aurum

AURUM MÜHENDİSLİK SİSTEMLERİ VE MİMARLIK DERGİSİ
AURUM JOURNAL OF ENGINEERING SYSTEMS AND ARCHITECTURE

Cilt 3, Sayı 1 | Yaz 2019
Volume 3, No 1 | Summer 2019, 51-64

ARAŞTIRMA MAKALESİ / RESEARCH ARTICLE

VERİ MADENCİLİĞİ YÖNTEMLERİ İLE ÜLKELERİ GELİŞMİŞLİK ÖLÇÜTLERİNE GÖRE KÜMELEME ÜZERİNE BİR UYGULAMA

Banu AKKUŞ[1]
[1]İAÜ Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı, Yüksek Lisans Öğrencisi
banuakkus8@windowslive.com ORCID No**:** 0000-0002-0040-1125

Metin ZONTUL[2]
[2]İstanbul Arel Üniversitesi, Mühendislik-Mimarlık Fakültesi, Bilgisayar Mühendisliği Bölümü, İstanbul.
metinzontul@arel.edu.tr ORCID No**:** 0000-0002-7557-2981

Geliş Tarihi/Received Date: 04/01/2019. **Kabul Tarihi/Accepted Date:** 03/04/2019.

Öz

Bir amaç doğrultusunda elde edilen verilerden anlamlı sonuçlar çıkarılması işlemine veri madenciliği denir. Kümeleme analizi de veri madenciliği alanında sıklıkla kullanılmaktadır. Bu makalede öncelikle kümeleme analizi kavramları açıklanmıştır. Çalışmada kullanılacak algoritmalar tanıtıldıktan sonra Dünya Bankası 'nın web sitesinden elde edilen verilere bu algoritmalar uygulanmıştır. Bu çalışmada amaç, önceden belirlenmiş parametreler göz önüne alınarak ülkelerin gelişmişlik ölçütlerine göre kümelenmesidir. Çalışma kapsamında 214 ülkeye ait 2015 verileri ele alınmıştır. Bu verilere Self Organizing Map ve K-Means kümeleme algoritmaları uygulanmış, sonrasında da elde edilen kümeler değerlendirilmiştir. Ayrıca ülkemizin bu kümelerdeki konumu da incelenmiştir.

Anahtar Kelimeler: Kümeleme analizi, K – Means algoritması, Self Organizing Map algoritması, İstatistik, Gelişmişlik.

AN APPLICATION ON CLUSTERING COUNTRIES WITH DATA MINING METHODS BASED ON DEVELOPMENT CRITERIA

Abstract

The process of extracting meaningful results from data obtained in the direction of a goal is called data mining. Clustering analysis is also frequently used in the field of data mining. In this article, firstly clustering analysis concepts are explained. These algorithms have been applied to the data obtained from the World Bank website after the algorithms to be used in the study have been introduced. The purpose of this study is to cluster countries according to their development criteria, taking into account pre-determined parameters. The study covered data from 214 countries. Self Organizing Map and K-Means clustering algorithms were applied to these data, and then the obtained clusters were evaluated. In addition, the position of our country in these clusters has been examined.

Keywords: Clustering Analysis, K- Means Algorithm, Self Organizing Map, Statistics, Development.

1. GİRİŞ

Sınıflar veya objelerin konsept olarak anlamlı grupları, benzer karakteristik özellikler taşırlar. Bu da insanların dünyayı analiz etmesi ve anlamlandırmasında büyük rol oynar. Aslında insanoğlu objeleri gruplama ve bu gruplara obje atama yeteneğine doğuştan sahiptir. Küçük çocuklar bile bir fotoğraf albümündeki objeleri insan, hayvan, bitki ve cansız varlıklar olarak sınıflandırabilir.

Kümeleme analizi, bir veri setini benzer özellikler taşıyan objeler aynı grupta yer alacak şekilde gruplamayı amaçlayan çok değişkenli bir metottur. Bu sayede veriler anlamlı ve kullanışlı kümelere bölünür. Bazı durumlarda ise veri özetleme gibi bir amaç için kullanışlı bir başlangıçtır.

Verilerin anlaşılması bağlamında, kümeler potansiyel sınıflardır ve kümeleme analizi de bu sınıfları otomatik olarak bulmayı sağlayan teknikler çalışmasıdır.

Kümeleme analizinin öneminin kavranabilmesi için kullanım alanlarını incelenmesi yeterlidir. Psikoloji ve diğer sosyal bilimler, biyoloji, istatistik, örüntü tanıma, makine öğrenmesi ve veri madenciliği kümeleme analizinin kullanım alanlarından sadece birkaçıdır.

Selim Çam 2014 yılında hazırladığı yüksek lisans çalışmasında, Sivas Cumhuriyet Üniversitesi Hastanesi'ne 2006-2011 yılında kayıt olmuş 18-65 yaş aralığındaki hastalara ait verileri kullanmıştır. Çalışma kapsamında K-Ortalamalar ve Yoğunluk Tabanlı Kümeleme Algoritması yöntemlerini kullanmış, demografik verilere ise Ki-Kare, Kruskal-Wallis H ve Mann-Whitney testlerini uygulamıştır. Sonuç olarak hastaları yaş, yaşadıkları yer, başvurdukları hastalık çeşitleri gibi sınıflara ayırmıştır. Çalışmanın uygulama alanı ise kayıt olan bir hasta için uygun ilaç ve personel tedariğinin doğru tespit edilebilmesidir (Çam, 2014).

Nesrin Alptekin ve Gözde Yeşilaydın'ın 2015 yılında sağlık alanında yaptığı bir çalışmada, veri kümesi olarak OECD' ye üye 34 ülkenin sağlıkla ilişkili 10 değişkeni ele alınmıştır. Çalışma kapsamında bulanık c-ortalamalar algoritması ve NCSS 10 paket programı kullanılmıştır.Farklı sayıda kümeler oluşturulmuş ve 5 küme sayısının en anlamlı olduğu tespit edilmiştir. Sonuç kümesinde Türkiye'nin Estonya, Meksika, Macaristan, Şili ve Polonya ile aynı kümede yer aldığı görülmüştür (Alptekin, Yeşilaydın, 2015).

Osman Kaya 2008 yılında hazırladığı yüksek lisans çalışmasında, özel bir kurumda çalışan 1100 personele ait 2 yıllık performans ölçütleri, bu ölçütlerin ağırlıkları,tanımı ve departman bilgisini veri kümesi olarak ele almıştır. Çalışma kapsamında bu veri kümesine c-mean ve x-mean kümeleme algoritmalarını uygulamış, 4 küme elde etmiş ve küme doğruluğunun test edilmesi için ROC eğrisinden faydalanmıştır. Çalışma sonucunda personellerin başarı ölçümünün en doğru şekilde yapılabilmesi, personelin zam ve terfi kararların adaletli olması ve kurumun başarısının değerlendirilmesini amaçlamıştır (Kaya, 2008)

Halil Darakçı 2011 yılında hazırladığı yüksek lisans çalışmasında, veri kümesi olarak özel bir akarkayıt firmasına ait 2008 ve 2009 yıllarındaki ürün ve müşteri verilerini kullanmıştır. Veri kümesi yaklaşık 48 milyon veri içerdiğinden günlük bilgiyi içerecek şekilde indirgeme yapmıştır. Uzaklık ölçümünde öklit metriği, kümeleme kısmında k-ortalamalar algoritması ve KNIME programını kullanmıştır. Sonuç olarak elde edilen kümelerdeki en verimsiz istasyonları tespit ederek kurumiçi performans değerlendirmesi yapmıştır (Darakçı, 2011).

Onur Değerli 2012 yılında hazırladığı yüksek lisans çalışmasında, veri kümesi olarak blog içeriklerini kullanmıştır. Bloglara ait içerikleri web-crawler teknolojisi ile veritabanına kaydetmiş, kelime kökünün tespit edilmesi için doğal dil işleme metotlarından faydalanmıştır. Bu içeriği Naive Bayes algoritması ile sınıflandırmış, kategorisi belli olmayan örneklerin hangi kategoriye ait olduğunu belirlemiştir. Kümeleme analizi algoritmalarının semantik web ve metin madenciliği alanında kullanılmasına dair bir çalışma yapmıştır (Değerli, 2012).

Gaffari Çelik 2013 yılında hazırladığı yüksek lisans çalışmasında veri kümesi olarak Ağrı İbrahim Çeçen Üniversitesi Meslek Yüksekokulu öğrencilerinin 2011-2012 yılına bilgilerini ele almıştır. Çalışma kapsamında K-Means, DBSCAN, OPTICS algoritmaları ve WEKA programını kullanmıştır. Çalışma sonucunda öğrenci başarısını etkileyen faktörleri tespit etmiştir. Bunlardan bazıları; öğrencinin sağlık sorununun olmaması, kardeş sayısının az olması, öğrencinin yurtta kalması ve annesi çalışmayan öğrencinin daha başarılı olması gibi çarpıcı sonuçlardır (Çelik, 2013).

Mahmut Karakaya 2012 yılında hazırladığı yüksek lisans çalışmasında, veri kümesi olarak Movielens, Jester ve Bookcrossing den aldığı verileri kullanmıştır. Çalışma kapsamında k-means ve yoğunluk tabanlı kümleme algoritmalarından faydalanmıştır. Bu çalışmada kullanıcılara müzik, film, kitap gibi öğeler önerilirken çeşitliliğin arttırılması hedeflenmiştir. Sonuç olarak, öneri sistemleri için kullanılan veri kümesine ait ortalama ve standart sapmayı değerlendiren bir çeşitlilik ölçümü geliştirmiştir (Karakaya, 2012).

Yasemin Akın 2008 yılında hazırladığı bir doktora çalışmasında, veri kümesi olarak 2004 yılında TÜİK tarafından yapılan "Hanehalkı Bütçe Anketi" verilerini kullanmıştır. Verilerin uygun olup olmadığını Ki-Kare Bağımsızlık Testi ile ölçmüştür. Çalışma kapsamında CLARA algoritması ve yoğunluk tabanlı kümeleme algoritmaları, S-Plus 2000 ve WEKA programlarını kullanmıştır. Farklı küme sayılarına ait verileri karşılaştırılarak en anlamlısının 5'li küme olduğunu tespit etmiştir. Sonuç olarak katılımcıların yaşadığı yerleşim yeri, sahip olduğu çocuk sayısı, eğitim düzeyi gibi özelliklere göre tüketicilerin harcama davranışları incelemiştir (Akın , 2008).

Tuna Vardar 2010 yılında hazırladığı yüksek lisans çalışmasında, özel bir bankadan alınan verileri ve bu verilere uygun finansal tablolardan belirlenen 13 adet değişkeni veri kaynağı olarak kullanmıştır. Çalışma kapsamında uzaklık ölçüsü olarak Öklid Metriği, küme sayısının belirlenmesinde BIC ve AIC kriterleri, kümelerin anlamlılığını ölçmek için Ki-Kare Bağımsızlık testi ve analizler için SPSS 15.0 programından faydalanmıştır. Çalışma sonucunda bankaların müşteri segmentasyonlarının bilimsel analizlerle elde edilen segmentasyonla uyuşmadığını ortaya çıkarmıştır (Vardar, 2010).

Ünzile Yılmaz 2011 yılında hazırladığı yüksek lisans çalışmasında veri kaynağı olarak Türkiye'deki 81 ilin 2008 yılına ilişkin DİE bültenlerini kullanmıştır. Çalışmada faktör analizi, kümeleme analizi yaklaşımları ve SPSS 12 programından faydalanmıştır. Çalışmanın amacı Türkiye' deki illerin gelişmişlik düzeylerine göre kümelenmesidir. Sonuç olarak 3 küme belirlemiş, İstanbul en gelişmiş iller kümesinde tek başına yer almıştır. İkinci kümede ise Bursa, İzmir, Ankara, Kocaeli yer almaktadır. Geri kalanlar ise üçüncü yani gelişmemiş iller kümesinde yer almıştır (Yılmaz, 2011).

1.1 Hiyerarşik Kümeleme Analizi

Hiyerarşik kümeleme analizinde toplamsal(agglomerative) ve parçalayıcı(disimisive) olmak üzere iki yöntem vardır.

Toplamsal yöntemde başlangıçta her nesne bir kümedir. Her adımda en yakın kümeler birleşir ve küme sayısı bir eksilir. Birleşen kümeler daha sonraki adımlarda kesinlikle ayrılmaz. Adımlar tamamlandığında, tüm nesneleri içeren tek bir küme elde edilir.

Parçalayıcı yöntemde başlangıçta tek bir küme vardır. Her adımda küme bir alt kümeye ayrılır ve küme sayısı bir artar. Ayrılan kümeler daha sonraki adımlarda kesinlikle birleşmez. Adımlar tamamlandığında her nesne bağımsız bir küme meydana getirir. Toplamsal yönteme göre daha az kullanılır.

1.2 Uzaklık Ölçümleri

Xi ve Xj gözlem vektörleri olsun, d (xi, xj) fonksiyonunun uzaklık fonksiyonu olabilmesi için aşağıdaki şartları sağlaması gerekir (Duran, Odell, 1974):

- E_p 'deki (p boyutlu öklit uzayındaki) tüm x_i ve x_j ler için d $(x_i, x_j) \geq 0$ 'dır.
- Ancak ve ancak $x_i = x_j$ ise d $(x_i, x_j) = 0$ dır.
- D (x_i, x_j) = d (x_j, x_i) dir.
- D $(x_i, x_j) \leq$ d (x_i, x_k) + d (x_k, x_j) dır. Burada x_i, x_j ve x_k vektörleri de E_p vektörlerdir.

En sık kullanılan uzaklık fonksiyonları aşağıdaki Tablo 1'deki gibidir (Duran, Odell, 1974):

Fonksiyon	Matematiksel Gösterim
Öklit	$d_2(x_i, x_j) = (\sum_{k=1}^{p}(x_{ki} - x_{kj})^2)^{1/2}$
B_1 norm	$d_1(x_i,x_j) = (\sum_{k=1}^{p}\lceil x_{ki} - x_{kj} \rceil)$
Sup-norm	$d_\infty(x_i,x_j) = sup_{k=1,2,\ldots,p}\{\lvert x_{ki} - x_{kj} \rvert\}$
B_p norm	$d_p(x_i,x_j) = \sum_{k=1}^{p} \lvert x_{ki} - x_{kj} \rvert^{p}]^{1/p}$
Mahalanobis	$D^2 = (\overline{x_A} - \overline{x_B})' W^{-1} (\overline{x_A} - \overline{x_B})$

Tablo 1. Uzaklık Fonksiyonları ve Matematiksel Gösterimleri

Benzerlik matrisi, gözlemlerin birbirine olan uzaklıklarından oluşan simetrik kare matristir. Nesnelerin birbirine olan uzaklığının 0 olduğu aşikardır. Nesneler x_i ve x_k ile ifade edilsin. (i ve k N kümesinin elemanlarıdır.) Uzaklık fonksiyonu d (x_i, x_k) ile ifade edilir. x_i ve x_k nın birbirine uzaklığı, x_k ve x_i nin birbirine uzaklığına eşittir. Bu nedenle benzerlik matrisi simetriktir. Bu bilgilere göre oluşturulan benzerlik matrisi Şekil 1'deki gibidir:

$$\begin{bmatrix} d_{11} & \cdots & \\ \vdots & \ddots & \vdots \\ d_{n1} & \cdots & d_{nn} \end{bmatrix}$$

Şekil 1. Benzerlik Matrisi Gösterimi

Nesnelerin birbirine olan uzaklığı ne kadar az ise aynı kümede olma ihtimalleri de o ölçüde fazladır.

1.3 Hiyerarşik Kümeleme Analizi Algoritmaları

1.3.1 Tek bağlantı tekniği

En yakın komşuluk tekniği olarak da bilinir. Bu teknikte, iki küme arasında birbirine en yakın elemanların uzaklığı, kümeler arasındaki mesafe olarak kabul edilir. Birbirine en yakın iki gözlem bulunur ve bu şekilde ilerlenerek kümeler oluşturulur. Algoritmanın zaman karmaşıklığı $O(n^2)$'dir (Manning ve ark., 2008)

1.3.2 Tam bağlantı tekniği

En uzak komşuluk tekniği olarak da bilinir. Bu teknikte, iki küme arasında en uzak elemanların uzaklığı, kümeler arasındaki mesafe olarak kabul edilir. Birbirine en uzak iki gözlem bulunur ve bu şekilde ilerlenerek kümeler oluşturulur.

1.3.3 Ortalama grup bağlantı tekniği

Bu teknikte, iki kümedeki elemanların uzaklıklarının ortalaması, kümeler arasındaki mesafe olarak kabul edilir. Bu yöntem, biyoloji alanında türlerin ortak kökenleri araştırmalarında kullanılmaktadır.

1.4 K-Ortalama Yöntemi

K-Means kümeleme metodu olarak bilinir. K- Means algoritması uygulama kolaylığından dolayı, en çok kullanılan algoritmalardan biridir. Büyük ölçekli veriler için kullanışlıdır. Buradaki k küme sayısıdır. Küme içi benzerliğin yüksek fakat kümeler arası benzerliğin düşük olması amaçlanır. (Yıldız, Çamurcu, Doğan, 2010)

K-Means algoritmasının adımları aşağıdaki gibidir:

1. K küme sayısı başlangıçta belirlenir.
2. K adet başlangıç noktası rastgele seçilir.
3. Küme merkezleri belirlenir. Burada uzaklık öklid uzaklığı kullanılarak ölçülür.
4. Her nesne en yakın olduğu kümeye atanır.
5. Küme merkezi yeniden ölçülür ve yeniden atama yapılır.
6. Nesnelerin yerleri artık değişmeyene kadar önceki adımlar tekrarlanır.

K-Means kümeleme analizi Karesel Hata Kriteri yardımıyla ölçülür. Başarı kriteri, karesel hatayı en küçük yapacak k adet kümenin elde edilmesidir.

K-Means algoritmasının dezavantajı, başlangıçta K sayısının belirlenme zorunluluğudur. K sayısının belirlenmesi kolay değildir. Algoritma farklı K değerleri için uygulanır ve sonuçlar doğruluk analizleri ile sınanır. Algoritma farklı K değerleri için çalıştırıldığında çok farklı sonuçlar üretebildiği için kararlı değildir.

1.5 Self Organizing Map (SOM) Yöntemi

SOM, bir gözetimsiz öğrenme algoritmasıdır. Yüksek boyutlu verilerin bir, iki veya üç boyutlu görselleştirilmesinde kullanılan bir yöntemdir. Öğrenme ve tahmin safhalarından oluşur. Öğrenme safhasında, harita oluşturulur ve eğitim verileri rekabetçi bir süreçten geçirilerek ağ oluşturulur. Tahmin safhasında, yakınsama haritasında yeni vektörlere bir lokasyon verilir ve yeni veriler hızlıca kategorilere ayrılır.

Öğrenme sürecinin adımları aşağıdaki gibidir:

1. Her düğüm için ağırlıklar belirlenir.
2. SOM'u temsil etmesi için eğitim verisinden rastgele bir vektör seçilir.
3. Girdi vektörü ile her vektörün ağırlığı arasındaki uzaklık hesaplanarak en iyi eşleşen birim (BMU-Best Matching Unit) bulunur.
4. BMU çevresindeki komşuluk yarıçapı hesaplanır. Komşuluk ölçüsü her tekrarlamada azaltılır.
5. BMU'nun komşuluğundaki her düğümün ağırlıkları, BMU'ya daha çok benzemek için ayarlanmıştır. BMU'ya en yakın düğümler komşulukta en uzaktaki düğümlerden daha fazla değiştirilir.
6. İkinci adımdaki işlemler yakınsama gerçekleşene kadar tekrar edilir.

2. VERİ MADENCİLİĞİ YÖNTEMLERİ İLE ÜLKELERİ GELİŞMİŞLİK ÖLÇÜTLERİNE GÖRE KÜMELEME

Dünya Bankası'nın internet sitesinden alınan veriler üzerinde, kümeleme algoritmalarından K-Means ve SOM uygulanarak ülkeler değerlendirilmiştir. Tezin amacı kümeleme algoritmaları kullanılarak önceden belirlenmiş parametrelere göre ülkelerin aldığı değerlerin karşılaştırılması ve anlamlı kümeler oluşturulmasıdır. Dünya Bankası verilerinin bulunduğu internet sitesinden 2015 yılına ait veriler incelenebilmesi için excel formatında indirilmiştir. Aşağıdaki Tablo 2'de, çalışmada kullanılan değişkenler ve kısaltmaları yer almaktadır.

Değişen Kısaltması	Değişken Adı
Tarım_GDP	Agriculture, value added (% of GDP) [NV.AGR.TOTL.ZS]
Ölüm_Oranı	Mortality rate, under-5 (per 1,000 live births) [SH.DYN.MORT]
GDP_endeksi	GDP per capita (current US$) [NY.GDP.PCAP.CD]
DısBorc_GNI	External debt stocks (% of GNI) [DT.DOD.DECT.GN.ZS]
Banka_Sube_Say	Commercial bank branches (per 100,000 adults) [FB.CBK.BRCH.P5]
Internet_Kullanıcı_Say	Internet users (per 100 people) [IT.NET.USER.P2]
IsYapılabilirlik	Ease of doing business index (1=most business-friendly regulations) [IC.BUS.EASE.XQ]
İnsan_Hakları_Endeksi	Strength of legal rights index (0=weak to 12=strong) [IC.LGL.CRED.XQ]
Teknoloji_İhracatı	High-technology exports (% of manufactured exports) [TX.VAL.TECH.MF.ZS]
Kadın_Milletvekili	Proportion of seats held by women in national parliaments (%) [SG.GEN.PARL.ZS]
Ihracat_GDP	Exports of goods and services (% of GDP) [NE.EXP.GNFS.ZS]
Temiz_Su	Improved water source, urban (% of urban population with access) [SH.H2O.SAFE.UR.ZS]

Tablo 2. Değişkenler ve kısaltmaları

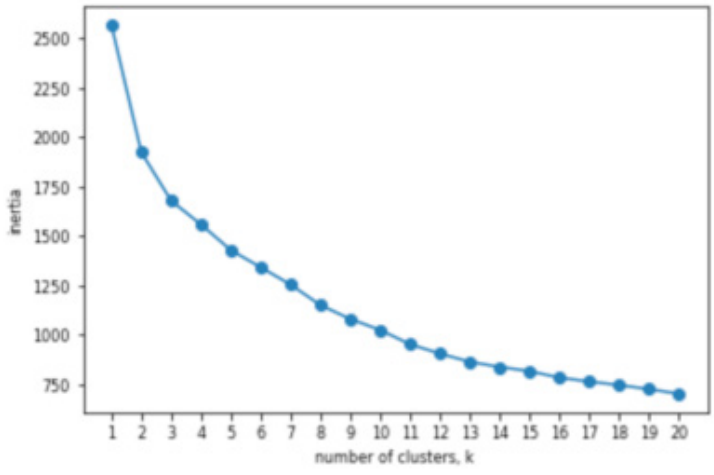
2.1 K-Means Algoritması Kullanılarak Yapılan Kümeleme

2.1.1 Küme sayısının belirlenmesi

K-Means algoritmasında uygun küme sayısının belirlenmesi için inertia(atalet) kriterinden faydalanılmıştır. İnertia'nın diğer bir ismi ise küme içi kareler ölçütü toplamıdır. Kümeleme algoritması sonucunda oluşan kümelerin tutarlılık ölçüsü olarak algılanabilir.

İnertia algoritması veriye uygulandıktan sonra oluşan noktalar incelenir. İki nokta arasındaki uzaklıkların hangi noktadan itibaren fazla değişmediği belirlenir.

Aşağıdaki Şekil 14'de, İnertia kriterinin Dünya Bankası'nın web sitesinden alınan verilere uygulanarak elde edilmiştir. Grafikte görüldüğü üzere uygun küme sayısı 16 olarak belirlenmiştir.



Şekil 2. İnertia Kriterinin Verideki Sonuç Grafiği

Dünya Bankasının web sitesinden alınan verilere K-Means algoritması uygulanarak elde edilen onaltı küme aşağıdaki gibidir.

2.1.2 K-Means Algoritması Sonucu Oluşan Kümeler

Dünya Bankasının web sitesinden alınan verilere K-Means algoritması uygulanarak elde edilen onaltı küme aşağıdaki gibidir.

Sıfırıncı küme: Iran, Islamic Rep., Venezuela, RB, Morocco, Egypt, Arab Rep., Sri Lanka, Paraguay, Jordan, Azerbaijan, Dominican Republic, Arab World, Lebanon, Maldives, Brazil, Turkey, Suriname, Chile, Argentina, Antigua and Barbuda, Oman, Saudi Arabia, Bahrain, Bahamas, The, Kuwait, Qatar

Birinci küme: Burundi, Malawi, Madagascar, Guinea, Mozambique, Ethiopia, Rwanda, Nepal, Uganda, Tanzania, Tajikistan, Senegal, Zimbabwe, Timor-Leste, Bangladesh, Cameroon, Kenya, Pakistan, South Asia, Sub-Saharan Africa, Lao PDR, Sudan, Algeria, Guyana,Turkmenistan

İkinci küme: Cambodia, Ghana, India, Solomon Islands, Uzbekistan, Honduras, Vanuatu, Micronesia, Fed. Sts., Tuvalu, Indonesia, Marshall Islands, Guatemala, Samoa, Tonga, Fiji, Botswana, Nauru

Üçüncü küme: Korea, Rep., Israel, France, New Zealand, Germany, Finland, Canada, Austria, United Kingdom, Iceland, Sweden, Denmark, North America, United States, Australia, Norway, Macao SAR, China, Switzerland

Dördüncü küme: Cuba, Andorra, Nicaragua, Bolivia, Tunisia, Namibia, Macedonia, FYR, Serbia, South Africa, Belarus, Ecuador, Grenada, Seychelles, Portugal, Spain, Italy

Beşinci küme: San Marino, Colombia

Altıncı küme: Eritrea, Mauritania, South Sudan, Haiti, West Bank and Gaza, Angola, Equatorial Guinea

Yedinci küme: Malta, Hong Kong SAR, China, Singapore, Ireland, Luxembourg

Sekizinci küme: Kyrgyz Republic, Moldova, Ukraine, Bhutan, Cabo Verde, Armenia, Georgia, Albania, Bosnia and Herzegovina, Belize, Jamaica, Mauritius, Panama

Dokuzuncu küme: Vietnam, Hungary, Central Europe and the Baltics, Poland, Latvia, Lithuania, Slovak Republic, Estonia, Czech Republic, Slovenia, Cyprus, Belgium, United Arab Emirates, Netherlands

Onuncu küme: Central African Republic, Niger, Congo, Dem. Rep., Togo, Afghanistan, Burkina Faso, Comoros, Sierra Leone, Mali, Benin, Chad, Cote d'Ivoire, Nigeria

Onbirinci küme: Liechtenstein, Bermuda, Korea, Dem. People's Rep., Virgin Islands (U.S.), Guam, Cayman Islands, French Polynesia, New Caledonia, Greenland, Aruba, Monaco, Thailand, St. Vincent and the Grenadines, Dominica, St. Lucia, China, Latin America & Caribbean, Russian Federation, East Asia & Pacific, Caribbean small states, Croatia, Barbados, Uruguay, St. Kitts and Nevis, Trinidad and Tobago, Greece, Brunei Darussalam, Japan

Onikinci küme: Sao Tome and Principe, Philippines, Malaysia, Kazakhstan, Palau

Onüçüncü küme: Mongolia

Ondördüncü küme: Libya, Papua New Guinea, Syrian Arab Republic, Gambia, The, Liberia, Somalia, Guinea-Bissau, Lesotho, Myanmar, Yemen, Rep., Zambia, Kiribati, Djibouti, Congo, Rep., Swaziland, Iraq, Gabon

Onbeşinci küme: Puerto Rico, Kosovo, El Salvador, Peru, Montenegro, Bulgaria, Romania, Mexico, Costa Rica

2.2 Self Organizing Map (SOM) Algoritması ile Kümeleme

Dünya Bankasının web sitesinden alınan verilere SOM algoritması uygulanarak elde edilen onaltı küme aşağıdaki gibidir.

Küme (1, 2): Liechtenstein, Virgin Islands (U.S.), Guam, Cayman Islands, French Polynesia, Greenland, Vietnam, South Africa, Thailand, Latin America & Caribbean, Argentina, Barbados, Seychelles, Uruguay

Küme (0, 2): Bermuda, Cuba, Philippines, China, Mexico, East Asia & Pacific, Costa Rica, Palau, Trinidad and Tobago

Küme (3, 1): Korea, Dem. People's Rep., Sri Lanka, Samoa, El Salvador, Peru, Nauru

Küme (2, 3): Puerto Rico, Monaco, Russian Federation, Croatia, St. Kitts and Nevis, Greece, Portugal, Bahrain, Cyprus

Küme (2, 2): New Caledonia, Aruba, Tunisia, Jordan, Azerbaijan, Belarus, Dominica, St. Lucia, Lebanon, Maldives, Turkey, Caribbean small states, Chile, Antigua and Barbuda, Oman, Saudi Arabia, Bahamas, The, Kuwait

Küme (1, 1): Libya, Syrian Arab Republic, Venezuela, RB, Nicaragua, Egypt, Arab Rep., Guyana, Ecuador, Dominican Republic, Arab World, Suriname

Küme (1, 3): Andorra, Hungary, Central Europe and the Baltics, Poland, Latvia, Lithuania, Slovak Republic, Slovenia, Spain, Italy, Brunei Darussalam, Japan

Küme (2, 0): Papua New Guinea, Yemen, Rep., Zambia, Ghana, India, Congo, Rep., Swaziland, Indonesia, Gabon

Küme (2, 1): Iran, Islamic Rep., Morocco, Paraguay, Fiji, St. Vincent and the Grenadines, Brazil

Küme (0, 0): Eritrea, Mauritania, Burundi, Madagascar, Central African Republic, Niger, Congo, Dem. Rep., Somalia, Guinea, Mozambique, Togo, Afghanistan, Burkina Faso, Ethiopia, Nepal, Uganda, South Sudan, Sierra Leone, Haiti, Mali, Benin, Tanzania, Chad, Cameroon, Kenya, Pakistan, Sub-Saharan Africa, Lao PDR, Sudan, Nigeria, West Bank and Gaza, Angola, Equatorial Guinea

Küme (3, 3): San Marino, Kyrgyz Republic, Moldova, Ukraine, Georgia, Mongolia, Bosnia and Herzegovina, Macedonia, FYR, Jamaica, Serbia, Colombia, Montenegro, Bulgaria, Romania, Mauritius, Panama

Küme (1, 0): Malawi, Gambia, The, Liberia, Guinea-Bissau, Comoros, Tajikistan, Zimbabwe, Myanmar, Bangladesh, Kiribati, Cote d'Ivoire, South Asia, Djibouti

Küme (0, 1): Rwanda, Senegal, Lesotho, Timor-Leste, Sao Tome and Principe, Bolivia, Algeria, Namibia, Iraq, Turkmenistan

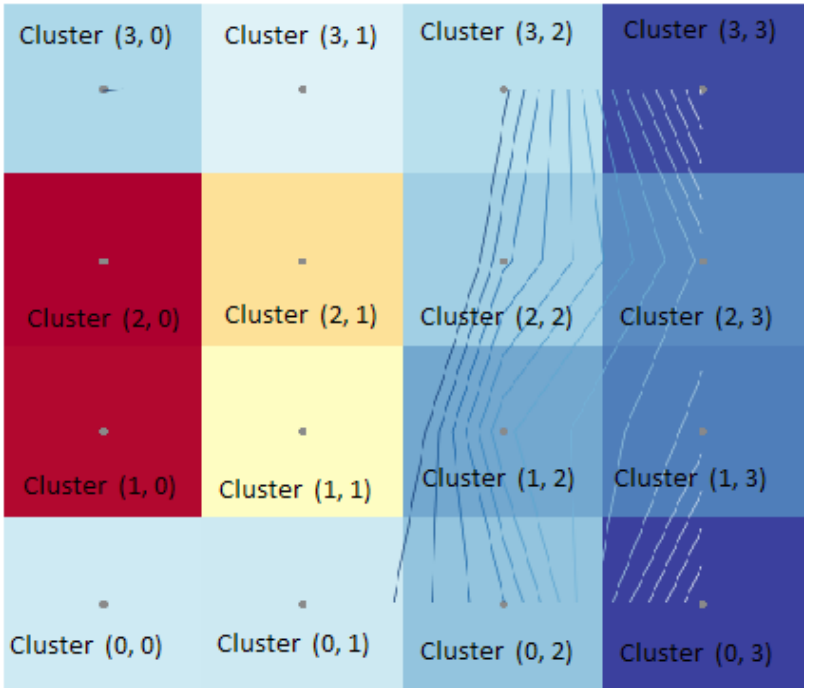
Küme (3, 0): Cambodia, Solomon Islands, Uzbekistan, Honduras, Vanuatu, Micronesia, Fed. Sts., Tuvalu, Kosovo, Marshall Islands, Guatemala, Tonga, Botswana

Küme (3, 2): Bhutan, Cabo Verde, Armenia, Albania, Belize, Grenada

Küme (0, 3): Malaysia, Kazakhstan, Estonia, Czech Republic, Malta, Korea, Rep., Israel, France, New Zealand, Belgium, United Arab Emirates, Germany, Finland, Hong Kong SAR, China, Canada, Austria, United Kingdom, Netherlands, Iceland, Sweden, Denmark, Singapore, North America, United States, Australia, Ireland, Qatar, Norway, Macao SAR, China, Switzerland, Luxembourg

2.2.1 SOM Algoritması Sonuçlarının Görselleştirimesi

SOM algoritması görselleştirilirken en çok kullanılan yöntemlerden biri bileşen düzlemleri yöntemidir. Bu görselleştirme şeklinde haritanın rengi, hedefin ortalamasına göre göreli değerini belirtir. Parlak kırmızı yüksek bir değeri, mavi ise düşük bir seviyeyi belirtir. Bazı renkler bazılarından daha parlaktır: dinamik aralık, simüle edilmiş boş hipotez ile belirlenir (Anonim 2013).



Şekil 3. Bileşen düzlemleri grafiğinin isimlendirilmiş hali

2.2.2 SOM Sonuçlarının Kalitesinin Ölçülmesi

Giriş verileri için her zaman en uygun harita mevcut olsa da, başlangıçtan itibaren doğru parametreleri seçmek zordur. Farklı parametreler ve başlatmalar farklı haritalara neden olduğundan, haritanın eğitim verilerine düzgün şekilde adapte olup olmadığını bilmek önemlidir. Haritanın kalitesini belirlemek, uygun öğrenme parametrelerini ve harita boyutlarını seçmek için yaygın olarak kullanılan iki kalite önlemi, ortalama nicemleme hatası(Average quantization error) ve topografik hata(Topographic error) dır (MOVSOM Research Lab).

Nicemleme Hatası, geleneksel olarak tüm vektör kuantizasyon ve kümeleme algoritmaları ile ilgilidir. Dolayısıyla, bu ölçüm tamamen harita topolojisini ve hizalamayı gözardı eder. Nicemleme hatası, örnek vektörlerin temsil edildiği küme merkezlerine ortalama mesafesini belirleyerek hesaplanır. SOM durumunda, küme merkezleri prototip vektörlerdir (Polzlbauer, 2004).

Verilen herhangi bir veri kümesi için, harita düğümlerinin sayısını arttırmakla nicemleme Hatası azaltılabilir, çünkü veri örnekleri daha seyrek olarak harita üzerinde dağıtılır.

Topografik Hata, topoloji koruma önlemlerinin en basitidir. Bu hesaplama şu şekilde yapılır: Tüm veri örnekleri için en ilgili ve en iyi eşleşen birimler belirlenir. Harita kafesinde bunlar bitişik değilse, bu bir hata olarak kabul edilir. Toplam hata daha sonra 0 ile 1 arasında bir aralıkta normalize edilir. Burada 0, mükemmel bir topoloji koruması anlamına gelir.

Elde edilen kümeler için topographic değer 0.056074766355140186 olarak bulunmuştur. Bulunan değer 0 ile 1 aralığında olduğundan topolojinin korunduğu sonucuna varılır.

3. SONUÇ VE ÖNERİLER

Bu çalışmada Dünya Bankası'nın web sitesinden alınan verilere kümeleme analizi yöntemlerinden olan K-Means ve Self Organizing Map algoritmaları uygulanmıştır. Bu algoritmalar sonucunda oluşan kümeler ve Türkiye' nin bu kümelerdeki yeri incelenmiştir.

K-Means algoritması sonucunda ülkemiz İran, Venezuela, Mısır, Sri Lanka, Paraguay, Ürdün, Azerbaycan, Dominik Cumhuriyeti, Arap Devletleri, Lübnan, Maldivler, Brezilya, Surinam, Şili, Arjantin, Antigua ve Barbuda, Umman, Suudi Arabistan, Bahreyn, Bahama Adaları, Kuveyt ve Katar ile aynı kümede yer almaktadır. Bu küme ağırlıkla Orta Doğu ülkelerinden oluşmaktadır.

K-Means algoritması sonucunda, her bir parametre için küme merkezleri incelenmiştir. Bu değerlere genel olarak bakıldığında ülkemizin genel olarak iyi bir noktada olduğu söylenebilir. Türkiye'nin daha iyi bir noktaya gelebilmesi için T.C. Kalkınma Bakanlığı tarafından kalkınma planları düzenlenmektedir. Güncel kalkınma planı 2014-2018 yıllarını kapsayan Onuncu Kalkınma Planı'dır (Kalkınma Bakanlığı, 2014).

SOM algoritması sonucunda ülkemiz Yeni Kaledonya, Aruba, Tunus, Ürdün, Azerbaycan, Belarus, Dominika, Saint Lucia, Lübnan, Maldivler, Karayipler, Şili, Antigua ve Barbuda, Umman, Suudi Arabistan, Bahama

Adaları ve Kuveyt ile aynı kümede yer almaktadır. Bu küme ağırlıkla Amerika kıtasındaki gelişmekte olan ülkelerden oluşmaktadır.

Değişkenlere ilk bakıldığında tarım parametresi ön plana çıkar. Sahip olduğumuz coğrafi konum ve iklim göz önüne alındığında, tarım alanı bu parametreler arasından en hızlı sonuç alınabilecek parametredir. Onuncu Kalkınma Planı'nda bu sektörün yıllık ortalama büyüme hızının yüzde 3,1 olması, toplam istihdam içerisindeki payının yüzde 21,9'a gerilemesi ve GSYH içerisindeki payının ise yüzde 6,8 olması beklenmektedir.

Bebek ölüm oranı parametresinde, ülkemiz diğer ülkelerin ortalamasına bakıldığında iyi bir noktadadır. Bu değerin daha da düşürülmesi için ana çocuk sağlığı ve aile planlama merkezleri sayısı arttırılmalıdır.

Kişi başına düşen milli gelir miktarına bakıldığında ülkemiz ortalamaya yakın bir değere sahiptir. Bu değer vatandaşların bireysel mutluluğuna etki eden en önemli parametrelerden biridir. Onuncu Kalkınma Planı'nda 2023 yılı hedefi olarak kişi başına düşen milli gelirin 16 bin dolara ulaşması beklenmektedir.

Dış Borç parametresi, Türkiye'nin bu parametreler içinde en çok geliştirmesi gereken alandır. Onuncu Kalkınma Planı'nda cari açığın kademeli olarak 5,2 seviyesine gerilemesi hedeflenmiştir.

Her yüz bin kişiye düşen banka şube sayısı parametresine bakıldığında, ülkemizin ortalamanın üzerinde bir değere sahip olduğu görülür. Türkiye' nin bankacılık sektöründe uluslararası standartlara sahip olması için Basel II standartları 2012 yılından itibaren uygulanmaktadır. Ayrıca Onuncu Kalkınma Planı dönemi sonunda İstanbul'un Küresel Finans Merkezleri Endeksinde ilk 25 içinde yer alması hedeflenmektedir.

İnternet kullanıcı sayısı parametresinde ülkemiz ortalamanın üzerinde bir değere sahiptir. 2009 yılından bu yana 3G hizmeti verilmeye başlanmış ve abone sayısı 12 milyonu aşmıştır.

İş yapılabilirlik parametresine bakıldığında ülkemizin bu alanda gelişmesi gerektiği görülür. Bu hedef için Onuncu Kalkınma Planı'nda İş ve Yatırım Ortamının Geliştirilmesi Programı oluşturulmuştur.

İnsan hakları parametresinde Türkiye 12 üzerinden 3 almıştır. Gelişmişlik düzeyi incelemesinde sosyal anlamda en önemli parametrelerden biri insan haklarıdır. 2013 yılında dördüncü yargı reformu paketi kabul edilmiştir. Bu pakette AİHM'in "yeniden yargılama" kararlarının tümü uygulanabilir olmuştur.

Teknoji ihracatı alanında Türkiye son yıllarda gelişme göstermektedir. Her geçen yıl yüksek teknolojiye yatırım yapan firma sayısı artmaktadır. Bu firmalardan Vestel, Venüs marka yerli cep telefonu ile kayda değer bir başarı elde etmiştir. Airties, yerel ağ ve internet üzerinden telefon ürünleri ile teknoloji ihracatında önemli bir paya sahiptir. Telekom sektöründe ise Netaş'ın ürettiği çözümler beş kıtada kabul görmektedir (Anonim, 2015).

Kadın milletvekili sayısı ülkemizde gelişmesi gereken alanlardan biridir. TÜİK'in 2014 yılında seçilmiş ülkeler için kadın milletvekili sayısını listelemiştir. Türkiye bu listede 45 ülke arasından %14.4 oranı ile 39. sıradadır. Listenin ilk üç sırasında ise İsveç, Finlandiye ve İzlanda bulunmaktadır.

Mal ve hizmetlerin ihracatı, dünyanın geri kalanına sağlanan malların ve diğer piyasa hizmetlerinin değerini temsil eder. Mal, nakliye, sigorta, nakliye, seyahat, gayrimaddi hak bedelleri, lisans ücretleri ve

iletişim, inşaat, finans, bilgi, iş, kişisel ve devlet hizmetleri gibi diğer hizmetlerin değerini içerirler. Türkiye ihracat alanında gelişme göstermektedir. TİM 'in 2015 yılı için hazırladığı sektör bazlı raporda, kimyevi maddeler ve mamulleri, otomotiv endüstrisi,hazırgiyim ve konfeksiyon ilk üç sırada yer almaktadır (Türkiye İhracatçılar Meclisi, 2016).

Geliştirilmiş bir su kaynağına erişim, gelişmiş bir içme suyu kaynağını kullanan nüfus yüzdesini ifade eder. İyileştirilmiş içme suyu kaynağında banyolarda su boruları ve diğer geliştirilmiş içme suyu kaynakları bulunmaktadır. Türkiye bu alanda gelişmiş ülkeler seviyesinde yer almaktadır.

Conflict of Interests/Çıkar Çatışması

Authors declare no conflict of interests/Yazarlar çıkar çatışması olmadığını belirtmişlerdir.

4. KAYNAKLAR

Akın, Y. 2008. Veri Madenciliğinde Kümeleme Algoritmaları Ve Kümeleme Analizi, Doktora Tezi, Marmara Üniversitesi, Sosyal Bilimler Esntitüsü.

Alptekin, N. Ve Yeşilaydın, G. 2015. "Oecd Ülkelerinin Sağlık Göstergelerine Göre Bulanık Kümeleme Analizi İle Sınıflandırılması", İşletme Araştırmaları Dergisi, Cilt 7/4, 137-155.

Anonim 2013. Bağlantı adresi: <http://www.finndiane.fi/wp-content/uploads/2013/01/help_plane.pdf>, Son Erişim Tarihi:: Mayıs 2017.

Anonim 2015. Teknoloji İhracatının Yıldızları Bağlantı adresi: <http://www.turkishtimedergi.com/ihracat/teknoloji-ihracatinin-yildizlari-2/>, Son Erişim Tarihi:: Mayıs 2017.

Çam, S. 2014. Veri Madenciliğinde Kümeleme Analizi Ve Sağlık Sektöründe Bir Uygulaması, Yüksek Lisans Tezi, Cumhuriyet Üniversitesi, Sosyal Bilimler Enstitüsü.

Çelik, G. 2013. Meslek Yüksekokulu Öğrencilerinin Başarı Durumlarını Etkileyen Faktörlerin Veri Madenciliği Kümeleme Teknikleri Kullanılarak Analizi: Ağrı Meslek Yüksekokulu Örneği, Yüksek Lisans Tezi, Atatürk Üniversitesi, Fen Bilimleri Enstitüsü.

Darakçı, H.Ç. 2011. Kümeleme Analizi Kullanılarak Benzin İstasyonlarının Operasyonel Değerlendirilmesi, Yüksek Lisans Tezi, Maltepe Üniversitesi, Fen Bilimleri Enstitüsü.

Değerli, O. 2012. Naive Bayes Yöntemi İle Blog İçeriklerinin Sınıflandırılması, Yüksek Lisans Tezi, Gazi Üniversitesi, Bilişim Estitüsü.

Duran, B.S. Ve Odell P.L. 1974. Cluster Analysis (Lecture Notes İn Economics And Mathematical Systems, Econometrics; Managing Editors: M. Beckmann And H.P. Kunzi). Springer-Verlag: New York.

Kalkınma Bakanlığı 2014. Onuncu Kalkınma Planı Bağlantı adresi: <http://www.kalkinma.gov.tr/Lists/Kalknma%20Planlar/Attachments/12/Onuncu%20Kalk%C4%B1nma%20Plan%C4%B1.pdf>, Son Erişim Tarihi:: Mayıs 2017.

Karakaya, M. Ö. 2012. *Clustering Based Diversity Improvement In Recommender Systems, Yüksek Lisans Tezi, Bahçeşehir Üniversitesi, Fen Bilimleri Enstitüsü.*

Kaya, O. 2008. Human Resource Performans Clustering Bu Using Self Regulating Clustering Method, Yüksek Lisans Tezi, Bahçeşehir Üniversitesi, Fen Bilimleri Enstitüsü.

Manning C. D. ve ark. 2008. Single-Link, Complete-Link & Average-Link Clustering. Bağlantı adresi: http://nlp.stanford.edu/IR-book/completelink.html, Son Erişim Tarihi: Şubat 2017

MOVSOM Research Lab, Bağlantı adresi: < http://rslab.movsom.com/paper/somrs/html/chapter4.php>, Son Erişim Tarihi:: Mayıs 2017.

Polzbauer, G. 2004. Survey And Comparison Of Quality Measures For Self-Organizing Maps, In Proc. 5th Workshop On Data Analysis (Wda'04), Pages 67–82.

Türkiye İhracatçılar Meclisi 2016. İhracat Rakamları Bağlantı adresi: <http://www.tim.org.tr/tr/ihracat-rakamlari.html>, Son Erişim Tarihi:: Mayıs 2017.

Vardar, T. 2010. Bankaların Tüzel Müşterileri Segmentasyonunun Niteliksel Ve Niceliksel Kümeleme Analizi, Yüksek Lisans Tezi, Marmara Üniversitesi, Sosyal Bilimler Esntitüsü.

Yıldız, K., Çamurcu, Y., Ve Doğan, B., 2010. Veri Madenciliğinde Temel Bileşenler Analizi Ve Negatifsiz Matris Çarpanlarına Ayırma Tekniklerinin Karşılaştırmalı Analizi, 10. Akademik Bilişim Konferansı Bildirileri.

Yılmaz, Ü. 2011. Türkiye'de İllerin Sosyoekonomik Gelişmişlik Düzeylerinin Faktör Analizi Ve Kümeleme Analizi İle İncelenmesi, Yüksek Lisans Tezi, Karadeniz Teknik Üniversitesi, Sosyal Bilimler Enstitüsü.