilen bu verilere algoritma uygulanır ve sınıflama modeline ulaşılır. İkinci adım ise ulaşılan modelin sınıflama için kullanılmasıdır.

1.7.2. Kümeleme analizi

Kümeleme, veritabanındaki verilerin benzer özelliklerine göre gruplanmasıdır. Oluşturulan kümelerde bulunan verilerin benzerliklerinin fazla olması amaçlanır. Bir diğer ifadeyle kümeler arası benzerliğin de mümkün olduğunca az olacağından söz edilebilir. Kümeleme analizinde sınıflandırma analizinden farklı olarak verilerin hangi gruplara ayrılacağı önceden belli değildir.

Yine sınıflama analizinden farklı olarak kümeleme tekniğinde, veriler sadece kendi değerlerine göre değil, diğer veriler ile olan ilişkilerine yani birbirlerine olan yakınlıklarına veya uzaklıklarına göre değerlendirilmektedir. Bu nedenden ötürü kümeleme analizi dinamik bir yöntem olarak addedilir [19].

Kümeleme analizi bu çalışmanın ikinci bölümünde detaylı olarak ele alınmaktadır.

1.7.3. Birliktelik kuralları ve ardışık zamanlı örüntüler

Büyük veri kümeleri arasındaki ilişkiler birliktelik kurallarının çalışma alanıdır. Toplanan ve depolanan verinin gün geçtikçe büyük boyutlara ulaşmasından dolayı, organizasyonlar veritabanlarındaki birliktelik kurallarını ortaya çıkarmak istemektedirler. Ortaya çıkarılan bu kurallar organizasyonlara karar alma süreçlerinde destek olurlar. Birliktelik kurallarına verilebilecek en genel örnek market sepeti uygulamasıdır (Market Basket Analysis). Bu uygulama, müşterilerin yaptıkları alışverişlerdeki ürünler arasındaki ilişkileri tanımlayarak müşterilerin satın alma alışkanlıklarını inceler. Buradaki birlikteliklerin keşfi, müşterilerin hangi ürünleri bir arada aldıklarını ortaya çıkarır ve bu doğrultuda daha etkili satış stratejileri geliştirebilir. Örneğin bir marketin müşterilerinin kahve ile şekeri birlikte satın alma oranı yüksek ise, işletme yöneticileri bu iki ürünün raflarını yan yana konumlandırarak satışlarını artırabilirler [1].

Birliktelik kuralları uygulamasında iki kriter bulunmaktadır. Bu kriterler güven (confidence) ve destek (support) ölçütleridir. Bilgilerin büyük hacimli veri yığınları içerisinden ayıklanarak anlamlı bilgilerin keşfedilmesinde bu iki kriter kullanılmaktadır. B olayı gerçekleştiği sırada A olayının da gerçekleşme olasılığı "Güven" olarak ifade edilir. Kesişen olayların ortaya çıkma sayısının toplam olay sayısı içerisinde oranı ise "Destek" olarak ifade edilir. İki değer de ne kadar yüksek olursa iki ürünün satın alınması bağıntısı o kadar yüksek olarak değerlendirilir. Appriori algoritması birliktelik kurallarında en bilinen ve en yaygın kullanılan algoritmadır [17].

Ardışık zamanlı örüntüler ise birbirleri ile ilişkili olup birbirini izleyen dönemlerde gerçekleşen ilişkilerin tanımlanmasında kullanılır. Bu örüntülere örnek olarak "çekiç satın alan bir kişinin izleyen dönemde %20 olasılıkla çivi de satın alacağı"; "X operasyonunu geçiren kimsenin Y süre içerisinde %30 olasılıkla Z enfeksiyonuna yakalanacağı" ifadeleri verilebilir [1].

2. KÜMELEME ANALİZİ

Nesneleri veya bir veri kümesindeki bilgileri saptanan yakınlık kriterlerine göre gruplama işlemine "Kümeleme Analizi" adı verilmektedir. Oluşan kümeler arasındaki benzerlik az, küme içi benzerlik fazla olmalıdır. Kümeleme analizinde sınıflar önceden belli olmayıp nesnelerin veya verilerin kaç faklı kümeye ayrılacağı, hangi kümeye ait olacağı mevcut nesnelerin/verilerin birbirlerine olan benzerliğine göre belirlenir [21].

Kümeleme analizi üzerine yapılmış ilk çalışmanın Linnaeus'un 1753 yılında hayvanların ve bitkilerin sınıflandırılması konulu çalışması olduğu varsayılır [22]. Kümeleme analizi veri madenciliğinin en önemli süreçlerinden biri olarak değerlendirilir. İstatistiksel işlemleri bünyesinde barındıran kümeleme analizi birçok disiplinde yapılan çalışmalara konu olmuştur ve olmaya da devam etmektedir. Bu yönüyle disiplinler arası bir yaklaşım olarak görülür.

Kümeleme analizinde neyin değişken olarak seçileceği oldukça önemlidir. Kümeler oluşturulurken kullanılan değişkenlere göre kümelere atama yapılacağından, seçilen değişkenlerin araştırma konusu ile tamamen uyumlu olması gerekir. Aksi takdirde, oluşan kümelerde güven düzeyi düşük olacaktır ve yanlış kümelere atama yapılması ile sonuçları ile karşı karşıya kalınabilmektedir. Analizin devamında gelecek diğer çalışmaların da zincirleme olarak bu durumdan etkilenmesi muhtemeldir. Değişkenlerin seçimi kümeleme analizinin en büyük zorluklarından biri olarak kabul edilir.

Kümeleme analizinin avantajları şu şekilde ifade edilebilir [23];

- İlişkilerin görüntülenmesi: Kümeleme analizinin en önemli özelliklerinden biri sonuçların grafik gösterim olarak verilmesidir. Görsel sonuçlardan benzerlikler kolay tespit edilir.
- Anormalliklerin tespiti: Grafiklerden aykırı değerler kolaylıkla tespit edilir, bu sayede aykırı değerler ile ilgili aksiyonlar alınabilir.

- Diğer veri madenciliği teknikleri için ön hazırlık: Karar ağaçları gibi bazı teknikler daha küçük veri yığınları üzerinde çalışabilir. Kümeleme analizi ile büyük veri yığınları daha küçük gruplara ayrıldığından bu tür tekniklerin kullanılabilmesi için zemin hazırlanır.

Bununla birlikte kümeleme analizinin bazı zayıf yönleri de vardır. Bunlar [23];

- Sonuçların anlaşılması zordur: Takip edilmesi gereken belirli kurallar olmadığı için tahminler belli bir hata payı ile yapılır.
- Farklı veri tiplerinde özellikler içeren nesnelerin karşılaştırılması zordur.

Kümeleme analizinde ilk öncelik, değişken seçimi ve amaca uygun uzaklık ölçüsünün belirlenmesidir. Seçilecek değişken, kümeleme analizinin en önemli noktalarından biridir. Kümeleme analizinde metrik veriler ile çalışılırken uzaklık ölçüleri, metrik olmayan yani nitelik belirten veriler ile çalışılırken benzerlik ve korelasyon ölçüleri kullanılmaktadır. Kümeleme analizinde veri yapısı matris formundadır. Bu matrisler [24];

a. Veri Matrisi (Data Matrix):

Bu matris n adet birim için p adet özelliğin tanımlandığı satırlar ile sütunların birleşmesinden oluşan (nxp) boyutunda bir matristir. Örnekle açıklamak gerekirse bir şehirde yaşayan insanların yaş, cinsiyet, boy, ağırlık, gibi özellikleri alt alta yazıldığında Şekil 2.1'de görünen yapıda bir matris oluşur. Bu matriste her bir sütun bir niteliği, her bir satır ise niteliklere karşı gelen değerleri içermektedir.

$$DAM = \begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

Şekil 2.1. Veri matrisi (DAM)

b. Uzaklık Matrisi (Dissimilarity Matrix):

Birimlerin birbirlerine olan uzaklık bilgilerinin yer aldığı (nxn) boyutunda aynı zamanda farklılık ve benzemezlik matrisi olarak da anılan kare matristir. Bu matrisin genel şekli aşağıda verilmiştir (Şekil 2.2).

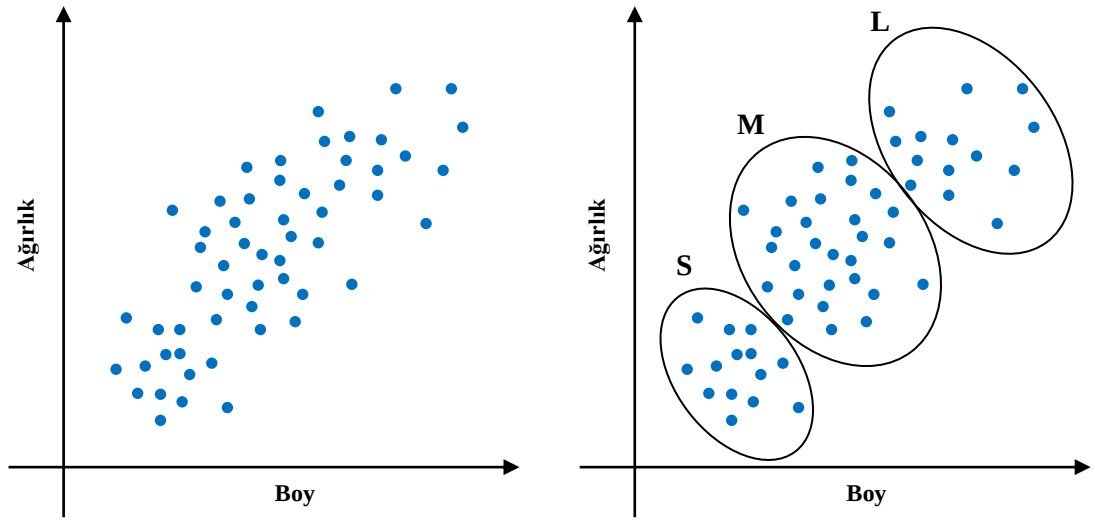
$$DIM = \begin{bmatrix} 0 & d(1,2) & d(1,3) & \dots & d(1,n) \\ d(2,1) & 0 & d(2,3) & \dots & d(2,n) \\ d(3,1) & d(3,2) & 0 & \dots & d(3,n) \\ \dots & \dots & \dots & \dots & \dots \\ d(n,1) & d(n,2) & d(n,3) & \dots & 0 \end{bmatrix}$$

Şekil 2.2. Uzaklık matrisi (DIM)

Birimler karşılıklı uzaklıkları birbirine eşit olduğu için matrisin asal köşegenleri altında kalan değerler üstünde kalan değerler ile simetriktir. Asal köşegenleri ise sıfır değerindedir. Bu sebeple bu matrise tek yönlü matris adı verilir ve yalnızca asal köşegen ve altında kalan elemanları içerir. Veri matrisi ise iki yönlü bir matristir. Uzaklık matrisinde verilerin birbirlerine olan uzaklık (mesafe) değerleri bulunmaktadır. Bu matristeki değerler uzaklık ve benzerlik ölçülerine göre farklı anlamlar içermektedir. Daha açık bir ifadeyle ilişkileri uzaklık ölçüleri kullanılarak yapılan çalışmalarda verilerin birbirleri ile aralarındaki uzaklıklar en az olacak şekilde kümeleme gerçekleşmektedir. Uzaklık değerinin düşük olması ilişkinin güçlü olduğu anlamına gelmektedir. Benzerlik veya korelasyon değerleri baz alınarak yapılan çalışmalarda ise verilerin birbirleri ile aralarındaki ilişkiyi gösteren ilişki değerleri yüksek olanlar aynı kümelerde yer alacaktır. Bu matriste ise değerinin yüksek olması güçlü ilişkiyi gösterir.

2.1. Uzaklık Ölçüleri

Uzaklık ve benzerlik ölçüleri tek boyutlu ya da çok boyutlu olabilmektedir. Her boyut verileri kümelemek/gruplandırmak için gereken anahtarı ifade eder. Örneğin kıyafetler beden ölçüleri gruplandırılmak istendiğinde dikkate alınabilecek ölçüler kişilerin ağırlıkları ve boylarıdır. Bu konuya ilişkin kümeleme işlemi örneği aşağıda gösterilmiştir (Şekil 2.3).



Şekil 2.3. Kümeleme işlemi örnek gösterimi [42]

İki ya da üç boyutlu uzayda iki nokta ya da birim arasındaki uzaklığı hesaplamada en çok Öklid Uzaklığı kullanılır. Bu ölçüm, en temel haliyle uzayda iki birimin arasındaki geometrik uzaklık olarak tanımlanır. Uzaklık fonksiyonunun genel özellikleri ise şöyledir [39];

- $d(i,j) \geq 0$; Uzaklık negatif olamaz
- $d(i,i)=0$; Her birimin kendisine olan uzaklığı sıfırdır
- $d(i,j)= d(j,i)$; Uzaklık fonksiyonu simetriktir
- $d(i,j) \leq d(i,h)+d(h,j)$; İki birimin arasındaki uzaklık bu iki birimin üçüncü bir birime olan uzaklıkları toplamından küçük olamaz (üçgen eşitsizliği).

Uzaklığın ölçülmesine yönelik teknikler üç başlık atında toplanmıştır. Bunlar [24];

1. Aralık Ölçekli Değişkenler
2. İkili Değişkenler
3. Ordinal Değişkenler

2.1.1. Aralık ölçekli değişkenler

Aralık ölçekli değişkenler doğrusal bir ölçek üzerinde gösterilebilen değişkenlerdir. Bunlara örnek olarak ağırlık, genişlik ve hava sıcaklığı gibi değişkenler verilebilir. Aralık ölçekli değişkenlerde dikkat edilmesi gereken en önemli husus tüm verilerin aynı birim ile ifade edilmesi gerekliliğidir. Aksi bir durum kümeleme işleminin başarısız olmasına yol açacaktır.

En yaygın kullanılan uzaklık ölçüleri Öklid, Minkowski, Karesel Öklid, Manhattan, Mahalanobis ve Chebyschev uzaklıklarıdır ve bunlara ait bilgilere aşağıda yer verilmiştir.

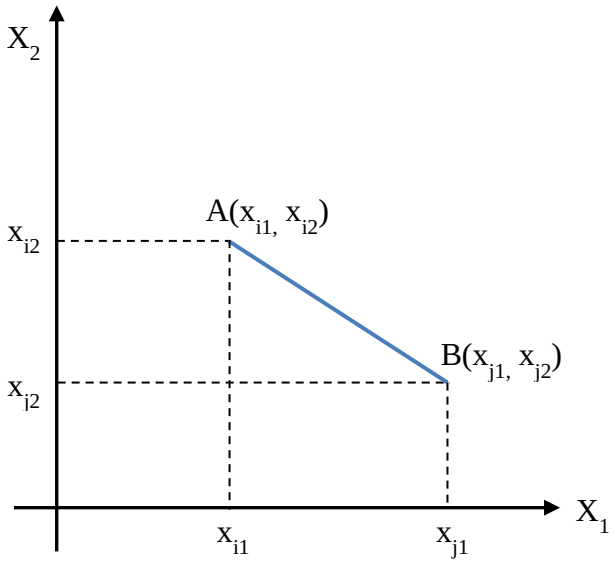
2.1.1.1. Öklid uzaklığı (euclidean distance)

En yaygın kullanılan uzaklık ölçüsüdür. Çok boyutlu uzayda birimlerin birbirlerine olan geometrik uzaklığıdır. Birimlerin konumlarını baz alarak ne kadar farklı olduklarını belirler. Öklid Uzaklığı, genel bilinirliliğiyle bir üçgendeki hipotenüs uzunluğudur. Yani noktalar arasındaki uzaklığın tam ölçüsü olarak ifade edilir. Bu uzaklık ölçüsünün kullanılmasında standartlaştırılmış veriler değil işlenmemiş veriler kullanılır.

n birim sayısı ve p değişken (boyut) sayısı olmak üzere i,j = 1,2,3……n ,

i. ve j. birimleri arası uzaklık Şekil 2.4'te de gösterilen Öklid uzaklık ölçüsü kullanılarak Formül (2.1)'deki gibi hesaplanır;

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots + (x_{ip} - x_{jp})^2} \qquad (2.1)$$



Şekil 2.4. Öklid Uzaklığı grafiksel gösterimi

2.1.1.2. Minkowski uzaklığı (minkowski distance)

Minkowski uzaklık ölçüsü genel bir formüldür. Formülde yer alan m değerinin alacağı her bir farklı değere göre yeni formüller ortaya çıkmış olur. Minkowski uzaklık ölçüsü Formül (2.2) kullanılarak iki birim arasındaki uzaklık hesaplanır.

$$d(i,j) = \left[\left|x_{i1} - x_{j1}\right|^m + \left|x_{i2} - x_{j2}\right|^m + \ \ldots \ + \left|x_{ip} - x_{jp}\right|^m\right]^{1/m} \quad (2.2)$$

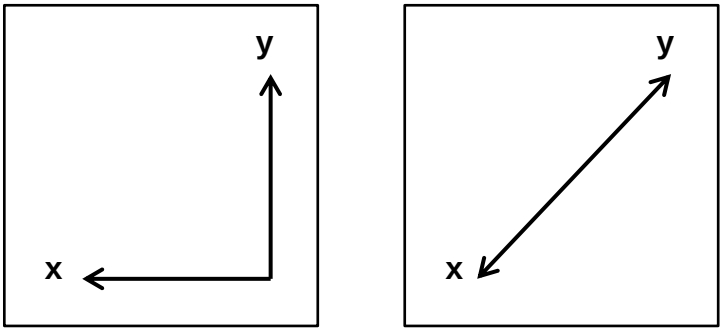
Bu uzaklık ölçüsü formülünde m değeri büyük ve küçük farklara verilen ağırlığı değiştirir. m değeri 1 olarak alınırsa formül, Manhattan uzaklık ölçüsü formülüne dönüşür. Eğer m değeri 2 olarak alınırsa formül Öklid uzaklık ölçüsü formülüne dönüşür [36].

2.1.1.3. Manhattan uzaklığı (manhattan city-block distance)

Bu uzaklık ölçüsünde birimler arasındaki mutlak uzaklık değerleri kullanılır. Birimlerin aynı değişkenleri arasındaki mutlak farkların toplanması Formül (2.3) ile hesaplanır.

$$d(i,j) = \left[\left|x_{i1} - x_{j1}\right| + \left|x_{i2} - x_{j2}\right| + \ldots + \left|x_{ip} - x_{jp}\right|\right] \quad (2.3)$$

Manhattan uzaklık ölçüsü uygulamada bazı sorunlara neden olmaktadır. Bu sorunlardan en dikkat çekeni değişkenler arası ilişkinin yok sayılmasıdır. Eğer değişkenler arasında korelasyon mevcutsa Manhattan uzaklık ölçüsüyle hesaplanan değerler temelinde yapılan kümeleme bir anlam ifade etmeyecektir. Bir diğer sorun da ölçüm yapılan değişkenlerin birimleri farklı olduğu durumda standartlaştırılmış Karesel Öklid uzaklık ölçüsü ile değerlendirme yapıldığında Manhattan uzaklık ölçüsünün pek anlamlı sonuçlar vermediği görülmektedir [40].



(a) Manhattan (b) Öklid

Şekil 2.5. Manhattan (a) ve Öklid (b) uzaklıklarının grafiksel gösterimi

2.1.1.4. Karesel öklid uzaklığı (squared euclidean distance)

Birimlere ait aynı değişkenler arası farkların karelerinin toplanması ile hesaplanır. Bu yöntemde uzaklık, Formül (2.4)'teki gibi tanımlanır:

$$d(i,j) = \left(x_{i1} - x_{j1}\right)^2 + \left(x_{i2} - x_{j2}\right)^2 + \ldots + \left(x_{ip} - x_{jp}\right)^2 \quad (2.4)$$

2.1.1.5. Mahalanobis uzaklığı (D^2)

Bu uzaklık ölçüsü doğrudan birleştirme yapmaktadır. İki değişken arasında ilişki olması durumunda, bu değişkenler arasında kovaryansı veya korelasyonu göz önüne alan Mahalanobis uzaklığı kullanılması gerekir. p değişken sayısı olmak üzere x ve y gözlem grupları arasındaki uzaklık Formül (2.5) ile hesaplanmaktadır.

$$d(x,y) = D^2 = \left(x - y\right)^T S^{-1} \left(x - y\right) \quad (2.5)$$

Buradaki T sembolü matrisin transpozunu, S sembolü, (p×p) kovaryans matrisini, S^{-1} sembolü ise kovaryans matrisinin tersini ifade etmektedir. Bu uzaklık ölçüsünün aykırı noktaları da hesaplaması nedeniyle diğerlerine göre bir avantajından söz edilebilir [35].

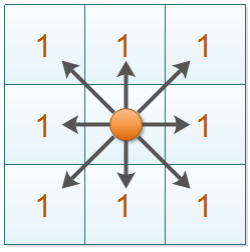
2.1.1.6. Chebyshev uzaklığı (chebyshev distance)

Chebyshev uzaklık ölçüsü farkların herhangi bir boyuttaki mutlak değerlerinin maksimumudur ve Formül (2.6) ile ifade edilir.

$$d_{ik} = d(x_i, y_k) = Max_i |x_{ji} - x_{jk}| \quad (2.6)$$

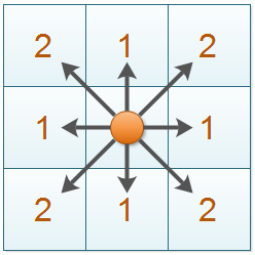
Bu uzaklık ölçüsüne ayrıca Minkowski uzaklığının özel bir durumu da denir [25].

(a) Chebyshev uzaklığı



$\max(|x_1 - x_2|, |y_1 - y_2|)$

(b) Manhattan uzaklığı



$|x_1-x_2| + |y_1-y_2|$

Şekil 2.6. Chebyshev (a) ve Manhattan (b) uzaklıklarının şekilsel gösterimi

2.1.2. İkili değişkenler

İkili değişkenler 1 veya 0, başka bir ifadeyle ise "var" ya da "yok" değerini alabilen değişkenlerdir. Bu tür değişkenler tanımlanan niteliğin mevcut olup olmadığını ortaya koyarlar. Örneğin kişinin okuma yazma bilmesi ile ilgili bir soru ancak evet ya da hayır değerini alacaktır. İkili değişkenler ile ilgili uzaklık hesaplaması için aşağıdaki gibi bir tablo geliştirilmiştir (Tablo 2.1). Bu tablo kontenjans tablosu olarak da adlandırılır [26].

Tablo 2.1. İkili değişkenler arası uzaklık hesaplama tablosu

		j.Gözlem		
i.Gözlem		Evet	Hayır	Toplam
	Evet	a	b	a+b
	Hayır	c	d	c+d
	Toplam	a+c	b+d	a+b+c+d

İkili değişkenlere ait benzerlik üzerine aşağıda Tablo 2.2'de bir örneğe yer verilmiştir.

Tablo 2.2. İkili değişkenler örnek tablosu

Meyve	Sulu	Tatlı	Ekşi	Kabuklu
Portakal (i)	Evet	Evet	Evet	Evet
Çilek (j)	Hayır	Evet	Hayır	Hayır

Bu örnekte portakalın koordinatı (1,1,1,1), çileğin koordinatı ise (0,1,0,0) olmuştur. Meyveler 4 değişken ile tanımlandığından dolayı bu nesneler 4 boyutlu olarak ifade edilir.

Örnekteki gözlemlerden hareketle a=1, b=3, c=0, d=0 olarak hesaplanmıştır.

İkili değişkenler için de farklı benzerlik ölçme metotları geliştirilmiştir. Bunlardan en yaygın kullanılanları Tablo 2.3'te gösterilmiştir.

Tablo 2.3. Yaygın kullanılan ikili değişkenler

BENZERLİK ÖLÇME YÖNTEMLERİ	DENKLEMLER
İkili Öklid	$\sqrt{\mathbf{b}+\mathbf{c}}$
İkili Karesel Öklid	$\mathbf{b}+\mathbf{c}$
Jaccard Benzerlik	$\frac{\mathbf{a}}{\mathbf{a}+\mathbf{b}+\mathbf{c}}$
Ochiai Benzerlik	$\frac{\mathbf{a}}{\sqrt{(\mathbf{a}+\mathbf{b})(\mathbf{a}+\mathbf{c})}}$
Rao	$\frac{\mathbf{a}}{\mathbf{a}+\mathbf{b}+\mathbf{c}+\mathbf{d}}$
Basit Eşleşme	$\frac{\mathbf{a}+\mathbf{d}}{\mathbf{a}+\mathbf{b}+\mathbf{c}+\mathbf{d}}$

2.1.3. Ordinal değişkenler

Sıralama ölçütünün keyfi olarak belirlendiği ve değerler arası farkların pek anlamlı olmadığı bir ölçme tekniğidir. Anketlere verilen cevaplarda bu tekniğe sıkça rastlanır. Örneğin bir hizmet anketinin 5 şıklı cevaplarının şöyle olduğu varsayılsın;

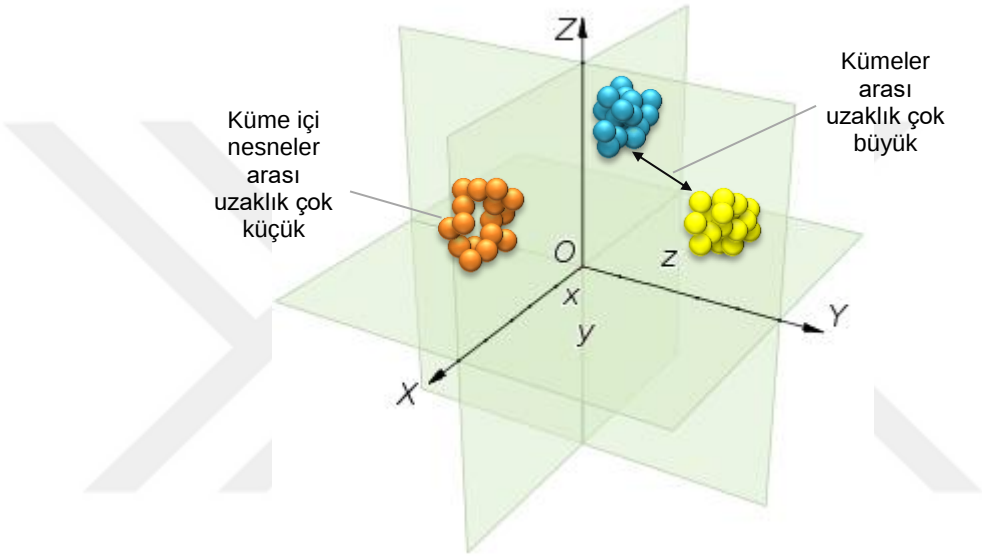
a. Çok memnun kaldım
b. Memnun kaldım
c. Kararsızım
d. Memnun kalmadım
e. Hiç memnun kalmadım

Bu cevaplar üzerinde matematiksel işlemleri yapmak mümkün olmamaktadır. Daha açık ifadeyle bir cevabın bir diğer cevaba belli bir oranda üstünlüğünden

bahsedilememektedir. Buradaki veriler 0 ve 1 değerleri arasına indirgenerek uzaklık/benzerlik değerleri hesaplanabilmektedir.

2.2. Kümeleme Teknikleri

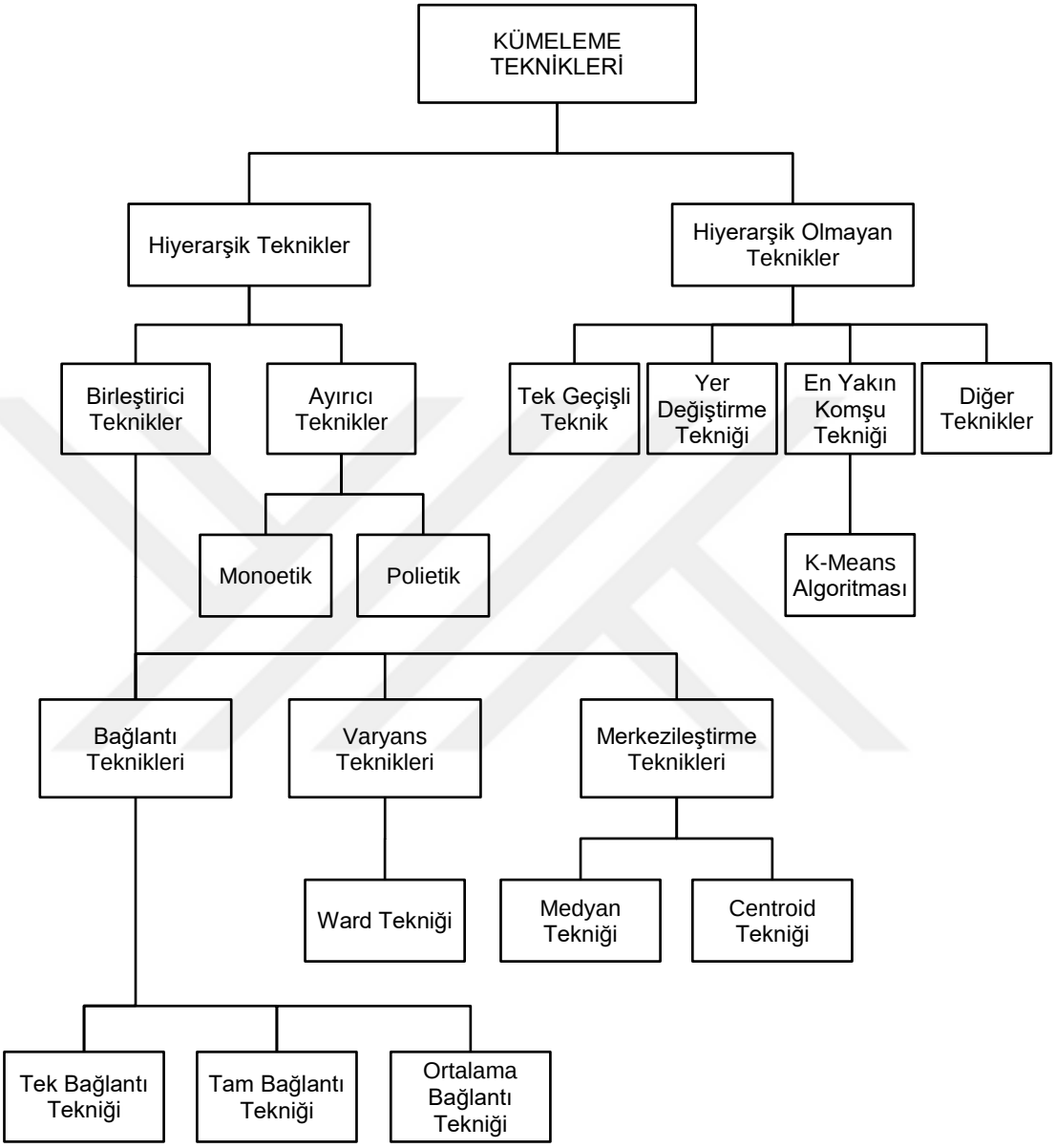
Kümeleme teknikleri, uzaklık matrisi yardımıyla nesnelerin veya verilerin, kendi içerisinde homojen, kendi aralarında ise heterojen gruplar oluşturulmasını sağlayan çalışmalardır [27]. Buna ilişkin görsel Şekil 2.7'de gösterilmiştir.



Şekil 2.7. Küme içi ve kümeler arası uzaklıklar

Kümeleme çalışmaları ile ilgili olarak özellikle 2000'li yıllardan itibaren bugüne kadar pek çok teknik ve varyantları geliştirilmiştir. Bu teknikleri standardize etmek ve onları kategorilere ayırmak oldukça zordur. Literatürde genel kabul gören kategorizasyon kümeleme tekniklerinin hiyerarşik teknikler ve hiyerarşik olmayan teknikler olarak iki gruba ayrıldığı şekildedir. Bazı tekniklerin birden fazla kategoride yer alabildiği de görülmektedir [28]. Hiyerarşik teknikler, veriler veya nesneler arasında ilişki düzeyini göstermesi bakımından dendogram (hiyerarşik ağaç diyagramı) adı verilen grafik ile ifade edilmektedir. Kümeler arasındaki benzemezliklerin ve küme içi benzerliklerin maksimizasyonu her iki tekniğin de ortak amacıdır. Bu küme içindeki homojenliğin artırılması ve kümeler arasındaki homojenliğin ise azaltılması anlamına da gelmektedir. Elde edilmek istenen küme sayısı, hangi tekniğin kullanılacağı kararının verilmesi noktasında en önemli etkendir. Bazı durumlarda her iki tekniğin de uygulanması ve en uygun sonucu veren

teknik üzerinden ilerleme yapılması yararlı görülmektedir. Kümeleme tekniklerinin sınıflandırılması (Şekil 2.8'de gösterilmiştir [28, 29].



Şekil 2.8. Kümeleme teknikleri ve sınıflandırması

2.2.1. Hiyerarşik kümeleme teknikleri

Hiyerarşik kümeleme tekniği, verileri/nesneleri ardı ardına birleştirme sürecinden ibarettir. Birleşerek küme oluşturan bir veri/nesne grubu sonraki adımlarda kesinlikle ayrılmaz. Bu tekniklerin sonuçları dendogram adı verilen hiyerarşik ağaç diyagramı ile görselleştirilir. Bu tekniklerin başlangıcı nesnelerin arasındaki uzaklıkları ifade eden uzaklık matrisidir. İlk başta her bir nesne bir küme olarak değerlendirilir. Daha

sonra yakınlık durumuna göre her bir küme kendisine en yakın olan ile sırasıyla birleştirilir. Art arda süren bu işlem bu tekniklerin temel yaklaşımını oluşturmaktadır. Bu tekniklerden yaygın olarak kullanılan ve genel kabul görenler; Tek Bağlantı Tekniği (En Yakın Komşuluk Algoritması, Single Link Clustering veya Single Linkage Clustering), Tam Bağlantı Tekniği (En Uzak Komşuluk Algoritması, Complete Link Clustering veya Complete Linkage Clustering), Ortalama Bağlantı Tekniği (Average Link Clustering veya Average Linkage Clustering), Ward Tekniği, Medyan Tekniği ve Centroid Tekniği'dir [30].

Tek Bağlantı, Tam Bağlantı ve Ortalama Bağlantı teknikleri grafik teknikleri olarak adlandırılır. Diğer teknikler ise kümelerin ifade edilmesinde geometrik merkezleri kullandıkları için geometrik teknikler olarak adlandırılır [32].

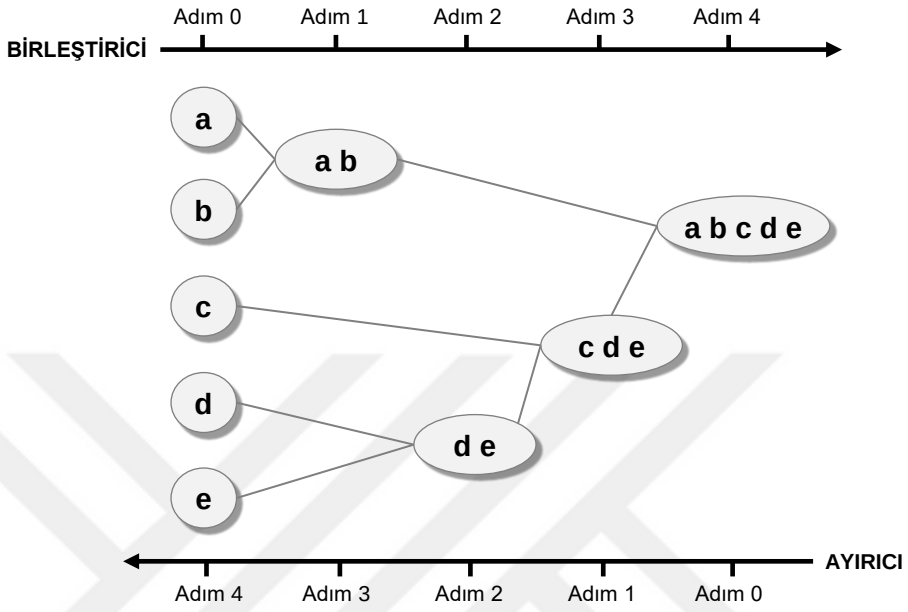
Bu tekniklerden birinin her durumda diğerlerinden üstün sonuçlar verdiği söylenememektedir. Dolayısıyla en iyi teknik olarak addedilen bir teknik bulunmamaktadır. Bir veri seti faklı tekniklerle farklı sonuçlar verebilmektedir.

Kümeleme, veri seti içinde verilen değişkenlere bağlı bir olaydır. Bu nedenden ötürü kümeler, kümelemenin amacı ile ilişkili olmalıdır. Bu konuya örnek vermek gerekirse görme konusunda rahatsızlık yaşayan hastaları gruplara ayırmak için onların boy ve kilo gibi fiziksel özelliklerini ölçmenin bir anlamı yoktur. Seçilen her değişkene göre kümeleme analizinin sonuçları farklı olacaktır. Bu nedenle değişkenler özenle seçilmelidir.

Hiyerarşik kümelemede tekniklerinin en çok kullanılan tekniklerinden olan birleştirici tekniklerin ilk adımı her bir veriyi veya nesneyi başlangıçta bir küme olarak varsaymaktır. Belli aşamalardan sonra sürecin sonunda tüm veriler veya nesneler bir kümede toplanır. Bu sürecin işleyişi şu şekilde özetlenebilir [31];

1. Öncelikle n adet birey, n adet küme olmak üzere işleme başlanır
2. En yakın iki küme (d(i,j) değeri en küçük olan) birleştirilir
3. Küme sayısı bir indirgenerek yinelenmiş uzaklıklar matrisi bulunur
4. 2 ve 3 numaralı adımlar n-1 kez tekrarlanır

Ayrıştırıcı tekniklerde ise bu durumun tamamen tersi söz konusudur. Hiyerarşik kümelemenin grafiksel gösterimi Şekil 2.9'da, Birleştirici Hiyerarşik Kümeleme Tekniklerinin genel akış diyagramı ise Şekil 2.10'da gösterilmiştir.



Şekil 2.9. Hiyerarşik kümelemenin grafiksel gösterimi [24]

Şekil 2.10. Birleştirici hiyerarşik kümeleme teknikleri genel akış diyagramı [33]

Hiyerarşik kümeleme tekniklerinde karşılaşılan en büyük sorun bireylerin birleşerek tek bir kümeye dahil olma süreci esnasında bu analizin hangi küme sayısında durdurulması gerektiğidir. Bu noktada ideal küme sayısı, küme iç uzaklıklarının minimum, kümeler arası uzaklığın maksimum olduğu andır. Fakat küme sayısı tespiti konusundaki karar uygulamayı gerçekleştiren kişiye bırakılmıştır. İdeal küme sayısı tespitinde kullanılan yöntemlerden birisi Formül (2.7)'de gösterilmiştir [31].

$$K = \left(\frac{n}{2}\right)^{1/2} \tag{2.7}$$

n: gözlem sayısı

K: küme sayısı

Hiyerarşik olmayan kümeleme tekniklerinde ise küme sayısı önceden belli olduğundan dolayı böyle bir tespite gerek yoktur. Bu tespitin özellikle örneklem

sayısı büyük olan durumlarda daha iyi sonuç verdiği söylenebilir. İdeal küme sayısının tespitine yönelik başka bir yaklaşım ise Formül (2.8)'de gösterilmiştir [41].

$$M = k^2|W| \qquad W = \sum_{j=1}^{k} \sum_{i=1}^{nj} \left(x_{ij} - \overline{x_j}\right) \left(x_{ij} - \overline{x_j}\right)' \tag{2.8}$$

W: Grup içi kareler toplamı matrisi
k: küme sayısı

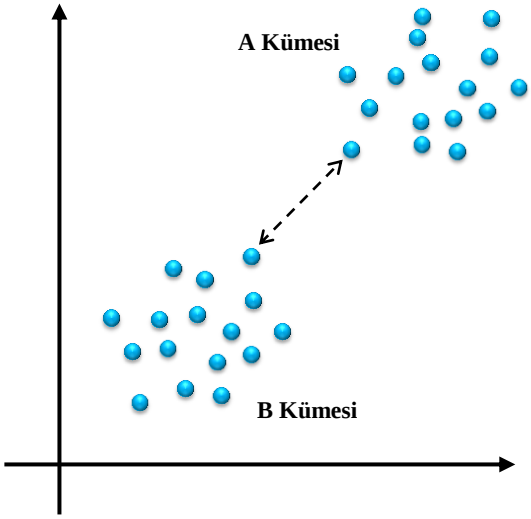
Bu yaklaşımda en küçük M değerini veren küme sayısının ideal küme sayısı olduğu ifade edilmektedir.

Hiyerarşik kümeleme tekniklerinin dezavantajları şu şekilde ifade edilebilir;
1. İki kümenin birleştirilmesine karar verildikten sonra geri alma işlemi yapılamamaktadır. Diğer bir ifadeyle analizin başında yapılan ayrıştırmada ortaya çıkan problemlerin analiz sonuna kadar devam etmesi söz konusudur.
2. Herhangi bir amaç fonksiyonu doğrudan minimize edilememektedir.
3. Farklı uzaklık ölçülerine göre farklı sonuçlar elde edilmektedir.

2.2.1.1. Tek bağlantı tekniği (SLC)

Birleştirici teknikler altında yer alan bu teknik, başlangıçta her bir değeri bir küme olarak varsaymaktadır. Bir algoritma formunda olan bu teknikte bir dizi adımla kümeler birleştirilmekte ve yeni kümeler elde edilmektedir. Bu teknikte kümeleme işlemi bütün kümelerin tek bir kümeye dahil olmasına kadar devam etmektedir.

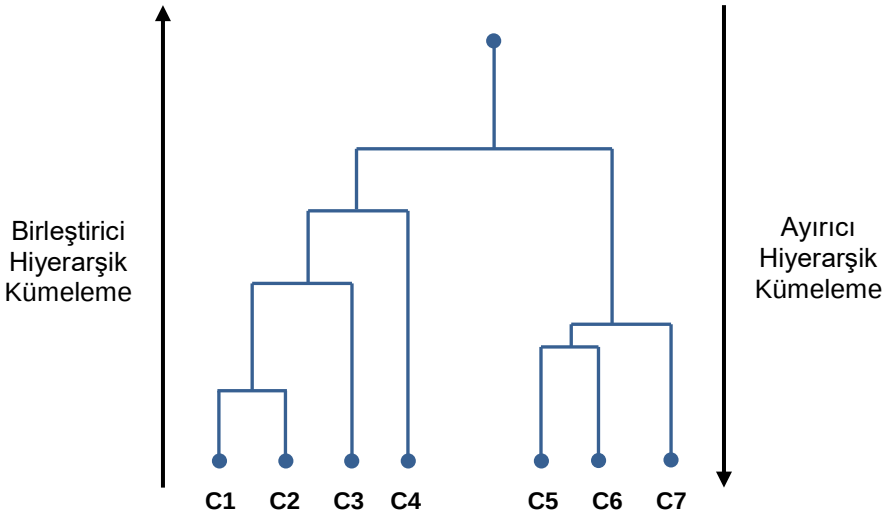
Bu teknikteki amaç, iki küme arasındaki uzaklığın minimum olduğu durumu belirlemek ve bu duruma göre tespit edilen iki kümeyi tek küme olacak şekilde birleştirmektir. Buna ilişkin görsel Şekil 2.11'de yer almaktadır. Bu işlem sürekli yapıldığında analiz sonunda tek bir küme elde edilecektir.



Şekil 2.11. Tek bağlantı tekniğinde iki kümenin birbirine olan uzaklığı

Tekniğin uygulama safhasında ilk işlem nesneler arasındaki uzaklıkların belirlenmesidir. i ve j indeksi ile gösterilen nesneler arasındaki uzaklıkların belirlenmesinde yaygın olarak Öklid Uzaklığı kullanılır.

Hiyerarşik kümeleme tekniklerinde sonuçlar, ağaç veri yapısı olarak da adlandırılan "Dendogram" ile gösterilir. Dendogram, görsel yapısıyla sonuçların kolay anlaşılmasını sağlar. Şekil 2.12'de örnek dendogram gösterimi yer almaktadır. Birbirleriyle benzer veya birbirlerine yakın olan nesneler dendogramda düşük değerlerde birleştirilirken, birbiriyle benzemeyen veya birbirlerine uzak olan nesneler ise yüksek bir değerde birleştirilmektedir.



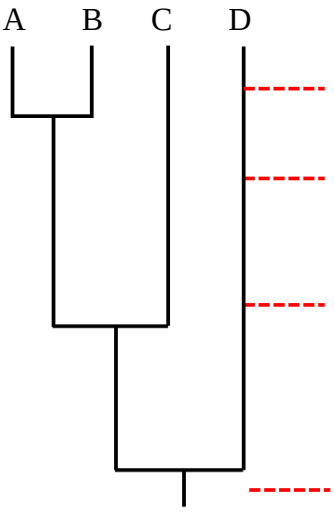
Şekil 2.12. Örnek dendogram gösterimi

Tablo 2.4'te uzaklık matrisi ile açıklanan örnek verilerin SLC Tekniği kullanılarak yapılan kümeleme analizi sonuçları Tablo 2.5' te verilmiştir.

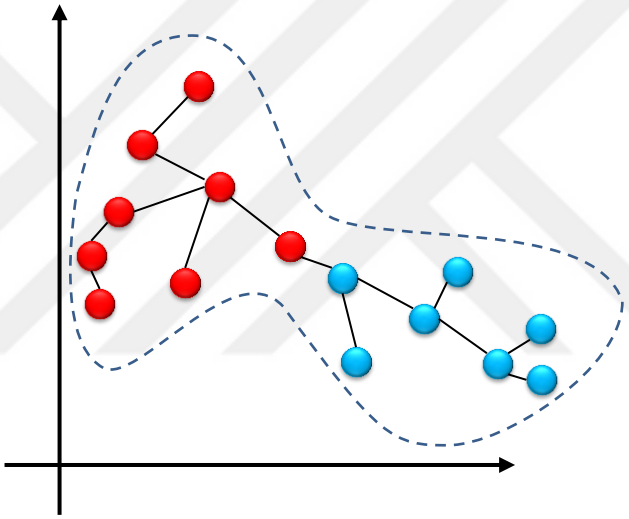
Tablo 2.4. Örnek veriler uzaklık matrisi (SLC)

	A	B	C	D
A	0	1	4	5
B		0	2	6
C			0	3
D				0

Tablo 2.5. Örnek verilerin SLC tekniği ile kümeleme analizi ve sonuçları

İterasyon	**Küme Sayısı**	**Kümeler**	**İşlem Adımları**	**Dendogram**
0	4	{A},{B},{C},{D}	Her veri bir küme	
1	3	{A,B},{C},{D}	A ve B arası uzaklık (d(A,B)=1) en küçük olduğu için {A} ve {B} birleştirilir.	
2	2	{A,B,C},{D}	B ve C arası uzaklık (d(B,C)=2) en küçük olduğu için {A,B} ve {C} birleştirilir.	
3	1	{A,B,C,D}	Son olarak {A,B,C} ile {D} birleştirilir.	

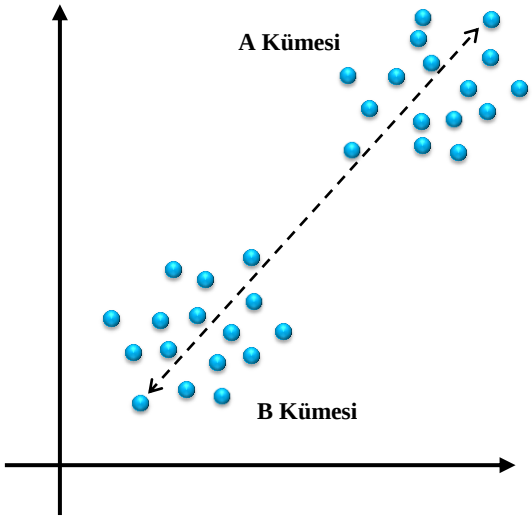
Aykırı değerlere karşı duyarlı olan bu teknik birbirinden yeterli düzeyde ayrık kümelerin tespit edilmesinde etkilidir. Ayrıca bu teknikte izlenen algoritma ile bir kümede toplanan nesneler diğer kümedeki nesnelere göre birbirlerine daha çok benzemektedir. Yine bu teknik dağılım özelliği göstermeyen elips şeklinde nesneleri aynı kümede toplayabilen ender kümeleme tekniklerinden biridir. SLC tekniği uzun kümeler oluşturma eğilimindedir. Buna "zincirleme etki" (chaining effect) de denilmektedir. Şekil 2.13'te zincirleme etki görseli yer almaktadır. Bu etki, teknik koordinat düzleminde U şeklinde dizilim gösteren noktaları aynı kümede toplayabilmektedir. Zincirleme etki sayesinde iki uzun dizilimin ucundaki nesneler birbirlerinden farklı bile olsa aynı kümede yer almaları mümkün olabilmektedir [34].



Şekil 2.13. Zincirleme etki

2.2.1.2. Tam bağlantı tekniği (CLC)

Bu teknik, SLC tekniği ile çok benzerlik göstermektedir. Fakat burada kümeler arası uzaklık tayin edilirken iki kümenin birbirine en uzak iki elemanı arasındaki mesafe iki küme arasındaki uzaklığı temsil etmektedir. Şekil 2.14'te bu durum görsel olarak ifade edilmiştir. İki küme birbirlerine en uzak elemanları kadar benzerdir yaklaşımını temel alır. SLC tekniğinde olduğu gibi her nesne başlangıçta bir küme olarak kabul edilir ve algoritma felsefesine göre adım adım nesneler birleştirilerek en sonunda tek bir küme elde edilir.



Şekil 2.14. Tam bağlantı tekniğinde iki kümenin birbirine olan uzaklığı

Bu teknikte de nesneler arasındaki uzaklıkların belirlenmesinde yaygın olarak Öklid Uzaklığı kullanılır ve sonuçlar dendogram yardımıyla görselleştirilir. Tablo 2.6'da uzaklık matrisi ile açıklanan örnek verilerin CLC tekniği ile yapılan kümeleme analizi sonuçları Tablo 2.7'de paylaşılmıştır.

Tablo 2.6. Örnek veriler uzaklık matrisi (CLC)

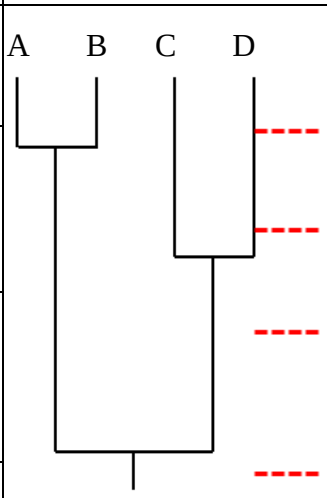
	A	B	C	D
A	0	1	4	5
B		0	2	6
C			0	3
D				0

Yukarıdaki verilere göre kümeleme adımında çizelgedeki en küçük değere sahip nesneler birleştirilir. Fakat kümeler birleştirilip tablo yenilenirken aralarındaki en uzak mesafe tabloya aktarılır. Şöyle ki;

min d(i,j) = d(A,B) = 1 olduğu için SLC tekniğindeki şekliyle A ve B nesneleri aynı kümeye dâhil edilir. Ancak tablo güncellenirken; max d(i,j) bulunur çizelge büyük değerlere göre güncellenir. Örneğin d(AB,D)'nin tespitinde d(A,D) ve d(B,D)'den hangisi büyükse o dikkate alınır.

Dolayısıyla d(A,D) = 5 ve d(B,D) = 6 olduğundan; büyük olan değer d(AB,D) = 6 tabloya aktarılır.

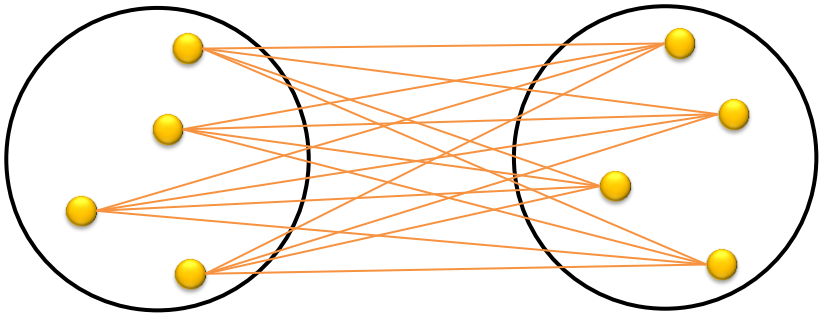
Tablo 2.7. Örnek verilerin tam bağlantı tekniği ile kümeleme analizi ve sonuçları

İterasyon	Küme Sayısı	Kümeler	İşlem Adımları	Dendogram
0	4	{A},{B},{C},{D}	Her veri bir küme	
1	3	{A,B},{C},{D}	A ve B arası uzaklık (d(A,B)=1) en küçük olduğu için {A} ve {B} birleştirilir.	
2	2	{A,B},{C,D}	C ve D arası uzaklık (d(C,D)=3) en küçük olduğu için {C} ve {D} birleştirilir.	
3	1	{A,B,C,D}	Son olarak {A,B} ile {C,D} birleştirilir.	

Bu teknikte izlenen algoritma verilerin birbirlerine yakın olduğu durumlarda daha etkilidir. Aykırı değerlerden etkilenebilmektedir. Genel ifadeyle küçük ve yoğun kümelerde ve birbirlerine yakın verilerde daha başarılıdır. Bu teknik, SLC tekniğinden farklı olarak küresel kümeler oluşturmaya daha elverişlidir.

2.2.1.3. Ortalama bağlantı tekniği

Bu teknikte iki küme arasındaki uzaklık ilk kümedeki nesnelerin ikinci kümedeki nesnelere olan uzaklıklarının aritmetik ortalaması olarak alınır. Aynı zamanda bu tekniğe aritmetik ortalama kullanan ağırlıksız grup çiftleri tekniği de denilmektedir. Ortalama Bağlantı Tekniği; SLC ve CLC tekniklerinin uç değerlere olan hassasiyetinin etkisini azaltmak maksadıyla bu iki tekniğe alternatif olarak önerilmiştir. Teknikte yer alan algoritmadaki işlem sırası, temelini baz aldığı SLC ve CLC teknikleri ile aynıdır.



Şekil 2.15. Ortalama bağlantı tekniğinde iki kümenin birbirine olan uzaklığında dikkate alınan uzaklıklar

Bu teknikte iki küme arası uzaklık aşağıda gösterilen Formül (2.9) ile ifade edilir;

$$d(C_i,C_j) = \frac{\sum_{y\in C_j}^{x\in C_i} d(x,y)}{m_i * m_j} \tag{2.9}$$

Burada C kümeleri, i ve j indeksli biçimde ifade etmektedir. d uzaklığı, m ise kümelerin eleman sayılarını göstermektedir.

2.2.1.4. Ağırlıklı ortalama bağlantı tekniği

Bu teknikte iki küme arasındaki uzaklık, Ortalama Bağlantı Tekniğinde olduğu gibi bir hesaplama ile gerçekleştirilir. Ortalama bağlantı tekniğinden farklı olarak yeni oluşan küme ile diğer kümeler arasındaki uzaklık her bir kümedeki eleman sayısı ile ağırlıklandırılır. Bu teknikteki uzaklığı gösteren formül aşağıda ifade edilmiştir (Formül (2.10)).

$$d(c_1,(c_i,c_j)) = \frac{n_i}{n_i+n_j} d(c_1,c_i) + \frac{n_j}{n_i+n_j} d(c_1,c_j) \tag{2.10}$$

2.2.1.5. Ward tekniği

Ward Tekniği, Joe H. Ward tarafından 1963 yılında ortaya atılmış olup minimum varyans tekniği olarak da adlandırılmaktadır. Bu teknikte kümeler arasındaki uzaklıklar hesaplanmayıp, küme içindeki homojenliğin maksimizasyonuna odaklanılmaktadır. Homojenliğin ölçüsü olarak da küme içi kareler toplamı kullanılmakta ve bunun minimize edilmesine çalışılmaktadır. Minimum değeri araştırılan küme içi kareler toplamı aynı zamanda Hata Kareler Toplamı (HKT) olarak da adlandırılmaktadır [35].

Küme birleştirme işlemine değişkenliğin en az olduğu kümelerden başlanmaktadır. Diğer tekniklere göre daha istatistiksel bir tekniktir. Bu teknik, birim sayısı az olan kümeleri birleştirme eğiliminde olup küme eleman sayılarını eşitlemeye çalışmaktadır. Yaklaşık veya eşit sayıda elemana sahip kümeler elde edilmesi isteniyorsa bu tekniğin kullanılması daha uygundur.

Küme merkezinin (ortalaması) hesaplanması Formül (2.11)'de gösterilmiştir.

$$m_i = \frac{1}{n_i} \sum_{x \in G_i} x \tag{2.11}$$

Burada n_i, G_i kümesindeki eleman sayısıdır.

HKT Formül (2.12) yardımıyla hesaplanmaktadır.

$$E = \sum_{k=1}^{K} \sum_{x \in G_k} \left\| x_i - m_k \right\|^2 \tag{2.12}$$

Burada K, küme sayısını ifade etmektedir.

İki kümenin birleşimi ile oluşan küme ile diğer kümeler arasındaki hata kareler farkı Formül (2.13)'te ifade edilmiştir [33].

$$\Delta E_{ij} = \frac{n_i n_j}{n_i + n_j} \left\| m_i - m_j \right\|^2 \tag{2.13}$$

Teknikte iki adımda kümeleme işlemi yapılmaktadır. Bunlar;

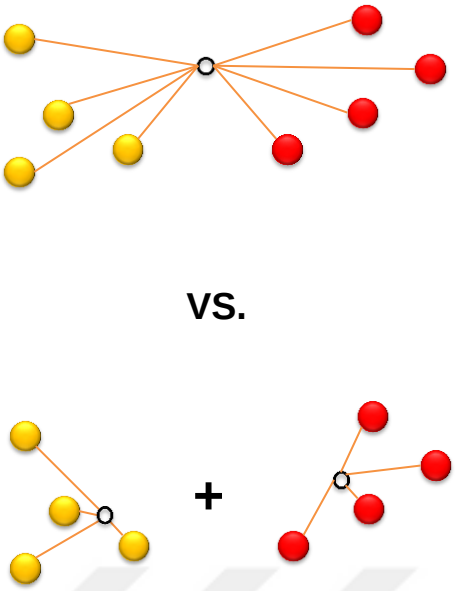
1. Başlangıçta her birim bir küme olarak kabul edilir. (HKT_k =0)
2. Daha sonra hata kareler toplamında en az artışı sağlayan {A} ve {B} kümeleri birleştirilerek {AB} kümesi oluşturulur.

HKT'daki bu artış ;

$$\Delta HKT_{AB} = HKT_{(AB)} - HKT_{(A)} - HKT_{(B)} \tag{2.14}$$

Formül (2.14)'teki şekilde hesaplanır ve n birim, (n-1) kümeye ayrılır.

3. Küme sayısı k=1 oluncaya kadar 2 numaralı adım tekrarlanır.

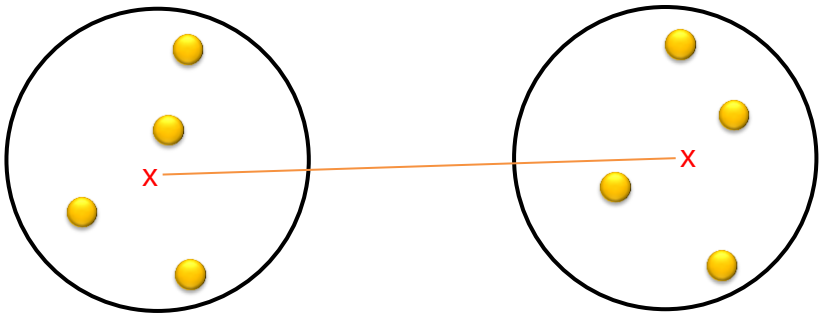


Şekil 2.16. Ward tekniği görseli

2.2.1.6. Centroid tekniği

Bu teknikte iki küme arasındaki uzaklık, kümelerdeki tüm nesneler arasındaki uzaklığın ortalamasını kullanmak yerine ilk adımda her bir kümenin küme merkezinin hesaplanması ve sonrasında kümeler arasındaki mesafe hesaplamasında bu merkezler arası Öklid Uzaklığı'nın kullanılmasıyla tanımlanır. Küme merkezi kümedeki tüm değerlerin değişkenlere göre ortalamasını ifade etmektedir. Teknik, yakınlık yerine veri matrisini kullanır. Merkez değeri kümedeki tüm nesnelerin değerleri hesaba katılarak hesaplandığından kümeye yeni bir birim eklendiğinde küme merkezi değişmektedir.

Uzaklık ölçüsü olarak Kareli Öklid Uzaklığı'nın bu teknikte daha mantıklı sonuçlar verdiği görülmüştür. Diğer uzaklık ölçüleri yorumlama açısından hatalı olabilmektedir [36].



Şekil 2.17. Centroid tekniğinde iki kümenin birbirine olan uzaklığında dikkate alınan uzaklıklar

Gi ile ifade edilen i. kümenin merkezi Formül (2.15) ile ifade edilir;

$$m_i = \frac{1}{n_i} \sum_{x \in G_i} x \tag{2.15}$$

İki küme arası uzaklık ise;

$$D(G_1,(G_i,G_j)) = \left\| m_1 - m_{ij} \right\|^2 \tag{2.16}$$

Formül (2.16)'da tanımlanan formül ile gösterilmekte olup bu uzaklık iki küme merkezi arasındaki Karesel Öklid Uzaklığı'na eşdeğerdir.

Hiyerarşik tekniklere ilişkin girdiler ve uzaklık tanımları ve özet bilgileri Tablo 2.8'de belirtilmiştir.

Tablo 2.8. Hiyerarşik kümeleme teknikleri özet bilgiler [37]

TEKNİK	ALTERNATİF ADI	KULLANILAN UZAKLIK	KÜMELER ARASINDAKİ UZAKLIĞIN TANIMI
Tek Bağlantı (SLC)	En Yakın Komşu	Benzerlik ya da uzaklık Matrisi	Nesne çiftleri arasındaki minimum uzaklıktır.
Tam Bağlantı (CLC)	En Uzak Komşu	Benzerlik ya da uzaklık Matrisi	Nesne çiftleri arasındaki maksimum uzaklıktır.
Ortalama Bağlantı	UPGMA	Benzerlik ya da uzaklık Matrisi	Nesne çiftleri arasındaki ortalama uzaklıktır.
Merkez(Centroid)	UPGMC	Uzaklık (veri matrisine ihtiyaç vardır)	Ortalama vektörleri(centroid) arasındaki kareli öklit uzaklığıdır.
Medyan	WPGMC	Uzaklık (veri matrisine ihtiyaç vardır)	Ağırlıklandırılmış merkezler(centroid) arasındaki kareli öklit uzaklığıdır.
Ward	Minimum Kareler Toplamı	Uzaklık (veri matrisine ihtiyaç vardır)	Birleştirmeden sonra küme içi kareler toplamındaki artışın tüm değişkenler için toplamıdır.

U:Ağırlıklandırılmamış; W:Ağırlıklandırılmış; PG:Grup çifti; M:Metod; A:ortalama; C:Merkez

2.2.2. Hiyerarşik olmayan kümeleme teknikleri

Bazı durumlarda elde edilmesi amaçlanan küme sayısı önceden bellidir ve bu küme sayısına sadık kalınarak çözümler üretilmesi gerekmektedir. Küme sayısı ile ilgili bir bilgi varsa ya da küme sayısına karar verilmiş ise bu durumda uygulanması için genellikle uzun zaman gerektiren hiyerarşik teknikler yerine hiyerarşik olmayan teknikler kullanılmaktadır [36].

Hiyerarşik olmayan kümeleme teknikleri, birimlerin K adet kümede toplanması için geliştirilmiştir. Hiyerarşik olmayan kümeleme teknikleri, hiyerarşik kümeleme tekniklerine oranla daha büyük sayıdaki nesnelere/verilere uygulanabilir. Hiyerarşik tekniklerden farklı olarak uzaklık matrisine ihtiyaç duyulmaz ve doğrudan ham verilerle işlem yapılabilir. Yine hiyerarşik tekniklerden farklı olarak nesneler atandıkları kümeden ayrılarak başka bir kümeye dahil olabilirler. Hiyerarşik olmayan kümeleme teknikleriyle oluşturulacak K kümenin her birinde en az bir nesne bulunur ve her birim hiyerarşik tekniklerde olduğu gibi yalnızca bir kümeye aittir. Bu tekniklerin ilk adımında rassal olarak K adet nesne başlangıç küme merkezi olarak seçilir. Nesnelerin belirlenen kümelerin merkezlerine olan uzaklıklarına göre yeni küme merkezleri oluşturulur. Dizi işlemler birbirleri arasında benzerliğin düşük

olduğu kümeler elde edilinceye kadar sürdürülür. Hiyerarşik olmayan kümeleme teknikleri arasında en yaygın olarak kullanılan teknik, K-Means Algoritması'dır.

2.2.2.1. K-Means algoritması

K-Means Algoritması'yla nesneler K adet kümeye bölünmektedir. Bu bölme işleminde küme içi hata kareler toplamının minimizasyonu hedeflenmektedir. Nicel verilere uygulanan K-Means Algoritması'nın en büyük özelliği küme sayısının önceden belirlenmesidir. Diğer tekniklerde olduğu gibi amaç küme içi benzerliklerinin maksimum ve kümeler arası benzerliklerinin minimum olmasını sağlamaktır. K-Means Algoritması'nda her biri tek nesneden oluşan K adet küme ile işleme başlanır. Her bir nesne en yakın ortalamalı kümeye dahil edilir ve küme merkezleri tekrar hesaplanır. Bu süreç tüm elemanlar bir kümeye atanıncaya kadar devam eder. Sonrasında ise hata karesi azaltılacak şekilde nesnelerin bulundukları küme ortalamasından daha yakın küme ortalaması varsa yerleri değiştirilir. Bu işlemler de kümeler arasında nesne alışverişi olmadığı duruma kadar devam eder. Algoritmanın işlem adımları şu şekilde tanımlanabilir:

1. K adet küme için K adet nesne rassal olarak seçilir ve küme merkezleri olarak belirlenir.
2. Her bir nesne belirlenen küme merkezlerine göre yakınlıkları belirlenerek en yakın kümeye dahil edilir.
3. Her bir küme merkezi yeniden hesaplanır.
4. 2.ve 3. adımlar tekrarlanır.

Hata kareler toplamının minimizasyonu Formül (2.17)'de tanımlanmıştır.

$$J_k = \sum_{k=1}^{K} \sum_{i \in C_k} (x_i - m_k)^2 \qquad (2.17)$$

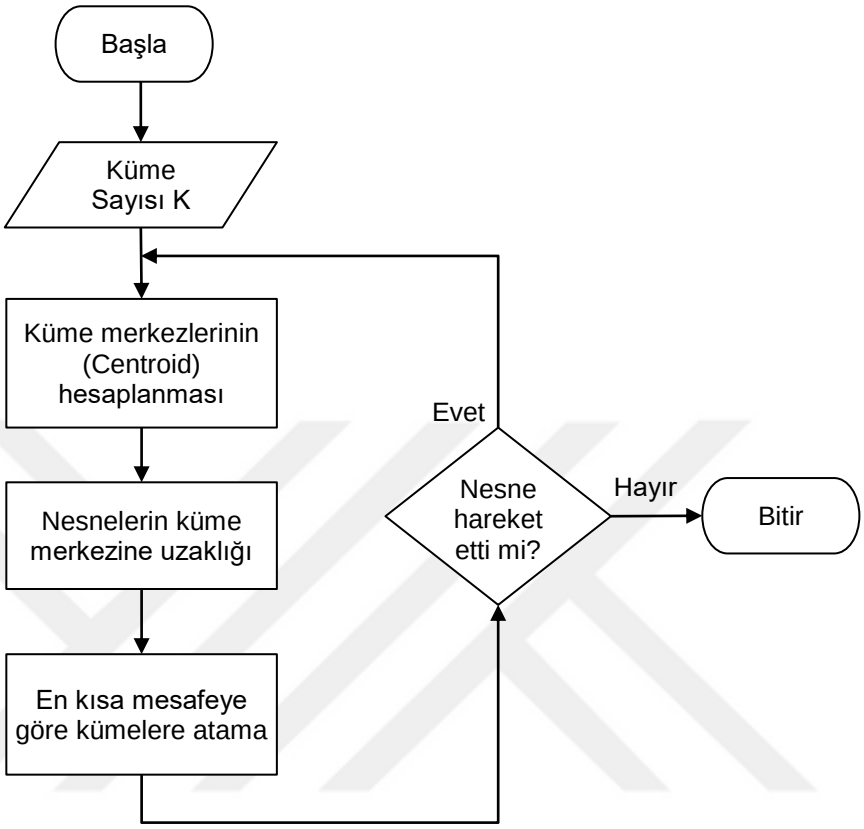
J_k: k. kümedeki nesnelerin hata kareleri toplamı

x_i: i. nesneye ait değer

m_k: k. kümenin merkez noktası

Küme merkezleri kararlı hale gelinceye kadar algoritma kendisini tekrar eder. Merkezin değişmeyecek hale gelmesi, hata kareler toplamının minimum değerine

ulaşmasıyla anlaşılır. K-Means Algoritması'nın akış şeması Şekil 2.18'de gösterilmiştir.

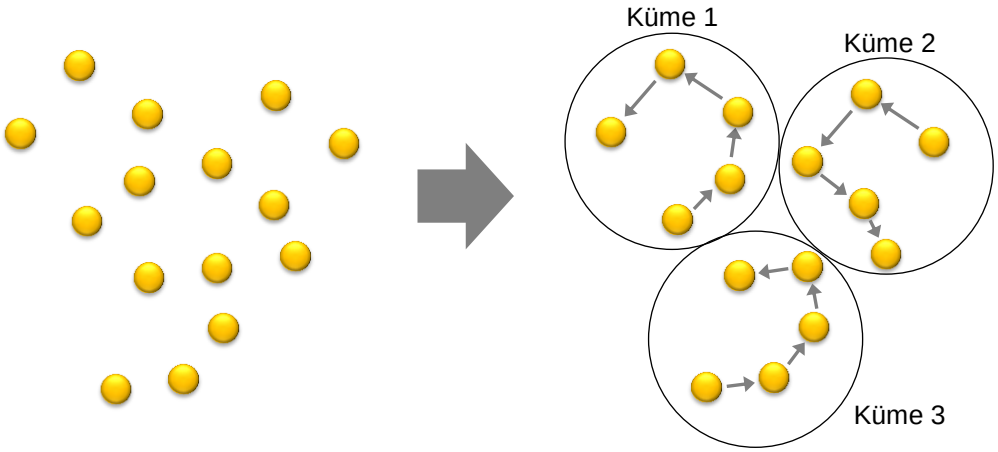


Şekil 2.18. K-Means algoritması akış şeması

3. TEKNİK GELİŞTİRME VE UYGULAMA

3.1. Çalışmanın Amacı

Bilgi teknolojileri alanındaki gelişmeler ve iletişim olanaklarının artması küresel rekabetin tetiklenmesi ve ivmelenmesi konusunda şüphesiz en önemli faktörlerdendir. Aynı zamanda gelişen bilgi teknolojileri sayesinde organizasyonlar ve ticari işletmeler faaliyetlerindeki verimliliğin ölçümlenmesini daha detaylı bir şekilde yapmakta ve bunun karlılık düzeylerine etkisini her geçen gün daha hassas bir şekilde görebilmektedirler. Geçmişte yapılan işin niteliğinin, işin tamamlanma zamanı ve maliyeti kavramlarının önüne geçtiği durumlar oldukça çoktur. Ancak günümüzde bir işin en düşük maliyetle, en kısa zamanda, istenen temel gereksinimleri karşılayacak şekilde yapılması daha büyük önem kazanmıştır. Bu işlerin ve projelerin en düşük maliyetle ve en kısa zaman diliminde bitirilebilmesi için gerçekleştirilen çalışmalar ve bu yönde yapılan geliştirmeler son derece önem arz etmektedir. Bu noktada planlama ve organizasyon çalışmaları, işlerin tamamlanmasından daha fazla öneme sahip olmaktadır. Optimum sonuçlara ulaşmak ancak doğru bir planlama ile mümkündür. Özellikle iyi bir planlama gereksinimi ihtiyacı da belli parçalara ayrılarak yapılması gereken işlerde/projelerde göze çarpmaktadır. Geniş bir alanda bir çok noktada gerçekleştirilmesi gereken ancak tek bir seferde yapılamayan işlerin, gruplara/kümelere ayrılarak bir takvim dahilinde yapılabilmesi mümkündür. Ancak bu işlerin gerçekleştirileceği işlem noktaları gruplanırken birbirlerine yakın olanların bir arada olacak şekilde gruplanması en optimal sonuca ulaşılmasını sağlayacaktır. İşlem noktaları arasındaki ulaşımın karayolu ile sağlandığı düşünüldüğünde noktalar arası ulaşımda katedilen(/gidilen) toplam mesafenin minimizasyonunu hedef alan ve işlem noktalarını birbirlerine en yakın olanların bir kümede toplanacağı ve bunu yaparken her kümede eşit sayıda nokta olmasını sağlayacak bir tekniğin ve buna bağlı olarak da algoritmanın geliştirilmesi bu çalışmanın amacını oluşturmaktadır. Elde edilmesi hedeflenen küme yapılarını gösteren örnek görsel Şekil 3.1'de verilmiştir.



Şekil 3.1. Geliştirilecek teknikle elde edilmesi hedeflenen kümeleme yapısı

3.2. Çalışmanın Kapsamı

Çalışma Kocaeli, Sakarya, Bolu ve Düzce illerinden oluşan geniş bir alana yayılmış bir saha projesi çerçevesinde ele alınmıştır. Bu illerde toplam 8168 noktada gerçekleştirilen bir montaj projesinin en az maliyetle ve en kısa zamanda tamamlanmasına yönelik yaklaşım bu çalışmanın temelini oluşturmaktadır. Bir saha montaj ekibinin toplamda 8168 noktada uygulama yapmasıyla projenin tamamlanması söz konusudur. Ancak montaj ekibinin günlük işlem kapasitesinin sınırlı olduğu göz önüne alındığında bu 8168 noktanın ziyaret edilerek projenin tamamlanması oldukça uzun bir zaman dilimi gerektirmekte ve projenin aylara yayılmış bir iş olmasını beraberinde getirmektedir. Sözü edilen projenin uygulanması esnasında iki temel problem ile karşılaşıldığı görülmüştür. Bunlardan birincisi noktaların hangi sıra ile ziyaret edilmesi gerektiği konusunda yaşanan karar verme problemidir. Burada montaj ekibine yönelik günlük iş planları oluşturulmasının önemi anlaşılmaktadır. İkinci problemin ekibin ziyaret etmek üzere çizelgelediği noktalar arası ulaşımda geçen sürenin uzunluğu nedeniyle günlük montaj kapasitesine ulaşılamaması ve ulaşım ile harcanan zaman ve dolayısıyla yakıt israfına neden olunmasıdır. Burada ikinci problemin ilk problemin sonucu olduğu göze çarpmaktadır.

Sahadaki noktaların birbirlerine yakın olanların bir arada gruplandırılması ve gruplandırma yapılırken ekibin günlük kapasitesi doğrultusunda her bir grubun eşit

sayıda nokta içermesi gereksinimi araştırma konusunun ve çalışmanın çerçevesini oluşturmaktadır.

3.3. Materyal

Çalışma kapsamında ele alınan Kocaeli, Sakarya, Bolu ve Düzce illeri dahilinde bulunan 8163 işlem noktasına ait yeryüzü konum verisi çalışmanın temel materyalidir. Enlem ve boylam nümerik değerleri ile ifade edilen yeryüzü konum verileri Coğrafi Koordinat Sistemi kapsamında ele alınmakta ve Coğrafi Bilgi Sistemi aracılığıyla kullanılabilir veriye dönüştürülmektedir.

3.3.1. Coğrafi koordinat sistemi ve konumlandırma

Coğrafi Koordinat Sistemi (GCS), dünya üzerindeki herhangi bir yeri topografik bir nokta olarak tanımlamayı sağlayabilen bir koordinat sistemidir. GCS'de dünya, enlem ve boylamlarla dikey ve yatay olarak adreslenmiştir. Dünya üzerindeki her bir noktanın konumu, bu enlem ve boylamların birlikte kullanılmasıyla belirtilebilmektedir.

Enlem ve boylamlar için geleneksel ölçü birimleri derece, dakika (1/60 derece) ve saniye (1/60 dakika) olmakla birlikte ondalık sisteme uyarlanmış yazım biçimleri de vardır. Bunlar aşağıda belirtilmiştir.

- DMS Derece:Dakika:Saniye (49°30'00"N, 123°30'00"W)
- DM Derece:Dakika-ondalık (49°30,0' -123°30,0'), (49d30,0m,-123d30,0')
- DD Dakika-ondalık (49,5000°,-123,5000°), genellikle 4-6 ondalık haneyle

Günümüzde en yaygın olarak kullanılan ondalık sistemde dakika yazımı olan DD'dir. Bu yazım biçimi Google maps ve GPS aygıtları tarafından da benimsenmiştir [38].

Çalışmamıza konu olan 8163 noktanın konum verilerine ilişkin koordinatlar Dakika-Ondalık (DD) gösterim biçiminde olup bazı noktalara ait değerler Tablo 3.1'de verilmiştir.

Tablo 3.1. Örnek koordinat verileri

Nokta (ID Label)	**Bölge (Region)**	**Enlem (Latitude)**	**Boylam (Longitude)**
P001	Kocaeli	40,7758	29,9648
P002	Kocaeli	40,7721	29,9605
P003	Kocaeli	40,8021	29,9567
P004	Kocaeli	40,7737	29,9412
P005	Kocaeli	40,7014	29,9426
P006	Kocaeli	40,6975	29,9453
P007	Kocaeli	40,7117	29,9325
P008	Kocaeli	40,6497	29,9301
…	…	…	…

3.3.2. Coğrafi bilgi sistemi ve navigasyon

Coğrafi Bilgi Sistemi (GIS), konuma dayalı gözlemlerle elde edilen grafik ve grafik-olmayan bilgilerin toplanması, saklanması, işlenmesi ve kullanıcıya sunulması işlevlerini bütünlük içerisinde gerçekleştiren bir bilgi sistemidir. Başka bir ifadeyle grafik, tablo ve metin şeklinde raporlamaların sağlayamadığı coğrafi yakınlık/uzaklık, benzerlik/farklılık ve ilişkilerin gösterilmesini sağlayan güncel bir teknolojidir. Farklı alanlarda değişik amaçlar için kullanılmakla birlikte genelde GIS aşağıdaki üç amaca ulaşmayı hedeflemektedir:

1. Harita ve coğrafi bilgileri kullanarak üretkenliği arttırmak
2. Coğrafi veritabanında yönetimi geliştirmek
3. Karar vermeyi destekleyen coğrafi verileri kullanacak daha iyi strateji yolları ortaya koymak

GIS, coğrafya, şehir planlama, kartografya, uzaktan algılama, arazi ölçümleme, fotogrametri, acil durum yönetimi, afet planlama, navigasyon, konumsal bilgi teknolojileri, kriminoloji, jojistik ve pazarlama gibi birçok alanda kullanılmaktadır. Bunlardan navigasyon içinde bulunduğumuz dönemde çok sık kullanılmakta ve bir noktadan başka bir noktaya gitmek için en uygun yolu bulan uygulama olarak tarif edilmektedir. Enlem ve Boylam değerleri noktaların dünya üzerindeki konumlarını tarif etmekte olup birbirleri ile arasındaki ilişkiye yönelik herhangi bir veri

içermemektedir. Bu noktada navigasyon, konum verilerini kullanarak gidilecek adrese en kısa ve hızlı şekilde ulaşımı sağlayan sistemdir.

Navigasyon sistemi donanım, yazılım ve harita olmak üzere üç farklı katmandan oluşmaktadır. GPS (Global Konum Belirleme Sistemi) sinyalleri ile çalışan navigasyon yeryüzünde iki nokta arası olası karayolu ulaşım alternatiflerini listelemekte ve bunlardan seçilen bir tanesi üzerinden kullanıcısına güncel yol tariflemesi yapmaktadır.

Bir noktadan başka bir noktaya gidilmesi istendiğinde çoğu zaman alternatif rotalar mümkün olmaktadır. Bu rotalar aynı noktaya ulaşılmasını sağlasa da farklı uzunluklarda olabilmekte ve bunun doğal bir sonucu olarak farklı sürelerde erişim sağlanmasına neden olabilmektedir. Şekil 3.2'de A noktasından B noktasına ulaşım sağlanması için navigasyon sisteminin önerdiği üç farklı rota görülmektedir.

Şekil 3.2. İki nokta arası karayolu ulaşım alternatifleri

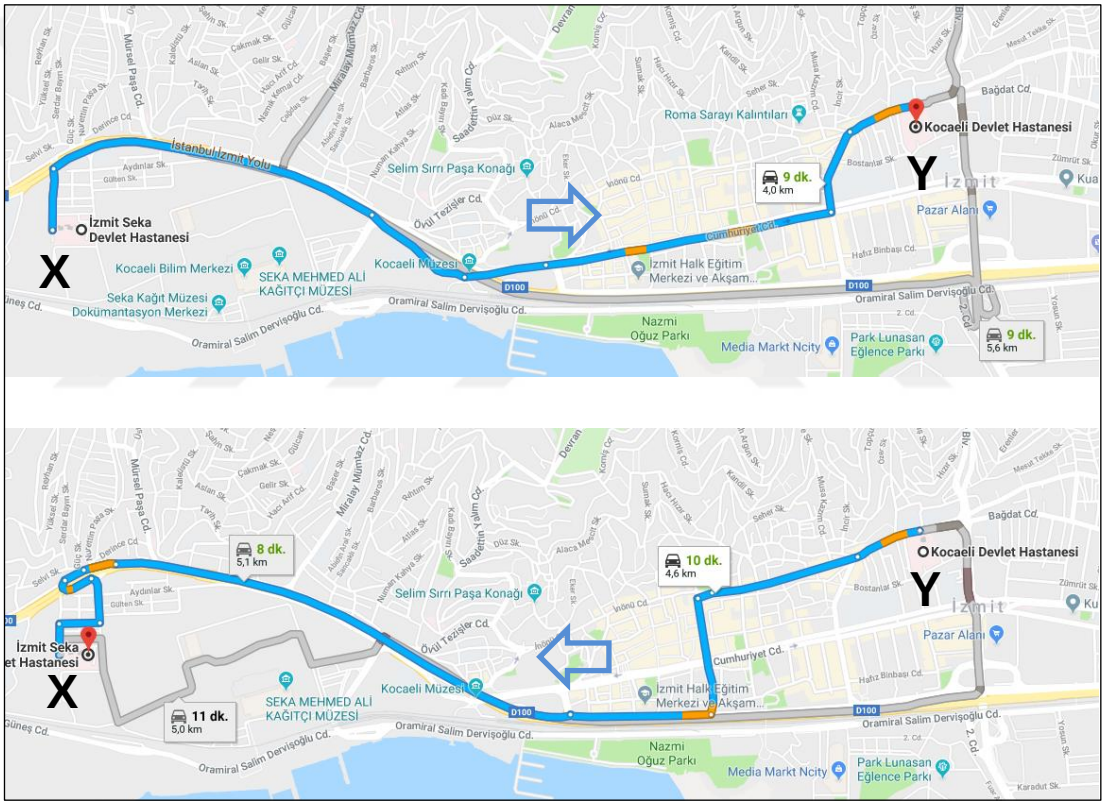
Bu noktaların koordinat bilgileri ile alternatif rotaların detaylı bilgileri Tablo.3.2'de verilmiştir.

Tablo 3.2.İki nokta arası karayolu ulaşım alternatifleri

Rota	**Kalkış Noktası Bilgileri (LAT, LONG)**	**Varış Noktası Bilgileri (LAT, LONG)**	**Ulaşım Mesafesi**	**Ulaşım Süresi**
Rota 1	40,822920, 29,921869	40,763808, 29,957120	15,3 km	19 dk
Rota 2	40,822920, 29,921869	40,763808, 29,957120	10,4 km	16 dk
Rota 3	40,822920, 29,921869	40,763808, 29,957120	13,0 km	16 dk

Buradan görüleceği üzere Rota 2 en kısa ulaşım mesafesine sahip olan yoldur. Ayrıca yol ve trafik şartlarından dolayı rotalama yapıldığı an itibarı ile Rota 3 ile aynı sürede ulaşımı öngörmektedir.

Navigasyon sistemi ile elde edilen verilerden hareketle iki nokta arası ulaşımda dikkati çeken başka bir husus da bazı durumlarda A noktasından B noktasına ulaşım mesafesinin, B noktasından A noktasına ulaşım mesafesinden farklılık arz etmesidir. Bu durumun nedenleri araştırıldığında bazı yolların tek yönlü olması ve kalkış/varış noktalarının karşı yönlü şeritte konumlanmış olma durumu nedeniyle ilave olarak katedilmesi gereken yollar kaynaklı olduğu görülmüştür.



Şekil 3.3. İki nokta arası karşılıklı en kısa ulaşım mesafeleri

Şekil 3.3'te örnek iki nokta arası en kısa rotalar verilmiştir. X noktasından Y noktasına gidilirken katedilen en kısa mesafe 4,0 km, Y noktasından X noktasına gidilirken katedilen en kısa mesafe ise 4,6 km olarak ölçümlenmiştir. Bunun nedeni tek yönlü olan yollardır.

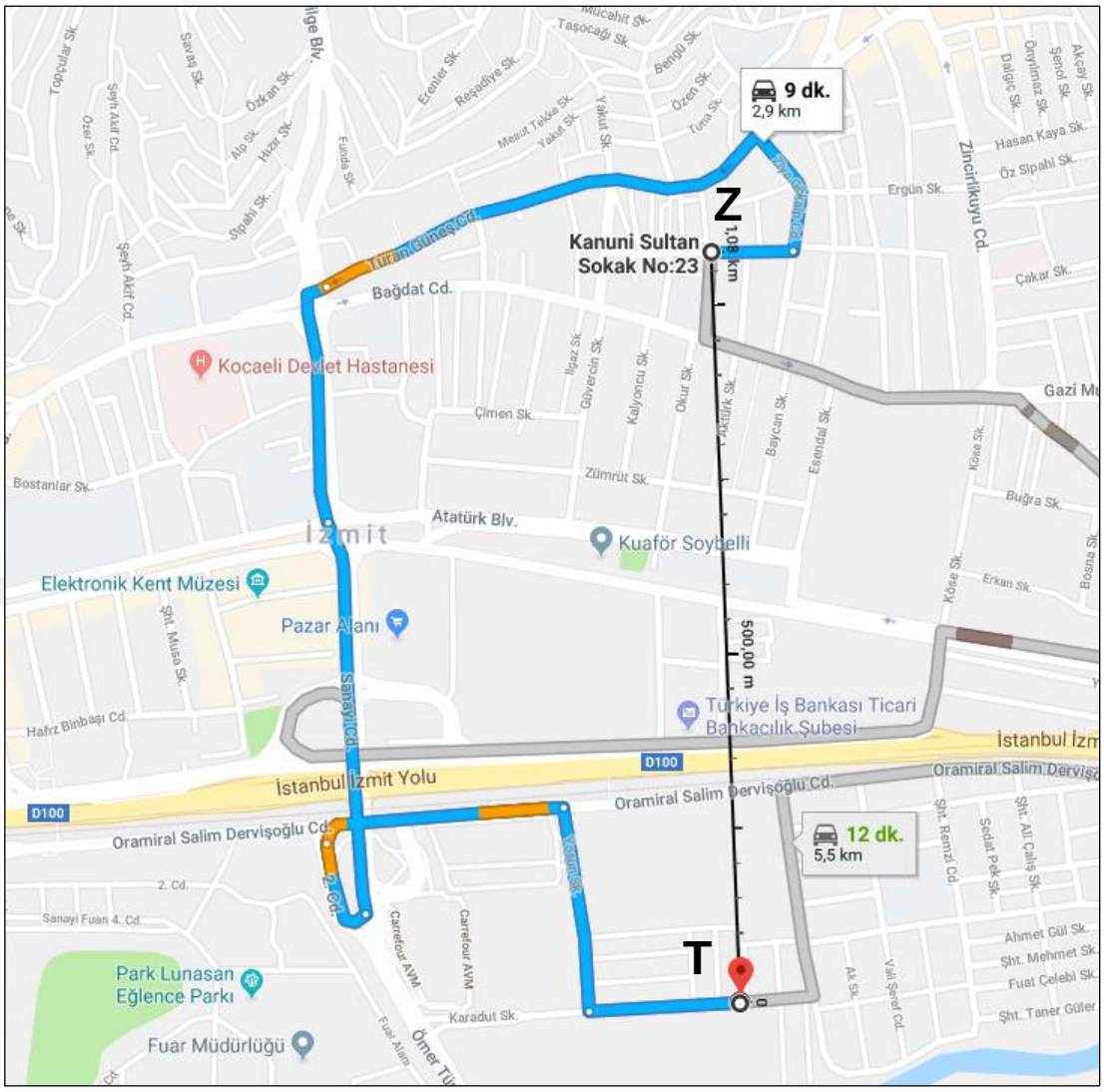
3.4. Kısıtlar ve Sınırlılıklar

Çalışma 8163 işlem noktası içeren bir saha projesi kapsamında ele alınmıştır. Saha ekibinin her bir işlem noktasını yalnız bir kez ziyaret etmesi yeterli gelmektedir. Bu durum işlem noktalarının gruplara ayrılması halinde her bir işlem noktasının sadece bir gruba ait olacağı anlamına gelmektedir. Ayrıca veri boyutunun yüksek olmasından dolayı uygulama veri seti içerisinden Kocaeli ili sınırları içerisinden rassal olarak seçilen 500 nokta üzerinde gerçekleştirilmiştir. Saha ekibinin günlük işlem kapasitesi 20 noktada işlem yapma ile sınırlıdır. 8163 işlem noktası çerçevesinden bakıldığında bir ekiple 409 günde, sınırlı olarak ele alınan 500 işlem noktası çerçevesinden bakıldığında ise 25 günde tamamlanması gereken bir proje olduğu görülmektedir.

3.5. Kullanılan Yöntem ve Teknikler

Geniş alana yayılan noktaların birbirleri ile aralarındaki yakınlık durumuna göre gruplara ayrılması işlemi bir kümeleme problemini tarif etmektedir. İkinci bölümde yaygın olarak kullanılan uzaklık/benzerlik ölçüleri ile kümeleme teknikleri ele alınmıştı. Kümeleme tekniklerinde kullanılan uzaklıklar, çeşitli uzaklık ölçüleri formülleriyle yapılan hesaplamalar sonucu elde edilen rakamsal değerlerdir. Uzaklık ölçülerinden en yaygın olarak kullanılanı Öklid Uzaklığı olarak karşımıza çıkmaktadır. Genel ifadeyle Öklid Uzaklığı, yeryüzü üzerinde iki nokta arası kuş uçuşu uzaklık anlamına gelmektedir. Ancak yeryüzünde iki nokta arası ulaşımın karayolu ile sağlandığı göz önüne alındığında noktaların kuş uçuşu uzaklıkların kullanılarak kümelenmesi optimal çözümden uzaklaşılmasına neden olacaktır. Dolayısıyla pratikle örtüşmesi açısından yeryüzü üzerindeki noktaların uzaklık ilişkisine dayalı kümelenmek istenmesi durumunda en doğru sonuç gerçek karayolu uzaklık verileri kullanılarak kümeleme çalışması ile elde edilebilir.

Şekil 3.4’te belirtilen örnek Z ve T noktaları arası en kısa karayolu ulaşım mesafesi 2,9 km iken bu iki nokta arası kuş uçuşu uzaklık 1,1 km olarak ölçümlenmiştir.



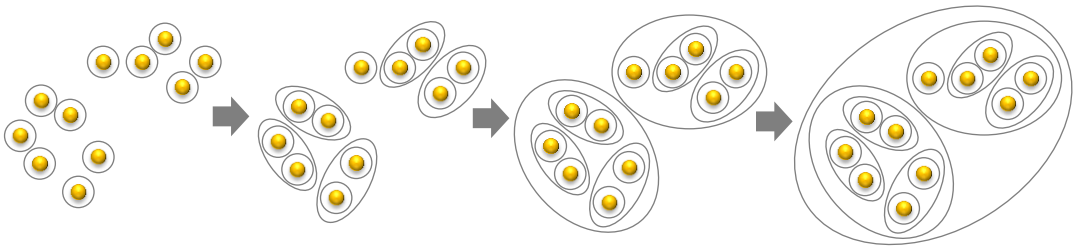
Şekil 3.4. İki nokta arası kuş uçuşu uzaklık ve en kısa karayolu ulaşım mesafesi

SLC ve CLC gibi bazı kümeleme teknikleri, küme elemanlarının kendisini baz alarak işlem yaptığı gibi K-Means Algoritması gibi bazı teknikler ise gerçekte var olmayan ve kümedeki elemanların konumuna göre hesaplanan hayali bir küme merkezini baz alarak işlem yapmaktadır. Bazı kümeleme teknikleri ise elemanların belli noktalara olan uzaklıklarının ortalamalarına göre işlem yapmaktadır. Ekibin günlük iş planı içerisinde yer alan noktaların sırayla ziyaret edilecek olması nedeniyle kullanılacak tekniğin uzaklıkların ortalamaları yerine uzaklıkların kendilerini baz almasının daha doğru olacağı düşünülmüştür.

Hiyerarşik kümeleme teknikleri, analiz sonunda tüm nesnelerin birbirleri ile olan yakınlık derecelerine göre hiyerarşik bir bağlantı ilişkisini gösteren tekniklerdir. Hiyerarşik kümeleme tekniklerinde küme sayısı önceden belirlenmemekte dendogram üzerinde uygun bir noktadan kesme yapılarak istenen sayıda küme elde edilmektedir. Bu tekniklerde kümelerdeki eleman sayılarını sınırlayıcı veya belirleyici herhangi bir unsur bulunmamaktadır. Hiyerarşik olmayan kümeleme tekniklerinde ise küme sayısı önceden belirlenebilmesine karşın bu tekniklerin de kümelerdeki eleman sayılarını sınırlayıcı veya belirleyici bir yönü yoktur. Burada ifade edilen nedenlerden ötürü tüm bu teknikler standart halleri ile çalışmaya konu olan saha projesine aranan çözüm için uygulanabilir durumda değildirler.

Birleştirici Hiyerarşik Kümeleme Tekniklerinden SLC ve CLC teknikleri üzerinde bazı geliştirmeler yapılarak ortaya çıkarılacak yeni bir teknikle problemin çözümüne yönelik bir yaklaşım elde edilebileceği düşünülmüştür.

Birleştirici Hiyerarşik Kümeleme Teknikleri'nin uygulama adımlarını gösteren akış Şekil 3.5'te verilmiştir.



Şekil 3.5. Birleştirici hiyerarşik kümeleme teknikleri uygulama akışı

3.6. Geliştirilen Teknikler

Birleştirici Hiyerarşik Kümeleme Tekniklerinden SLC tekniği ve CLC tekniğinde görülen bazı dezavantajlara ikinci bölümde değinilmişti. Bu dezavantajlara ve bunların elimine edilerek bu tekniklerin temelindeki algoritmanın geliştirilmesine yönelik temel yaklaşımlara aşağıda yer verilmiştir.

Dezavantaj: Algoritmada iki küme birleştirildikten sonra herhangi bir şekilde kümeler bozulamamakta ve bir geri alma işlemi yapılamamaktadır.

Önerilen Geliştirme: Eşit sayıda eleman içeren kümeler oluşturulabilmesi için kümeler arasında eleman alış verişi yapılabilmelidir. Toplam eleman sayısının istenen küme sayısına bölümü bir tamsayı ise tüm kümeler eşit sayıda eleman içermelidir. Tamsayı değil ise son küme hariç diğer tüm kümeler eşit sayıda eleman içermelidir.

Dezavantaj: Herhangi bir amaç fonksiyonu doğrudan minimize edilememektedir.

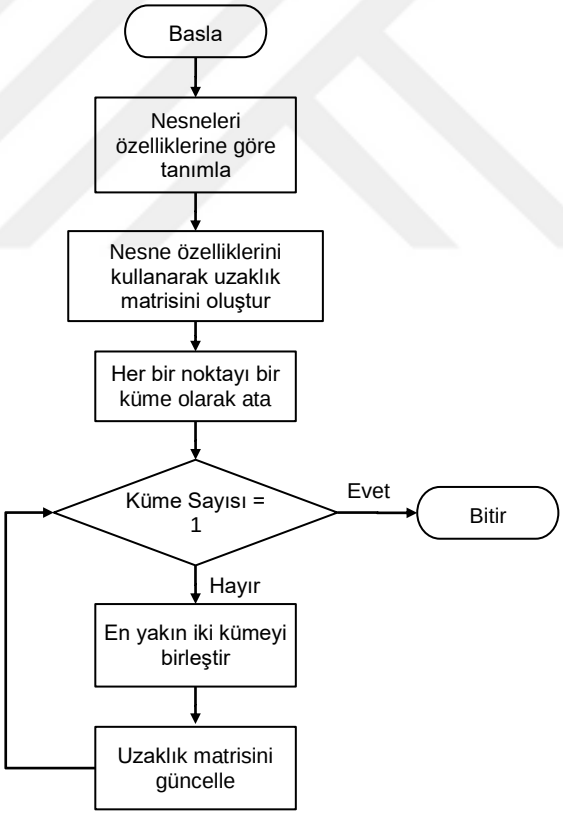
Önerilen Geliştirme: Gerçek uzaklık verileri üzerinden hareket edildiği için algoritma toplamda katedilen mesafenin minimizasyonunu hedef alacak şekilde yapılandırılmalıdır. Ayrıca küme içi en kısa rotayı bulmak için ikinci bir algoritma daha kullanılmalı ve toplamda katedilen mesafe bu çerçevede hesaplanmalıdır.

Dezavantaj: Farklı uzaklık/benzerlik ölçülerine göre farklı sonuçlar elde edilmektedir.

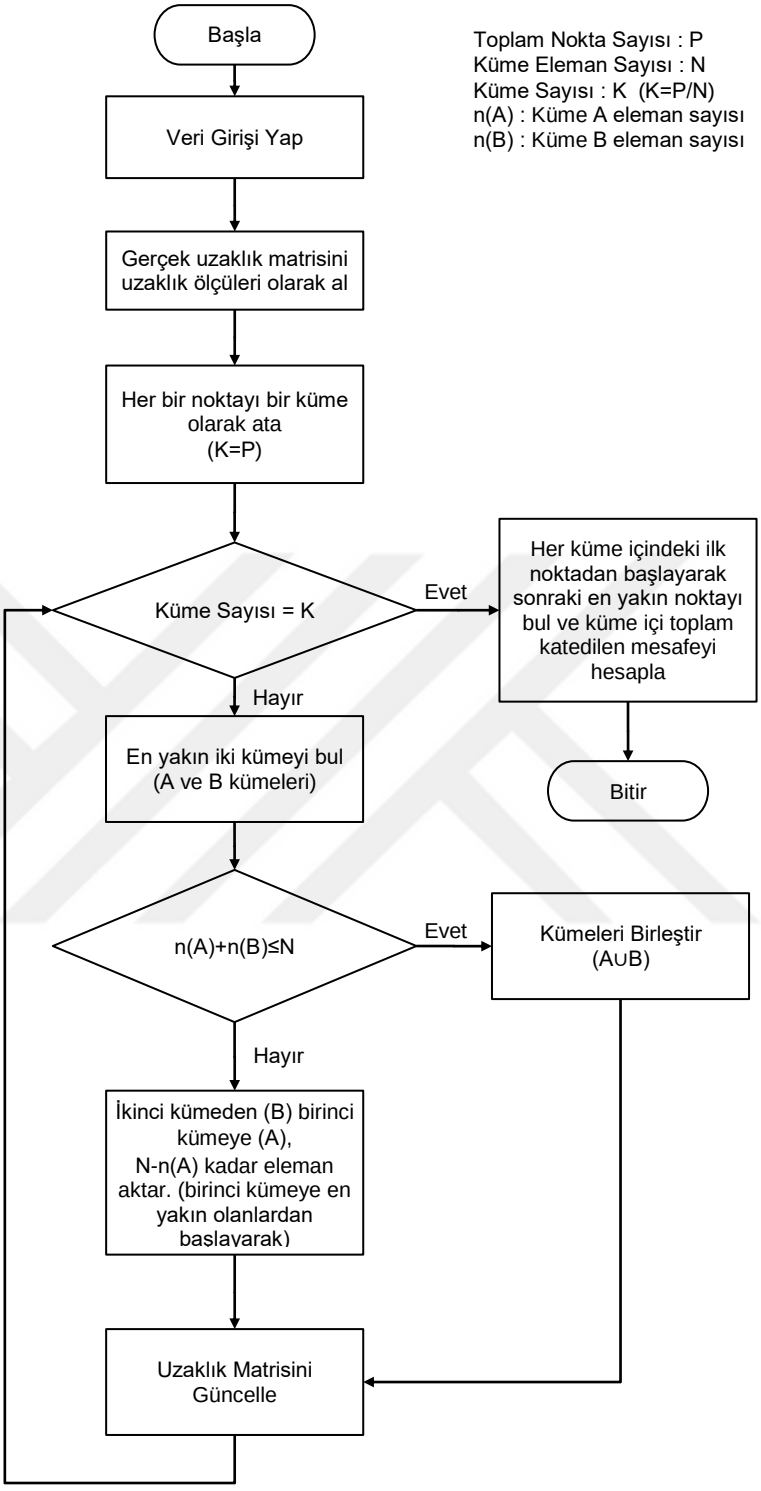
Önerilen Geliştirme: Algoritmanın standart yaklaşımında ilk adım olarak çeşitli uzaklık ölçülerinden biri (yaygın olarak da Öklid Uzaklığı) kullanılarak nesneler arası uzaklıklar hesaplanmakta ve uzaklık matrisi oluşturulmaktadır. Nesneler arası uzaklık hesaplaması yapılmadan doğrudan gerçek karayolu uzaklık verileri ile oluşturulmuş matris kullanılarak kümeleme analizine başlanmalıdır.

Kümeleme analizleri etkin olarak bilgisayar yazılımları ile gerçekleştirilmektedir. Bu konuda pek çok yazılım bulunmakla birlikte yaygın olarak kullanılanlar SPSS, R, Rapidminer, Weka, Xlstat gibi yazılımlardır. Hiyerarşik kümeleme teknikleri etkili sonuçlar üretse de hesaplama karmaşıklığı nedeniyle büyük veriler ile çalışma konusunda yetersiz kalmaktadır. Ayrıca bu yazılımlar seçilen uzaklık ölçüsüne göre uzaklık değerlerini kendisi hesapladığından gerçek karayolu uzaklık verilerini bu yazılımlar üzerinde uzaklık değerleri girdisi olarak kullanmak mümkün olmamaktadır. Bunun yanı sıra tekniklerin standart hali ile kümeler arası eleman alışverişi mümkün olmadığından yine bu yazılımlar hiyerarşik tekniklerle istenen eşit elemanlı kümeler oluşturulması ihtiyacına cevap vermemektedir. Bu nedenlerden ötürü hiyerarşik kümeleme tekniklerinden SLC ve CLC teknikleri üzerinde geliştirmeler yapılarak yeni teknikler türetilmesi ve bunun da uygulanabilmesi için bu teknikleri kullanan özel bir yazılım geliştirme zorunluluğu ortaya çıkmıştır.

Yapılan araştırma ve erişim sağlanan kaynaklar çerçevesinde bu şekilde gerçek karayolu uzaklık verilerinin girdi olarak kullanılabildiği bir yazılıma ve hiyerarşik kümeleme teknikleri ile eşit sayıda eleman içeren kümeler oluşturulmasına yönelik geliştirilen bir tekniğe rastlanmamıştır. Geliştirmeler neticesinde küme sayısının girdi parametresi olarak kullanılabildiği ve küme eleman sayısının belirlenebildiği bu teknikler standart tekniklerin kısaltılmış isimlerine getirilen Kn eki ile isimlendirilmiştir. SLC'nin geliştirilmiş haline SLC-Kn, CLC'nin geliştirilmiş haline ise CLC-Kn adı verilmiştir. Bu teknikleri kullanmak üzere Microsoft SQL platformu üzerinde amaca yönelik özel bir yazılım geliştirilmiş olup bu yazılıma da Cluster Limited (CL) adı verilmiştir. Hiyerarşik kümeleme tekniklerinin standart yapısını gösterir akış diyagramı ile geliştirilen tekniklerin temelinde yer alan algoritmanın akış diyagramı Şekil 3.6 ve Şekil 3.7'de gösterilmiştir.



Şekil 3.6. Birleştirici hiyerarşik kümeleme standart akış şeması



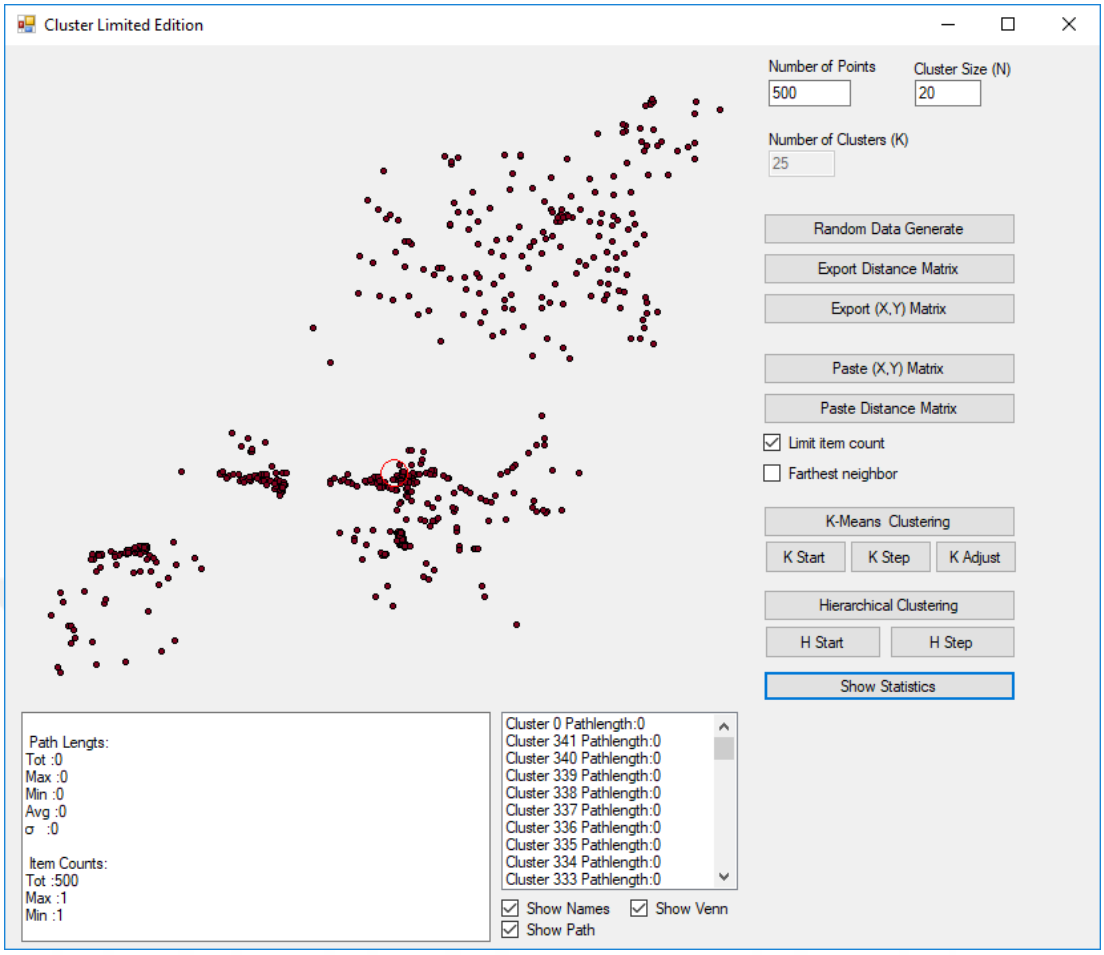
Şekil 3.7. Geliştirilen tekniğe ait algoritma akış diyagramı

3.7. Uygulama

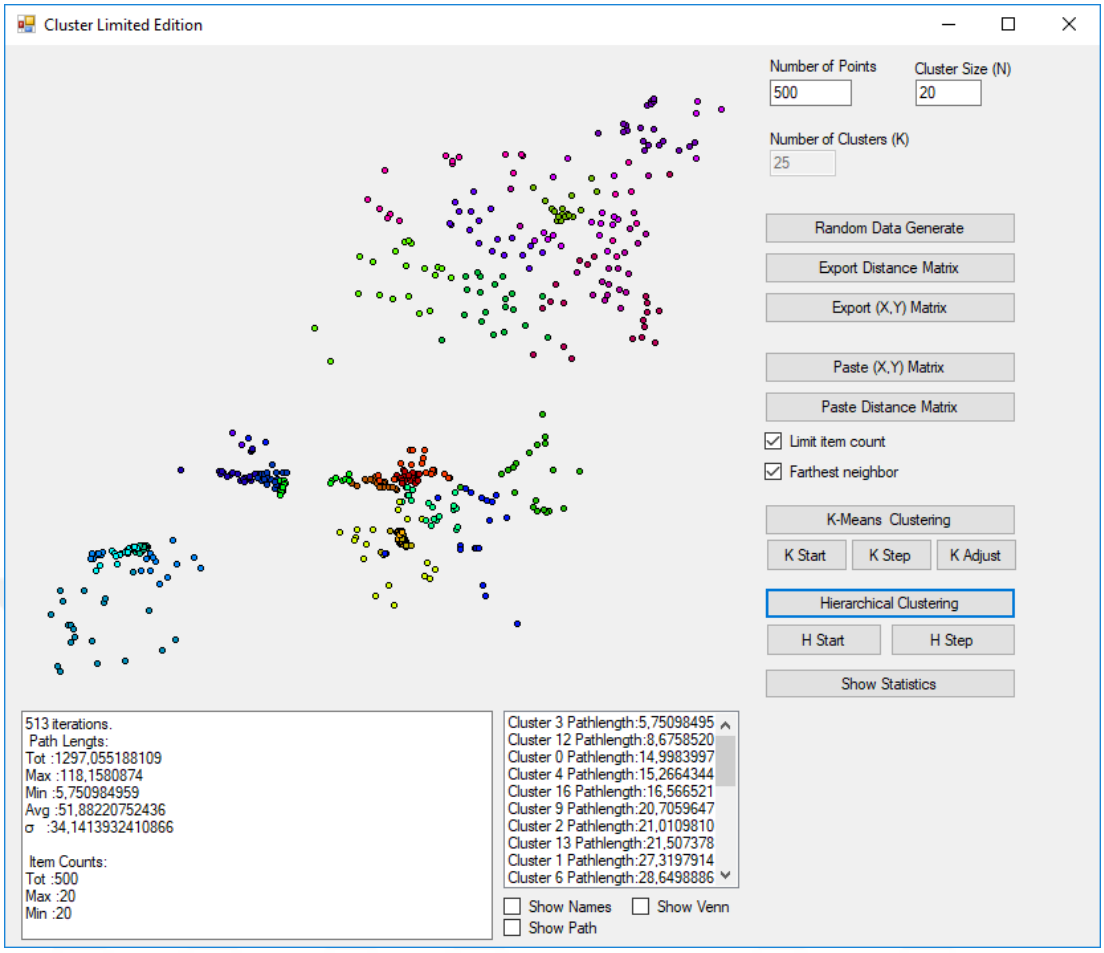
Saha projesine konu olan 8163 noktaya ait koordinat verileri bir Coğrafi Bilgi Sistemi operatörüne iletilmiş ve bir noktanın diğer tüm noktalar ile arasındaki en kısa karayolu uzaklıkları tespit edilmiştir. İşlem sonucunda oluşan gerçek karayolu uzaklık verileri 8163x8163 büyüklüğünde bir kare matris oluşturmuştur. Ancak hiyerarşik tekniklerle kümeleme analizi gerçekleştirebilmek için bu matrisin asal köşegenlerine göre simetrik olması gerektiğinden ve herhangi iki nokta arası gidiş ve geliş uzaklıkları arasında farklılıklar olabildiğinden dolayı bu farklılık gösteren noktalar arasındaki gidiş geliş mesafelerinin ortalaması alınarak uzaklık matrisi asal köşegenlerine göre simetrik hale getirilmiştir. İşlem yoğunluğunun yüksek olması nedeniyle Kocaeli ili sınırları içerisinden rassal olarak seçilen 500 noktaya ait koordinat verileri ile 500x500 boyutunda bir kare matris formunda olan uzaklık verileri CL yazılımına yüklenerek kümeleme analizi gerçekleştirilmiştir.

3.8. Uygulama Sonuçları ve Bulgular

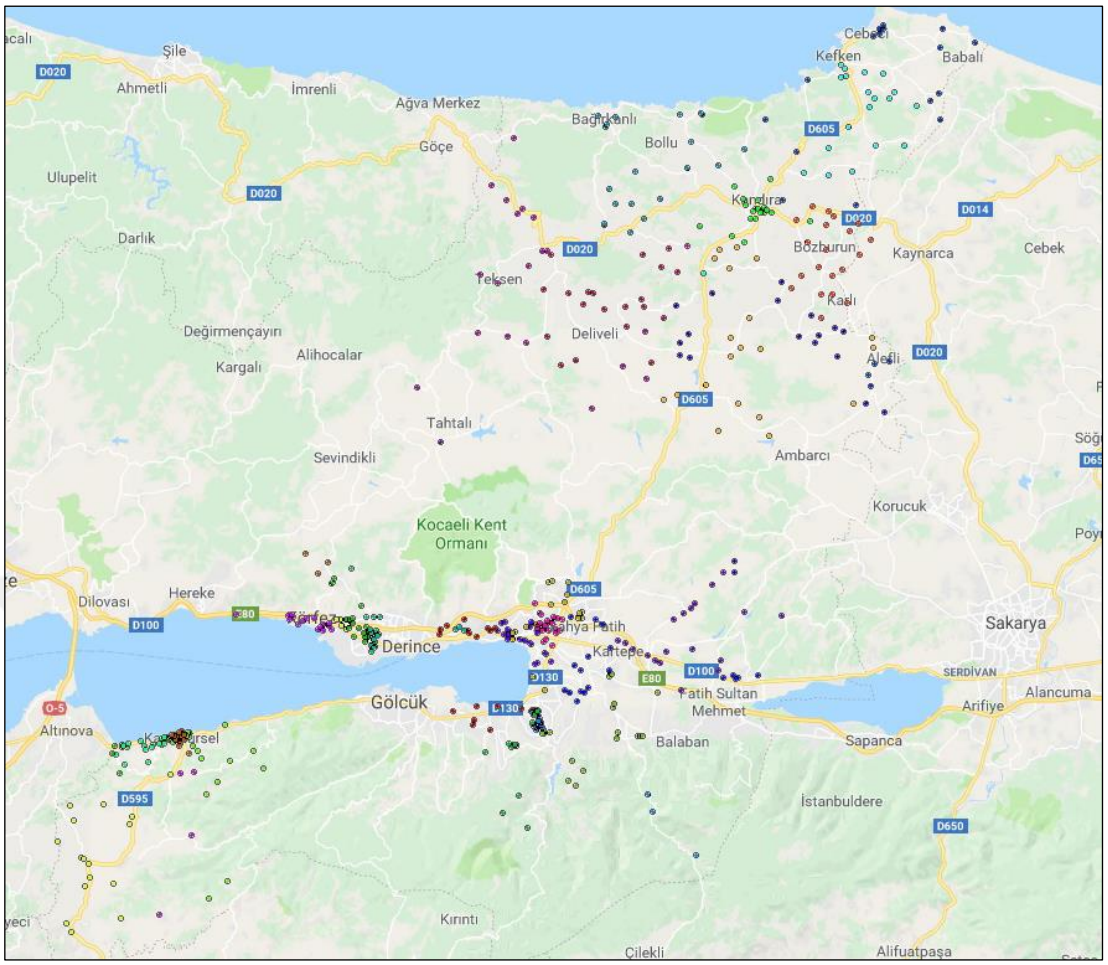
500 noktaya ait koordinat verileri ile noktalar arası uzaklıkları içeren 500x500 büyüklüğündeki matris CL programına yüklenerek SLC, CLC, SLC-Kn ve CLC-Kn teknikleri ile kümeleme analizleri gerçekleştirilmiştir. Bu analizlere ilişkin bazı CL programı ekran görüntüleri Şekil 3.8, Şekil 3.9 ve Şekil 3.10’da verilmiştir.



Şekil 3.8. Koordinatlar ve uzaklık matrisi yükleme işlemi sonrası CL programı ekran görüntüsü

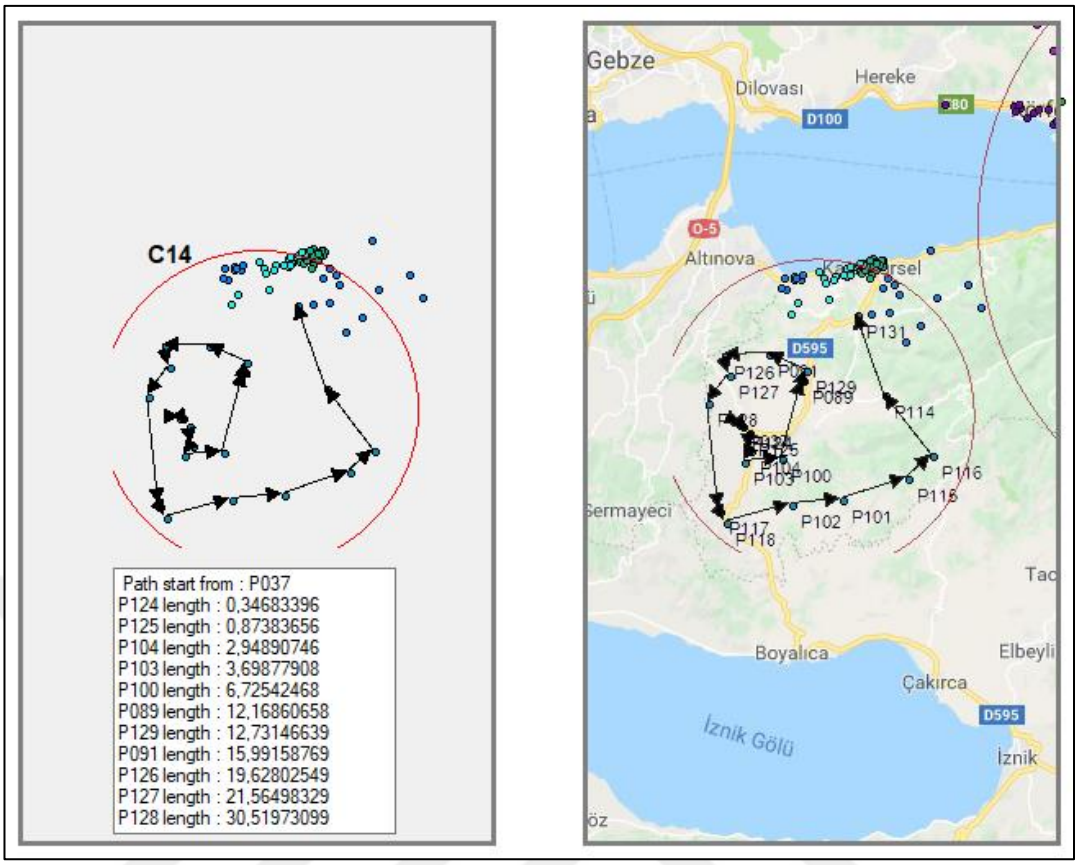


Şekil 3.9. CLC-Kn tekniği ile kümeleme analizi sonucu CL programı ekran görüntüsü



Şekil 3.10. Kocaeli ili sınırları içerisinde yer alan 500 noktanın harita üzerine uyarlanmış görüntüsü

Kümeleme analizi sonucunda girilen parametrelere göre 500 nokta 25 kümeye ayrılmış olup küme içi rota ve uzaklıkların gösterimine örnek olarak CLC-Kn tekniği ile yapılan analiz sonucu oluşan kümelerden 14 no.lu küme görüntüsü ve bunun harita üzerine uyarlanmış görüntüsü Şekil 3.10'da gösterilmiştir.



Şekil 3.11. 14 no.lu küme içi rota ve uzaklıkları gösterir CL programı ekran görüntüsü ile harita üzerine uyarlanmış görüntüsü

3.9. Kümeleme Analizi İle İlgili Kabul ve Varsayımlar

Gerçekleştirilen kümeleme analizi ile ilgili bazı kabul ve varsayımlar yapılmıştır. Bunlar aşağıda maddeler halinde belirtilmiştir.

- Analizde her noktada gerçekleştirilecek işlem süresi sabit olarak alınmıştır. Gerçek hayatta yürütülen işlerde bazı işler öngörülenden daha erken ve daha geç sürede tamamlanabilmektedir. Bu da projenin planlanan takvime göre daha erken veya daha geç bitmesi anlamına gelmektedir. Analiz ideal çözümü ortaya koymak amacıyla yürütülmüştür.

- İşlem süreleri sabit olarak alındığından amaç fonksiyonu noktalar arasında geçen seyahat süresinin minimizasyonu temeli baz alınarak dizayn edilmiştir. Araç ile karayolu üzerinde seyir hızı da sabit olarak alınmıştır. Trafik yoğunluğu, yol çalışması gibi seyahat süresini uzatıcı unsurlar göz ardı edilmiştir. Buradan hareketle

noktalar arası en kısa karayolu uzaklıklarının minimizasyonu amaç fonksiyonu olarak tanımlanmıştır.

- Gerçek hayatta iki nokta arası gidiş ve geliş mesafeleri farklılık gösterebilmektedir. Uzaklık matrisinin kare ve asal köşegenlerine göre simetrik olma zorunluluğundan ötürü tüm noktalar arası ikili uzaklık tanımlamalarında gidiş ve geliş uzaklıklarının ortalamaları alınarak yeni bir matris elde edilmiş ve uzaklık matrisi olarak bu matris kullanılmıştır.

- Her bir küme bir günlük iş programını ifade etmekte olup ekibin ikamet ettiği noktadan ilk uygulama noktasına ve son uygulama noktasından ikamet noktasına ulaşımı ile ilgili süre ve ulaşım mesafeleri kapsam dışında tutulmuştur.

4. SONUÇ VE ÖNERİLER

4.1. Kümeleme Analizi Sonuçları

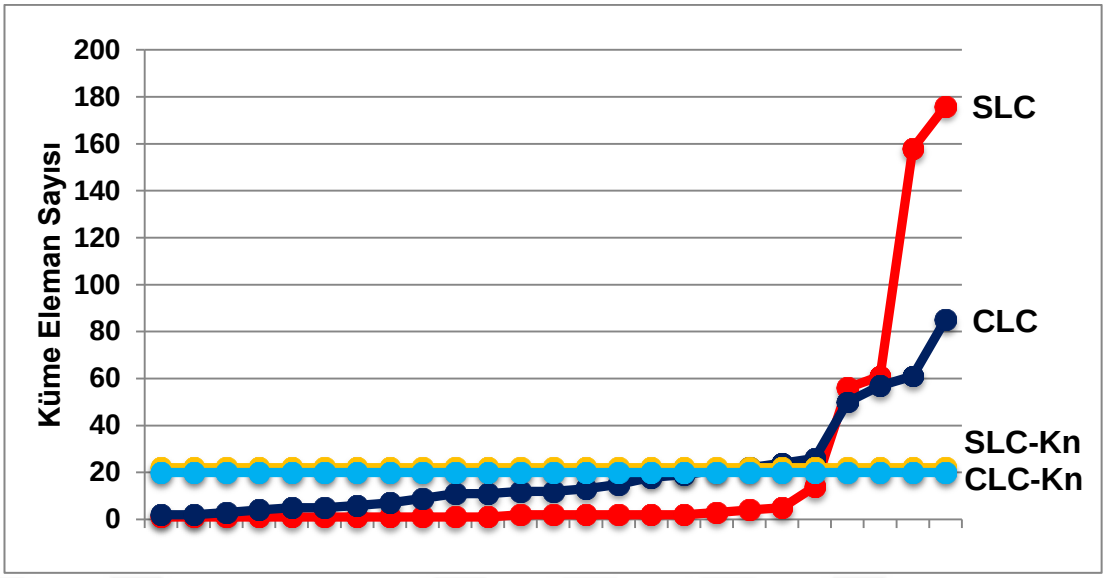
Tek Bağlantı Tekniği (SLC), Tam Bağlantı Tekniği (CLC) ve bunların geliştirilmiş şekli olan SLC-Kn ve CLC-Kn teknikleri ile Kocaeli ili sınırları içerisinden seçilen 500 noktaya ait veriler kullanılarak 4 farklı analiz gerçekleştirilmiştir. Söz konusu analiz sonuçları Tablo 4.1'de gösterilmiştir.

Tablo 4.1. Analiz Sonuçları

	Küme Eleman Sayısı				**Küme İçi Katedilen Mesafe (km)**			
Küme	**SLC**	**CLC**	**SLC-Kn**	**CLC-Kn**	**SLC**	**CLC**	**SLC-Kn**	**CLC-Kn**
C0	176	85	20	20	248,9	100,9	15,0	15,0
C1	2	50	20	20	2,6	44,9	24,9	27,3
C2	158	7	20	20	492,0	31,2	5,8	21,0
C3	1	26	20	20	0,0	48,1	16,3	5,8
C4	1	2	20	20	0,0	6,4	50,0	15,3
C5	1	4	20	20	0,0	7,2	67,3	78,4
C6	1	22	20	20	0,0	60,0	30,0	28,6
C7	61	61	20	20	87,4	87,4	94,7	74,4
C8	14	11	20	20	39,8	41,0	39,3	52,0
C9	2	6	20	20	4,3	23,2	21,2	20,7
C10	1	2	20	20	0,0	4,6	53,7	76,5
C11	2	57	20	20	4,6	69,0	28,0	36,6
C12	56	15	20	20	57,8	35,7	133,6	8,7
C13	1	21	20	20	0,0	42,7	7,2	21,5
C14	2	3	20	20	1,6	9,2	15,7	90,0
C15	1	11	20	20	0,0	27,0	71,7	55,9
C16	3	24	20	20	5,3	55,8	99,1	16,6
C17	2	13	20	20	1,7	36,1	10,4	118,2
C18	1	12	20	20	0,0	30,3	15,3	32,0
C19	1	18	20	20	0,0	52,6	127,2	108,2
C20	1	9	20	20	0,0	18,7	76,0	39,5
C21	5	12	20	20	10,3	41,6	39,5	91,0
C22	4	19	20	20	3,5	45,7	85,3	93,2
C23	2	5	20	20	2,5	10,3	67,7	92,0
C24	1	5	20	20	0,0	12,3	68,5	78,9
Toplam	**500**	**500**	**500**	**500**	**962,2**	**942,0**	**1263,5**	**1297,1**
Min	1	2	20	20	0,0	4,6	5,8	5,8
Max	176	85	20	20	492,0	100,9	133,6	118,2
Ortalama	20	20	20	20	38,5	37,7	50,5	51,9
σ	46,1	20,7	0	0	105,8	24,3	36,6	34,1
R	175	83	0	0	492,0	96,3	127,8	112,4

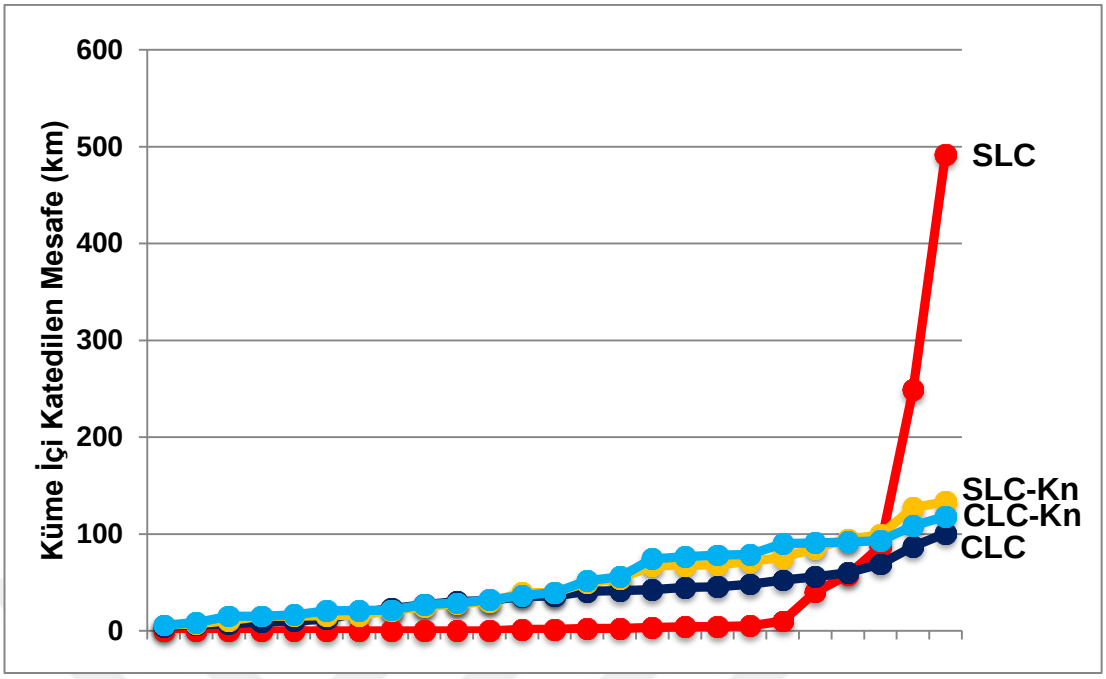
500 elemanlı veri kümesi üzerinde gerçekleştirilen çalışmada saha ekibinin bir günde 20 farklı noktada uygulama yapma kapasitesi ile 25 günlük iş planı alternatifleri araştırılmıştır. SLC ve CLC teknikleri ile yapılan analizlerde kümeleme işlemi, küme sayısı 25'e ulaştığında durdurulmuştur. SLC-Kn ve CLC-Kn tekniklerinde ise küme eleman sayısı bir parametre olduğundan analizler 500 noktanın her biri 20 eleman içeren 25 kümeye ayrılması çerçevesinde gerçekleştirilmiştir.

Analiz sonucunda SLC ve CLC teknikleri ile elde edilen sonuçlar incelendiğinde kümelerdeki eleman sayıları içerisinde büyük değişkenlik olduğu gözlenmiştir. Şöyle ki: SLC tekniğinde 11 küme sadece 1 nokta içermekte, 1 küme ise 176 eleman içermektedir. Bu teknikte algoritma sistematiğine göre tek elemanlı kümelerdeki noktalar diğer noktalardan hiçbiri ile birleşememiş ve yalnız başlarına birer küme oluşturmuşlardır. Bu teknikle yapılan analizde dikkat çeken bir diğer husus ise 176 elemanlı bir kümenin varlığıdır. Küme eleman sayısındaki bu varyasyon neticesinde eleman sayılarına ilişkin değişkenliği gösteren standart sapma değeri 46,1 olarak hesaplanmıştır. CLC tekniğinde ise en az elemana sahip kümedeki eleman sayısı 2, en çok elemana sahip kümedeki eleman sayısı ise 85 olarak kaydedilmiştir. Küme eleman sayısındaki değişkenlik 20,7 olarak ölçümlenmiş olup bu değer SLC tekniği ile elde edilen sonuçtan daha iyidir. SLC-Kn ve CLC-Kn teknikleri ile oluşan kümelerde ise her bir kümedeki eleman sayısı eşittir. Bu tekniklerle 500 nokta 20'şer elemanlı 25 ayrı kümeye ayrılmıştır. Küme eleman sayılarındaki değişimi gösteren grafik Şekil 4.1'de verilmiştir.



Şekil 4.1. Küme eleman sayıları değişimi grafiği

Küme içi katedilen mesafeler açısından bakıldığında SLC tekniği ile oluşturulan bütün kümeler içerisindeki toplam katedilen mesafe 962,2 km olarak ölçülmüştür. Bu teknikle elde edilen kümelerden 11 tanesi yalnızca bir eleman içerdiğinden küme içi katedilen mesafe bu kümelerde sıfırdır. Aynı teknikte elde edilen, içerisinde 158 eleman bulunan kümede, küme içi toplam katedilen mesafe ise 492 km olarak ölçülmüştür. CLC tekniği ile elde edilen sonuçlara bakıldığında bütün kümeler içerisindeki toplam katedilen mesafe 942,0 km olarak ölçülmüş olup küme içi katedilen mesafelerden en küçüğünün 4,6 km, en büyüğünün ise 100,9 km olduğu görülmüştür. Küme içi katedilen mesafelerde değişkenlik SLC ve CLC tekniklerinde sırasıyla 105,8 km ve 24,3 km olarak gerçekleşmiştir. SLC-Kn ve CLC-Kn teknikleri ile elde edilen sonuçlar birbirlerine oldukça yakın olup bütün kümeler içerisindeki toplam katedilen mesafe sırasıyla 1263,5 km ve 1297,1 km olarak ölçülmüştür. Küme içi katedilen mesafelerdeki değişimi gösteren grafik Şekil 4.2'de verilmiştir.



Şekil 4.2. Küme içi katedilen mesafe değişimi grafiği

Sonuçlar üzerinden bakıldığında SLC tekniğinde küme eleman sayılarındaki değişkenliğin çok fazla olduğu görülmüştür. Buna bağlı olarak küme içi katedilen mesafelerin en yüksek değeri de yine SLC tekniği sonuçları arasındadır. Bu açıdan bakıldığında CLC tekniğinin SLC tekniğine göre eleman sayıları açısından daha dengeli kümeler oluşturduğu söylenebilir. Küme eleman sayısına sınırlama getirerek eşit sayıda eleman içeren kümeler oluşturulmasına olanak sağlayan ve bu çalışma kapsamında geliştirilen ve önerilen SLC-Kn ve CLC-Kn teknikleri, birbirlerine yakın olan noktalar bir arada olacak şekilde kümeler oluşturmuştur. Toplam katedilen mesafe açısından bakıldığında SLC-Kn tekniğinde, tekniğin standart haline göre %31 düzeyinde, CLC-Kn tekniğinde ise tekniğin standart haline göre %38 düzeyinde artışlar olduğu gözlenmiştir. Toplam katedilen mesafenin minimizasyonu, saha projesi çalışmasında amaç fonksiyonu olarak belirlenmiştir. Her ne kadar toplam katedilen mesafenin en küçük değeri sezgisel bir yaklaşım olan SLC tekniği ile elde edilmiş olsa da, her kümenin saha ekibinin bir günlük çalışma programını belirlediği ve ekibin günlük uygulama kapasitesinin sınırlı olduğu göz önüne alındığında küme eleman sayısı bakımından değişim genişlikleri çok yüksek değerlerde olan SLC ve CLC teknikleri bu ve benzer saha projelerinde lokasyonların gruplandırılması için kullanılabilecek uygun teknikler olamamaktadırlar. Geliştirilen ve önerilen SLC-Kn ve CLC-Kn tekniklerinin, ekstrem değerleri yalnız başına bırakmayarak bir kümeye

dahil etmesi ve istenenden çok sayıda eleman içeren kümeleri parçalayarak az sayıda eleman içeren kümelere eleman kaydırması ve bunlar neticesinde küme eleman sayılarında bir dengeleme yaklaşımı izlemesinin doğal bir sonucu olarak bu tekniklerle yapılan analizlerde toplam katedilen mesafelerde artış gözlenmiş olup bu mesafe artışının oldukça makul düzeyde olduğu değerlendirilmiştir. Lokasyonların birbirleri ile arasındaki gerçek karayolu uzaklıkları dikkate alınarak eşit elemanlı kümeler oluşturulması hedeflenen çalışmalara çözüm üretme kabiliyeti dikkate alındığında, geliştirilen SLC-Kn ve CLC-Kn tekniklerinin oldukça başarılı, makul ve uygulanabilir sonuçlar ortaya koyduğu gözlenmiştir.

Geliştirilen tekniklerin bu çalışmada olduğu gibi geniş sahaya yayılmış işlerin/projelerin lokasyon bazlı gruplamasında, çeşitli dağıtım veya toplama faaliyetleri yapan firmaların lojistik operasyonlarında, bayilik sistemi ile çalışan firmaların ideal bayi konumlarını belirlemesi gibi uygulamalarda etkili bir şekilde kullanılabileceği düşünülmektedir.

4.2. Öneriler

Çalışmada SLC ve CLC teknikleri üzerinde geliştirme yapılarak yeni teknikler türetilmiş ve bir uygulama üzerinde denenmiştir. Benzer şekilde diğer sezgisel ve metasezgisel teknikler üzerinde de küme eleman sayısının sınırlanmasına yönelik yaklaşımlar araştırılabilir. Ayrıca küme içi daha optimal rotaların araştırılması için Gezgin Satıcı Problemi (TSP) gibi tekniklerle kümeleme tekniklerinin bir arada kullanılacağı hibrit uygulamalar da gerçekleştirilebilir.

KAYNAKLAR

[1] Çankırı S., Kartal E., Yıldırım K., Gülseçen S., Organizasyonlarda Bilgi Yönetimi Sürecinde Veri Madenciliği Yaklaşımı, *ÜNAK 2009 Bilgi Çağında Varoluş:"Fırsatlar ve Tehditler" Sempozyumu*, Yeditepe Üniversitesi, İstanbul, Türkiye, 1-2 Ekim 2009.

[2] Öğüt S., Veri Madenciliği Kavramı ve Gelişim Süreci, *Veri Madenciliği Paneli*, İstanbul, Türkiye, 5 Mart 2005.

[3] Fayyad U., Piatetsky-Shapiro G., Smyth P., From Data Mining to Knowledge Discovery in Databases, *AI Magazine*, 1996, **17**(3), 37-54.

[4] Cabena P., Hadjinian P., Stadler R., Verhees J., Kamber M., *Discovering Data Mining: From Concept to Implementation*, Prentice Hall, New Jersey, 1998.

[5] Chung H., Gray M., Special Section: Data Mining, *Journal of Management Information Systems*, 1999, **16**(1), 11-16

[6] Hui S.C., Jha G., Data Mining for Customer Service Support, *Information & Management*, 2000, **38**, 1-13.

[7] Berry M.J.A., Linoff G.S., *Mastering Data Mining*, John Wiley & Sons, New York, 2000.

[8] Grossman R.L., Kamath C., Kegelmeyer P., Kumar V., Namburu R. R., *Data Mining For Scientific and Engineering Applications*, Kluwer Academic Publishers, Netherlands, 2001.

[9] Mackinnon M. J., Glick N., Data Mining and Knowledge Discovery in Databases- An Overview, *Australian & New Zealand Journal of Statistics*, 1999, **41**(3), 255-275.

[10] Wang X., Abraham A., Smith K. A., Web Traffic Mining Using a Concurrent Neuro-Fuzzy Approach, *Soft Computing Systems: Design, Management and Applications*, Santiago, Chile, 1-4 Aralık, 2002.

[11] Inmon W. H., *Building the Data Warehouse*, John Wiley & Sons, New York, 1993.

[12] Baykal A., Application Fields of Data Mining, *D.Ü. Ziya Gökalp Eğitim Fakültesi Dergisi*, 2006, **7**, 95-107.

[13] Vatansever M., Görsel Veri Madenciliği Tekniklerinin Kümeleme Analizlerinde Kullanımı ve Uygulanması, Yüksek Lisans Tezi, Yıldız Teknik Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul, 2008, 237170.

[14] Shearer C., The Crisp-DM model: The New Blueprint For Data Mining, *Journal of Data Warehousing*, 2000, **5**, 13-22.

[15] Linacre J. M., Data Mining and Rasch Measurement CRISP-DM, *Rasch Measurement Transactions*, 2001, **15**(2), 826-833.

[16] Ristoski P., Paulheim H., Semantic Web in Data Mining and Knowledge Discovery: A Comprehensive Survey, *Web Semantics: Science, Services and Agents on the World Wide Web*, 2016, **36**, 1-22.

[17] Diler S., Veri Madenciliği Süreçleri ve Karar Ağaçları Algoritmaları ile Bir Uygulama, Yüksek Lisans Tezi, Yüzüncü Yıl Üniversitesi, Fen Bilimleri Enstitüsü, Van, 2016, 433080.

[18] Fakı B. M., Veri Madenciliği Yöntemlerini Kullanarak Anemi Sınıflandırılmasına Yönelik Bir Uygulama, Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul, 2015, 389460.

[19] Aydemir B., Veri Madenciliği Yöntemleri Kullanarak Meslek Yüksek Okulu Öğrencilerinin Akademik Başarı Tahmini, Yüksek Lisans Tezi, Pamukkale Üniversitesi, Fen Bilimleri Enstitüsü, Denizli, 2017, 486838.

[20] Arabacı G., Veri Madenciliğinde Appriori, Tahminci Appriori ve Tertius Algoritmalarının Weka ve Yale Programları ile Karşılaştırılması ve Bir Uygulama, Yüksek Lisans Tezi, İstanbul Ticaret Üniversitesi, Sosyal Bilimler Enstitüsü, İstanbul, 2007, 239276.

[21] Selvi H. Z., Çağlar B., Çok Değişkenli Haritalama İçin Kümeleme Yöntemlerinin Kullanılması, *Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi*, 2017, **6**(2), 415-429.

[22] Hofman I., Jarvis R., Robust and Efficient Cluster Analysis Using a Shared Near Neighbours Approach Proceedings, *14. International Conference on Pattern Recognition*, Washington, USA, 16-20 Ağustos 1998.

[23] Kaufman L., Rousseeuw P.J., *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley-Interscience, New York, 1990.

[24] Han J., Kamber M., *Data Mining Concepts and Techniques*, 3rd ed., Morgan Kaufmann Publishers Inc., Massachusetts, 2001.

[25] Dahal S., Effect of Different Distance Measures in Result of Cluster Analysis, Yüksek Lisans Tezi, Aalto University, School of Engineering, Helsinki, 2015.

[26] Choi S. S., Cha S. H., Tappert C. C., A Survey of Binary Similarity and Distance Measures, *Systemics, Cybernetics And Informatics*, 2010, **8**(1), 43-48.

[27] Akın Y.K., Veri Madenciliğinde Kümeleme Algoritmaları ve Kümeleme Analizi, Doktora Tezi, Marmara Üniversitesi, Sosyal Bilimler Enstitüsü, İstanbul, 2008, 221364.

[28] Ma E.W.M., Chow T. W. S., A New Shifting Grid Clustering Algorithm, *Pattern Recognition*, 2004, **37**(3), 503-514.

[29] Grabmeier J., Rudolph A., Techniques of Cluster Algorithms in Data Mining, *Data Mining and Knowledge Discovery*, 2002, **6**, 303-360.

[30] Khattre R., Naik D. N., *Multivariate Data Reduction and Discrimination with SAS Software*, 1st ed., Wiley-SAS, North Caroline, 2002.

[31] Tatlıdil H., *Uygulamalı Çok Değişkenli Analiz*, Akademi Mat, Ankara, 1996.

[32] Ergüt Ö., Uzaklık Ve Benzerlik Ölçülerinin Kümeleme Sonuçlarına Etkisi, Yüksek Lisans Tezi, Marmara Üniversitesi, Sosyal Bilimler Enstitüsü, İstanbul, 2011, 291467.

[33] Servi T., Çok Değişkenli Karma Dağılım Modeline Dayalı Kümeleme Analizi, Doktora Tezi, Çukurova Üniversitesi, Fen Bilimleri Enstitüsü, Adana, 2009, 244347.

[34] Wendly L. M., Angel, R. M., *Exploratory Data Analysis with MATLAB*, Chapman & HallPress, Florida, 2005.

[35] Sharma S., *Applied Multivariate Techniques*, John Wiley and Sons, Newyork, 1996.

[36] Anderberg M. R., *Cluster Analysis for Applications*, Academic Press, London, 1973.

[37] Everitt B. S., Landaau S., Leese M., *Cluster Analysis*, 4th ed., Oxford University Press, London, 2001.

[38] İlvan A., Mersin İli Toroslar İlçesi Örneğinde Lokal Datum Dönüşüm Parametrelerinin Belirlenmesi Üzerine Bir Çalışma, Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul, 2014, 389409.

[39] Dinler M., Kümeleme Analizi Yöntemlerinin Hayvancılık Verilerinde Karşılaştırılmalı Olarak İncelenmesi, Yüksek Lisans Tezi, Bingöl Üniversitesi, Fen Bilimleri Enstitüsü, Bingöl, 2014, 372613.

[40] Atbaş A. C. G., Kümeleme Analizinde Küme Sayısının Belirlenmesi Üzerine Bir Çalışma, Yüksek Lisans Tezi, Ankara Üniversitesi, Fen Bilimleri Enstitüsü, Ankara, 2008, 233366.

[41] Marriott, F.H.C., Practical Problems in a Method of Cluster Analysis, *Biometrics*, 1971, **27**, 501–514.

[42] http://www.holehouse.org/mlclass/13_Clustering.html (Ziyaret tarihi: 10 Mayıs 2018).

[43] http://www.data-miners.com (Ziyaret tarihi: 15 Nisan 2018).

KİŞİSEL YAYINLAR VE ESERLER

[1] **Taşatan A.**, Baynal K., Hiyerarşik Kümeleme Teknikleri ile Eşit Sayıda Eleman İçeren Kümeler Oluşturulmasına Yönelik Bir Yaklaşım ve Karayolu Uzaklık Verilerine Dayalı Kümeleme, *4. Uluslararası Mühendislik Mimarlık ve Tasarım Kongresi*, Kocaeli, 4-5 Mayıs 2018.

ÖZGEÇMİŞ

1982 yılında Gaziantep'te doğdu. İlk, orta ve lise öğrenimini Kocaeli'de tamamladı. 2000 yılında girdiği Eskişehir Osmangazi Üniversitesi Mühendislik-Mimarlık Fakültesi Endüstri Mühendisliği Bölümü'nden 2004 yılında Endüstri Mühendisi olarak mezun oldu. 2006-2016 yılları arasında özel sektörde üretim, mali işler ve finansal kontrol alanlarında yöneticilik görevlerinde bulundu. 2016 yılından beri kurucusu olduğu şirketi yönetmektedir.