T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

ÖZDÜZENLEYİCİ HARİTALARIN GÖRSELLEŞTİRİLMESİ

EKREM ÖNCEL KORKMAZ

YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ PROGRAMI

DANIŞMAN
YRD.DOÇ. DR. SONGÜL ALBAYRAK

İSTANBUL, 2011

T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

ÖZDÜZENLEYİCİ HARİTALARIN GÖRSELLEŞTİRİLMESİ

Ekrem Öncel KORKMAZ tarafından hazırlanan tez çalışması 28.11.2011 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı'nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Tez Danışmanı

Yrd. Doç. Dr. Songül ALBAYRAK

Yıldız Teknik Üniversitesi

Jüri Üyeleri

Prof. Dr. Tülay YILDIRIM

Yıldız Teknik Üniversitesi _____

Doç. Dr. Banu DİRİ

Yıldız Teknik Üniversitesi _____

Yrd. Doç. Dr. Songül ALBAYRAK

Yıldız Teknik Üniversitesi _____

Bu çalışma, Yıldız Teknik Üniversitesi Bilimsel Araştırma Projeleri Koordinatörlüğü' nün 2011-04-01-YULAP01 numaralı projesi ile desteklenmiştir.

ÖNSÖZ

Bu çalışmada, özdüzenleyici haritaların görselleştirilmesi ve kümeleme metotları üzerinde durulmuştur. Özdüzenleyici haritalar metotlarının oluşturulan bir uygulama veri kümesi üzerindeki sonuçları incelenerek bunlardan anlamlı veriler üretmek amaçlanmıştır.

Özdüzenleyici haritalar, son zamanlarda yaygın olarak kullanılan eğiticisiz-öğrenmeye dayalı bir Yapay Sinir Ağı algoritmasıdır. Özdüzenleyici haritalar ile yüksek-boyutlu veride bulunan lineer olmayan istatiksel ilişkileri, düşük boyutlu (genellikle 2-boyutlu) örgüsel sisteme yansıtıp verinin analizi yapılabilmektedir. Bu çalışmada, başta Ekonomik Kalkınma ve İşbirliği Örgütü (OECD) üyesi ülkeler olmak üzere, bazı gelişmekte olan ülkelere ait eğitim, enerji, çevre, küreselleşme, işgücü, nüfus, fiyatlar, üretim ve tüketim, kamu maliyesi, yaşam kalitesi ve bilim ve teknoloji gibi çeşitli başlıklar altındaki parametreler kullanılarak bir veri kümesi oluşturulmuştur. Oluşturulan veri kümesine Özdüzenleyici haritaların metot ve yöntemleri uygulanılarak ülkeler arası benzerlikler ve farklılıklar ortaya konmaya çalışılmıştır.

Tez çalışmam sürecinde, bana destek olmaya çalışan, benden sabrını, bilgisini esirgemeyen değerli hocam Yrd. Doç.Dr. Songül Albayrak'a teşekkürü bir borç bilirim.

Ağustos, 2011

Ekrem Öncel KORKMAZ

İÇİNDEKİLER

Sayfa

SİMGE LİSTESİ

$\alpha(t)$	t anında öğrenme katsayısı
σ	Topolojik komşulukta efektif genişlik, komşuluk yarıçapı
κ	Sinaptik komşuluk derecesi
x_i	i'inci girdi vektörü
w_i	i'inci nöronun ağırlık vektörü
$d(x, y)$	SOM algoritmasında x ve y noktaları arasındaki uzaklık
$BK(r)$	r nöronuna ait birincil komşular
l	Örgü üzerindeki toplam nöron sayısı
A	Örgü üzerindeki nöronları kapsayan küme
$p(x, X)$	X veri uzayının x noktasındaki deneysel yoğunluk tahmini
m	Girdi uzayının boyutu
$d(x)$	x girdi vektörüne ait kazanan nöron
$h_{j,i}$	j ve i nöronları arasındaki komşuluk fonksiyonu
$d_{j,i}$	j ve i nöronları arasında öklid uzaklığı
$\eta(k)$	k anında öğrenme katsayısı
N	Girdi örneklerinin toplam sayısı
V_j	Voronoi kümesi
n_k	Voronoi kümesinin kütle merkezi
N_k	V_j'deki örnek sayısı
ζ_i	i nöronuna ait komşuların sayısı
M	Ağırlık vektörlerinin toplam sayısı
K_x	x ağırlık vektörüne haritalanan örneklerin kümesi
$d'_{j,i}$	Girdi uzayında j ve i birimleri arası uzaklık
s	SDH tekniğinde veri birimlerinin kümelere olan üyelik derecesi
c_j	K-ortalama'da kümeler
k	K-ortalama'da küme sayısı

KISALTMA LİSTESİ

AB	Avrupa Birliği
BMU	Best Matching Unit (Kazanan Nöron)
BRIC	Brazil, Russia, India and China (Brezilya, Rusya Federasyonu, Hindistan ve Çin)
CCA	Curvilinear Component Analysis (Eğrisel Bileşen Analizi)
GTM	Generative Topographic Mapping (Üretken Topografik Haritalama)
KGS	Kendini Güncelleme Süreci
KNN	K-Nearest Neighbor (K-En Yakın komşu)
MAR	Missing at Random (Rastgele Kayıp)
MCAR	Missing Completely at Random (Hepsi Rastgele Kayıp)
NI	Nonignorable (Göz Ardı Edilemez)
OECD	Organization for Economic and Co-Operation Development (Ekonomik Kalkınma ve İşbirliği Örgütü)
PCA	Principal Component Analysis (Temel Bileşen Analizi)
RF	Receptive Field (Algı Alanı)
SDH	Smoothed Data Histograms (İşlenmiş Veri Histogramları)
SOM	Self Organizing Map (Özdüzenleyici Haritalar)
SVD	Singular Value Decomposition (Tekil Değer Ayrıştırma)
VN	Vektör Nicemleme (Vector Quantization)
VY	Vektör Yansıtımı (Vector Projection)
YSA	Yapay Sinir Ağları

ŞEKİL LİSTESİ

Sayfa

ÇİZELGE LİSTESİ

Sayfa

ÖZET

ÖZDÜZENLEYİCİ HARİTALARIN GÖRSELLEŞTİRİLMESİ

Ekrem Öncel KORKMAZ

Bilgisayar Mühendisliği Anabilim Dalı

Yüksek Lisans Tezi

Tez Danışmanı: Yrd.Doç. Dr. Songül ALBAYRAK

Özdüzenleyici haritalar (SOM), eğiticisiz öğrenmeye dayalı bir Yapay Sinir Ağı algoritmasıdır. Özdüzenleyici haritalar ile yüksek-boyutlu veride bulunan lineer olmayan istatistiksel ilişkileri, düşük boyutlu (genellikle 2-boyutlu) örgüsel sisteme yansıtıp verinin analizi yapılabilmektedir. Bu tez kapsamında, başta OECD ülkeleri olmak üzere, yakın gelecekte güçlü ülkeler konumuna gelecek olan çeşitli ülkelere ait (iktisadi, sosyal, kültürel, enerji, eğitim, bilim ve teknoloji vb. çalışmalar sonucu elde edilen) nümerik veriler kullanılarak oluşturulan çok boyutlu bir veri kümesi ile çalışılmıştır. Oluşturulan çok boyutlu veri kümesine, SOM teknikleri ve yöntemleri uygulanılarak verinin iki ve üç-boyutlu örgüsel sistemler üzerinde analizinin gerçekleştirilmesi sağlanmaktadır. Bu görselleştirme ile mevcut ülkelerin (40 ülke) çeşitli açılardan gelişmişlikleri değerlendirilerek, bu ülkeler hakkında yorumlar yapmak mümkündür. Yine tez kapsamında, elde edilen veri kümesinin yanısıra bilinen veri havuzlarından alınan yapay ve gerçek veri kümeleri de incelenmiştir.

Özdüzenleyici haritaların görselleştirilmesi ve kümeleme çalışmaları bu tezin asıl ilgi alanıdır. Dolayısıyla bu çalışmada, görselleştirme teknikleri, çeşitli görselleştirme yöntemleri ele alınmış, bunların OECD veri kümesi ve seçilen diğer veri kümeleri üzerindeki uygulamaları verilmiştir.

SOM algoritmasının artı ve eksilerini ortaya koymak için ilgili alanlardaki farklı vektör nicemleme ve vektör yansıtım algoritmaları burada anlatılarak, SOM algoritmasının bu algoritmalar ile karşılaştırılması yapılmış ve sonuçları verilmiştir.

Özdüzenleyici haritaların kullanım amaçlarından bir diğeri de kümeleme metotları ile veri kümelerindeki kümelerin ortaya çıkartılmasıdır. Çeşitli veri kümeleri kullanılarak Özdüzenleyici haritaların kümeleme üzerindeki başarısı ortaya koyulmuştur.

Anahtar Kelimeler: Özdüzenleyici Haritalar, Yapay Sinir Ağları, Görselleştirme, Vektör Nicemleme, Vektör Yansıtımı, Veri Madenciliği, Kümeleme

ABSTRACT

VISUALIZATION OF SELF-ORGANIZING MAPS

Ekrem Öncel KORKMAZ

Department of Computer Engineering

M.Sc. Thesis

Advisor: Assist. Prof. Dr. Songül ALBAYRAK

A self-organizing map (abbreviated "SOM", also known as "Kohonen map") is a type of artificial neural algorithm and based on unsupervised learning. Self-organizing map is trained to produce a low-dimensional (typically two-dimensional) lattice which is discretizated representation of the input space of the training samples. It is generally difficult to analyze high-dimensional data and make a command about that, so that visualizing low-dimensional views of high-dimensional data is very useful.

The initial idea behind this thesis is applying the SOM algorithm to a dataset which is composed of various specifications of OECD (Organisation for economic co-operation and development) countries. We designed a data set which is made up of 40 countries and 35 indicators, which involve financial, social, culturel, educational, scientific and technological information about countries. This thesis aims to discuss what can be learned from the created data set with the help of self-organizing map. At the end of this thesis, people can assess the levels of the development of the existing countries. Besides created data set, we choose different data sets from various data pools and analyse them.

The actual scope of this thesis is visualization of the self-organizing map, so that here we present several visualization techniques and methods, and the applications of these methods using created OECD data set are presented.

Also, we mentioned other vector quantization and vector projection methods to compare with self-organizing map, and exposed their advantages and disadvantages. We seek for solutions to missing values and present examined ways for this problem.

Self-organizing map is also considerable in the name of clustering. Self-organizing map can be used to reveal cluster structures in the data sets and it is very efficient in this issue. And we used different and various data sets to show the success of SOM in clustering.

Key words: Self-Organizing Maps, Artificial Neural Networks, Visualization, Vector Quantization, Vector Projection, Data Mining, Clustering

BÖLÜM 1

GİRİŞ

1.1 Literatür Özeti

Veri madenciliği, elde edilmiş olan verileri işleyip analiz yoluna giderek, veri içerisinde saklı olan önemli bilgileri ortaya koymayı ve anlamlı sonuçlar üretmeyi amaçlayan bir araştırma dalıdır. Eldeki veriyi araştırma, anlama ve bu veriden sonuç üretme tüm bilimsel dallar için son derece önemlidir. Bilgisayar teknolojisindeki gelişmeler ile birlikte, çeşitli kaynaklardan elde edilen ve saklanan veri miktarında önemli artışlar meydana gelmiştir. Dolayısıyla, veri madenciliği çatısı altında, daha etkin çalışabilen ve veri analizi yapabilen metotlar geliştirilmiştir. Bu metotlardan bir tanesi de özdüzenleyici haritalardır [1].

Özdüzenleyici haritalar, eğiticisiz-öğrenmeye dayalı yapay sinir ağı algoritmasıdır. İlk olarak 1981 yılında Teuvo Kohonen tarafından ortaya atılmış olan SOM haritaları günümüze kadar yoğun olarak kullanılmış ve veri görselleştirmesi adına bu konuda birçok çalışma yapılmıştır.

Özdüzenleyici haritalar, yüksek-boyutlu veride bulunan linear olmayan istatistiksel ilişkileri, düşük-boyutlu (genellikle 2-boyutlu ve dikdörtgenimsi) örgüsel sisteme yansıtır. Daha yüksek boyutlu (3-boyutlu ya da daha fazla) örgüsel sistemler de kullanılabilmekte fakat bunların görselleştirilmesi problemli olduğundan dolayı çoğu zaman tercih edilmemişlerdir. Yüksek-boyuttaki topolojik ilişkilerin düşük-boyuta geçişte korunması son derece önemlidir.

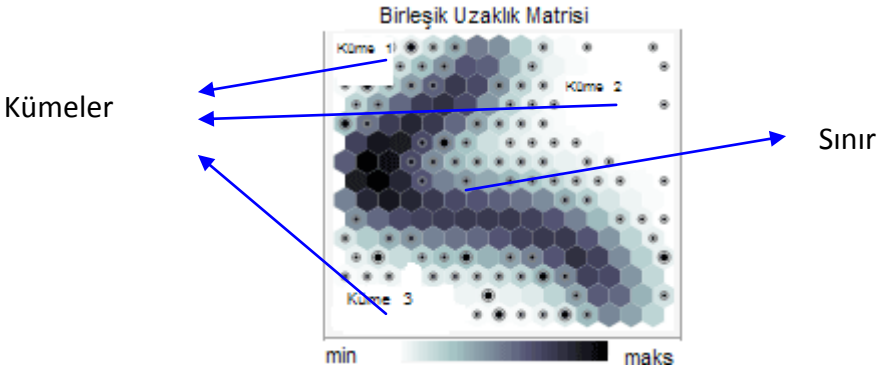
Çeşitli kaynaklardan toplanan ve nümerik formattaki veri yığınlarından oluşan girdi uzayları, veri madenciliği çerçevesinde günümüzde yoğun olarak kullanılmakta ve bu kaynaklardan anlamlı sonuçlar çıkartılmaya çalışılmaktadır. Toplanan veriler tıp [7], biyoloji, telekomünikasyon[20] ve ekonomi gibi çeşitli dallarda gerçekleştirilen bilimsel araştırmalar sonucu elde edilmiş bilgiler olabilir. Geniş bir çalışma sahası bulunması nedeniyle elde edilen veriler de bu yönde farklılık ve çeşitlilik göstermiş, dolayısıyla bu konuda yapılan çalışmaların sayısı dikkat çekici boyutlara ulaşmıştır.

Bilindiği üzere girdi uzayındaki mevcut topolojinin çıktı haritasında da korunması, özdüzenleyici haritalar temel prensiplerinden birisidir. Taşdemir ve Merenyi [9], [10], topolojiler ve özdüzenleyici haritalarda meydana gelen topolojik bozulmaların ortaya çıkartılması konusunda çalışmalar sunmuşlardır. Bu çalışmalarda veri topolojisi örgü üzerine yerleştirilerek, bu işlemin veri yapısının ortaya çıkartılmasına nasıl bir katkı sağladığı gösterilmiştir. Eğitimli bir SOM'da ağırlık vektörleri (w_i), veri örneklerinin temsilcisi haline gelirler (bu vektörler için farklı adlandırmalar mevcuttur; model vektör, codebook vektör, referans vektörü, ağırlık vektörü). Burada i, ağırlık vektörünün özdüzenleyici harita üzerindeki konumunu belirtmektedir ve her bir w_i'ye ait bir algı alanı ("receptive field", RF) mevcuttur. Kazanan nöronun (BMU) belirlenmesinden sonra ikinci BMU (komşu nöron)'lar seçilmelidir. İkinci BMU'lar da önemlidir çünkü özdüzenleyici haritalar üzerindeki nöronlar kazanan nöron ve komşuları şeklinde uyarlanmaktadırlar. Netice olarak, BMU ve ikinci BMU'ların arasında teşekkül eden ilişkiler, algı alanında bir tertip oluşturur. Algı alanındaki bu düzenin gösterilmesi için bir "toplam komşuluk matrisi" oluşturulmuştur [9]. Sonuçta da belirlenmiş olan eşik değerlerine göre, özdüzenleyici harita üzerinde, w_i ve komşuları arası ilişkiler, belli renk ve kalınlıktaki çizgiler ile gösterilmiş, topolojik bozulmaların görsel olarak elde edilmesi sağlanmıştır. Aynı zamanda, özdüzenleyici haritaların kümeleme ve görselleştirme çalışmalarında veri topolojisinin etkisi ortaya koyulmuştur [10]. Özellikle karmaşık verilerin kullanıldığı çalışmalarda bu yöntem çok efektif sonuçlar ürettiği için veri kümesindeki topolojik bozulmalar açık bir şekilde incelenebilecektir.

Özdüzenleyici haritaların görselleştirilmesinde kullanılan yaygın yöntemlerden birisi, uzaklık matrislerinin bir versiyonu olan U-Matris (Unified distance matrix, Birleşik uzaklık matrisi)'tir. U-matris, bir nöronun diğer nöronlara olan ortalama uzaklıklarını göstermektedir. Neticede Şekil 1.1'e benzer bir grafik ortaya çıkar. U-matris, SOM ile kaç adet küme yapısının ortaya çıkartıldığını ortaya koyan kullanışlı bir anahtardır denilebilir. Silva [11], jeo-uzamsal veriler kullanıp, eldeki verilere ait U-matris ve bileşen düzlemlerini ortaya koyarak verilerin analizini gerçekleştirmiştir. Yine son zamanlarda yoğun olarak kullanımına başvurulan bileşen düzlemleri ise, veri kümesindeki her bir özelliğin (değişken) ayrı ayrı renk haritalarını çıkartılmasıdır. Bu şekilde veri kümesindeki özellikler arası ilişkiler ortaya çıkartılabilmekte ve her bir değişkenin özdüzenleyici haritalara katkısı görülebilmektedir. Benzer bir çalışma da, Pölzlbauer [6] tarafından gerçekleştirilmiş, fakat bu kez benzer nitelikteki değişkenler gruplama yoluna gidilmiştir. Değişkenler arası ilişkiler "Pearson Correlation", "Kendall Ranking" veya "Gradient Field" yöntemleri ile ölçülebilmektedir. [6], ise gradient field metotu tercih etmiştir. Araştırmalarda kullanılan bazı veri kümelerinde, değişkenlerin semantik olarak gruplandığı ve benzer nitelikteki değişkenlerin aynı değişken grubu içerisinde yer aldığı görülmektedir.

Özdüzenleyici haritaların kullanım amaçlarından bir diğeri de kümeleme metotları ile veri kümelerindeki kümeleri ortaya çıkartmaktır. Günümüze kadar, harita yapılarındaki kümelerin elde edilmesi için çeşitli tekniklerin uygulandığı görülür. Bunlardan birisi de uzaklık matrisleridir. Uzaklık matrisleri veriye ait yerel ağırlıkları tuttuğu için özdüzenleyici haritalarda küme yapısının ortaya koyulmasında uzaklık matrisleri günümüze değin kullanılagelmiştir. Bunun için [12], uzaklık tabanlı kümeleme için bir algoritma tasarlamıştır. Bu algoritmaya göre, başlangıçta uzaklık matrisindeki en küçük yerel ağırlık değerine sahip harita birimleri küme merkezi olarak seçilmekte, daha sonra geri kalan harita birimleri, hangi küme merkezi kendisine daha yakın ise o kümeye atanmaktadır. Uzaklık matrisindeki yüksek değerler, kümeler arasında bulunan sınırları gösterirken, sınırların arasında kalan bölgeler de kümeleri belirtirler. Bir başka ifade ile koyu renkler birimler arası mesafenin yüksek olduğu bölgeler iken, açık renkler birbirine yakın birimlerin bir arada olduğu bölgeleri gösterir (Şekil 1.1). Genellikle oluşan kümeler "vadi" olarak anılır ve bu vadinin en derin yeri (en açık renkli nokta) o

kümenin merkez noktasını temsil eder. Şekil 1.1'de üç beyaz bölge göze çapmakta, bu da bize özdüzenleyici haritalar tarafından üç adet kümenin ortaya çıkartıldığını gösterir. Renkli U-matris grafiklerinde ise kırmızı renk tonu yüksek değerleri temsil ederken, mavi renk tonu ise düşük değerleri temsil eder.



Şekil 1. 1 Birleşik uzaklık matrisi (U-matris)

U-matrislerin nasıl oluşturulduğu detaylı olarak 4.bölümde verilmiştir. Yüksek U-yükseklik değerine sahip nöronlar girdi uzayında diğer vektörlerden daha uzakta olurken, daha küçük U-yükseklik değerine sahip birimler ise girdi uzayında diğer nöronlar ile yakın mesafelerde yer almaktadırlar. Dolayısıyla, girdi uzayında diğer nöronlardan daha uzakta bulunan vektörlerin "gürültü" olması olasıdır. Bu birimler, genellikle U-matrisi üzerindeki "sınır" olarak nitelenen bölgelerde yer alır.

U-matris kullanımları verinin kümelenmesi için fevkalade yararlı olmaktadır. Fakat, SOM/U-matris kullanımı, bazı kümeleme çalışmalarında olumsuz sonuçlar doğurabilmektedir [3]. Bunun nedeni de, küme-içi uzaklık kavramının kümeler-arası uzaklık kavramı ile aynı şekilde değerlendirilmesidir. Ultsch [3], seçtiği bir veri kümesine hem hiyerarşik kümeleme algoritmasını hem de U-matris yöntemini sınamış, sonuç olarak da hiyerarşik kümeleme algoritması mevcut kümeleri ortaya çıkarırken, U-matris yönteminin kümeleri ortaya çıkarmada başarısız olduğu gözlenmiştir. Bunun üzerine, U-matris görselleştirmelerini iyileştirmek amacıyla veri kümesindeki veri yoğunluğunu dikkate alan P-matris yöntemi [4], U-matris ile birleştirilerek yeni bir metot olan U*-matris yöntemi ortaya koyulmuştur. U*-matris yöntemi hiyerarşik kümeleme algoritması ile bağdaşan bir metottur ve veri kümesi içindeki küme yapıları bu yöntem ile elde edilebilir.

Veri uzayındaki yapısal özelliklerin ortaya koyulması için geliştirilmiş görselleştirmelerden bir tanesi de P-matristir. P-matris, U-matris ile benzer bir çalışma prensibine sahiptir. Fakat bu kez yükseklik, U-matriste kullanılan yerel uzaklıklar yerine veri uzayındaki veri yoğunluğuna bağlı olarak hesaplanmaktadır [3]. P-matrisin nasıl oluşturulduğu yine 4. bölümde verilmiştir.

Genel olarak P-matrise ait özellikler U-matrisin özelliklerine terstir. U-matrisler veriler arası öklid uzaklığına dayanırken, P-matrisler ise veriler arası yoğunluğa bağlı olarak çalışır.

Kayıp Değer Sorunu

Özdüzenleyici harita uygulamalarında sıkça karşılaşılan problemlerden birisi "kayıp değerler" dir. Kayıp değer sorunu, veri kümesinde olması gereken özelliklerden veya kayıtlardan bir ya da bir kaçının bulunmaması demektir. Örneğin, veri toplama esnasında insanların bazı düşüncelerini ortaya koymamasının sonucu olarak veri eksik kalabilir. Kaliteli bir analiz ve sonuçların doğrulanması açısından kayıp değerlerin dikkatli bir çalışma ile ele alınması gerekir. Özellikle sınıflandırma alanındaki çalışmalarda tartışma konusu olmuştur. Çözüm olarak, eksik veri olan satırı silme, eksik veri(ler) yerine ortalamayı koyma, geçerli sabit bir değer ile doldurma, kayıp değer yerine sınıfa ait tüm değerlerin ortalamasının verilmesi gibi çok çeşitli yöntemlere başvurulur. Tabiki burada hangi tekniğin uygulanacağı büyük ölçüde kayıp değer mekanizmasına bağlıdır. Little ve Rubin [14], kayıp değerleri üç kategoride incelemektedir; göz ardı edilemez (Nonignorable), rastgele kayıp (Missing at random) ve hepsi rastgele kayıp (Missing completely at random).

1. *Göz ardı edilemez (NI):* Bu gruba örnek olarak insanların düşüncelerini ifade etmek istemedikleri durumlar örnek gösterilebilir. Örneğin, araştırmada bazı insanlar yüksek bir ekonomik gelire sahip olduğunu saklayabilir ve bunu ifade etmekten kaçınır. Bu durumda bu değişkenin göz ardı edilmesi söz konusu olmamalıdır.
2. *Rastgele kayıp (MAR):* Kayıp değerler örnek uzay boyunca rastgele olarak dağılım göstermektedir. Dolayısıyla değerlerin kaybolmasında, herhangi bir tahminde bulunmak güç olacaktır. Kayıp değerler arasında tam anlamıyla bir

ilişki kurmak da zordur. Bu noktada herhangi bir atıf analizinde bulunmak da geçersiz olur.

3. *Hepsi rastgele kayıp (MCAR):* İki değişken ele alındığında durum bu iki değişkenden bağımsız ise kayıp değer MCAR olarak düşünülür. Örneğin, verinin olmaması araştırmacıdan kaynaklanabilir ya da araştırmaya katılanlar dikkatsizlikle bazı durumları atlayabilir. Genellikle çoğu kayıp değer bu grubun dışındadır.

[15], özdüzenleyici harita temelli veri sınıflandırmasındaki kayıp değerlerin tespit edilmesi ve görselleştirmeleri için bir model önermektedir. İstatistiksel kümeleme metotlarının uygulanması zor olduğunda, özdüzenleyici haritaların görselleştirilmiş kümeleme analizinde ne kadar yararlı olduğu görülmektedir.

1.2 Tezin Amacı

Bu çalışmanın amacı; herkese açık olan OECD ülkelerine ait verilerden, belli kriterler ve değerlendirmeler neticesinde oluşturulan veri kümesi üzerinde SOM algoritmasını uygulamak ve SOM algoritmasının yardımıyla veri kümesine ait yapısal nitelikleri ortaya koyabilmektir. Ayrıca, çeşitli görselleştirme teknikleri ve yöntemleri ile veri kümesi ile ilgili görsel sonuçlar türetip analiz, değerlendirme ve yorumlamalarda bulunabilmektir.

1.3 Hipotez

Bugüne değin, bir çok araştırma ve çalışma bu konu çerçevesinde yapılmıştır. Bu tez kapsamında ise özdüzenleyici haritaların, gerçek-dünyaya ait veriler kullanılarak yapılacak olan bir çalışmadaki başarısı sergilenecektir. Yani, çalışmanın sonunda, veri kümesi gerçek-dünya gözlemleri ile sorgulanabilecektir (Örneğin, ülkelerin özdüzenleyici harita üzerindeki dağılımı). Aynı zamanda, herhangi bir kriterin, ülkelerin gelişmişlik seviyesini ne şekilde etkilediği gözlemlenebilecektir. (Örneğin, “Eğitim seviyesi” veya “nüfus ve göç miktarları” nın ayrı ayrı etkilerinin harita üzerinde görülmesi.)

Bu tezin asıl kapsamı; özdüzenleyici haritaların yardımıyla OECD veri kümesinden nelerin öğrenilebileceğini araştırmak ve tartışmaktır. Bilhassa, özdüzenleyici haritalara

ait güncel gelişmeler ve uzantılar vurgulanacaktır. Özdüzenleyici haritalar yapısı gereği herhangi bir görselleştirme tekniğine sahip değildir. Fakat haritamsı yapısından dolayı, anlamlı ve sezgisel yollar kullanılarak görselleştirilebilir. Yani önceden eğitilmiş olan haritanın görselleştirilmesi mümkün olabilmektedir. Bunu gerçekleştiren bazı yöntemler burada tartışılacaktır.

Özdüzenleyici haritalar çok amaçlı bir yapıya sahiptir. Vectör nicemleme, vektör yansıtımı, yapay sinir ağları ve eğiticisiz öğrenme kavramları ile bir yakınlık içindedir. Dolayısıyla özdüzenleyici haritalar bu kavramların her biri ile kıyaslanabilir. Bu çalışmada bu konuya da değinilecektir.

Proje kapsamında gerçekleştirilmesi düşünülen bütün hesaplamalar ve uygulamalar, bir yazılım aracı ve bilimsel hesaplamalar için bir bilgisayar dili olan Matlab ile yapılacaktır. Aynı zamanda, şekillerin ve grafiklerin oluşturulması için Matlab geniş imkanlar sağlamaktadır. Tezin büyük çoğunluğunda, J.Vesanto, J.Himberg, E.Alhoniemi, J. Parhankangas ve daha bir çok kişi tarafından gerçekleştirilen SOM toolbox [16] kullanılmaktadır. SOM toolbox, güçlü bir uygulama ve görselleştirme paketidir denilebilir.

Çalışma sırasında, mevcut değişkenler ve ölçülmesi gereken büyüklükler çok fazla miktarda (100'den fazla sayıda değişken) olduğu için, irdelenmesi gereken konular ve elde edilmesi öngörülen veriler çerçevesinde değişkenler indirgenme yoluna gidilerek daha az sayıda değişken (35 adet) ile çalışılacaktır.

Temel olarak özdüzenleyici haritaların eğitiminde iki farklı türde eğitim algoritması kullanılır. Bunlar, Sıralı öğrenen algoritması ve Batch yöntemi'dir. Genellikle SOM eğitim algoritması denilince akla, *sıralı öğrenen algoritması* gelir. Fakat, bu çalışmadaki uygulamalarda sıralı algoritma yerine Matlab'da daha efektif çalışan *batch eğitim algoritması* tercih edilmiştir.

Özdüzenleyici haritaların görselleştirmeleri, Vesanto [17] tarafından iki grupta incelenmiştir; harita vektörlerinin görselleştirilmesi ve örgü görselleştirilmesi. Burada da görselleştirme çalışmaları bu iki ana başlık üzerinden yürütülmektedir. Bunun

yanısıra kümelerin görselleştirilmesi ele alınacaktır. Daha önce yapılmış olan çalışmalara ek olarak yeni yaklaşımlar ve teknikler üzerinde durulacaktır.

Gerçekleştirilecek olan tez sonucunda, Türkiye'nin de içinde bulunduğu OECD ülkelerinin gelişmişlik düzeyleri, ülkelerin diğer ülkeler arasında bulunduğu konum, ekonomik, kültürel, teknolojik ve bilimsel vb. açıdan nasıl bir seviyede oldukları bu çalışma sonucunda elde edilecektir. Bunun yanında, farklı parametrelerin tekil olarak ya da grupsal olarak etkileri gözlemlenebilecektir.

Tez çalışmasının birinci bölümünde tez konusu, yapılacak olan çalışma, çalışmanın amaç ve hedefleri ve bu alanda daha önce yapılmış olan çalışmalardan bahsedilmektedir. İkinci bölümde, veri kümesi detaylı bir şekilde ele alınmıştır. Veri kümesinde kullanılan ülkeler ve değişkenler bu bölümde açıklanmıştır. Üçüncü bölümde, yapılan çalışmanın daha anlaşılır olması için SOM metodu ve konuyla ilgili metotlar açıklanmaktadır. SOM algoritması, SOM'un çalışma mantığı, SOM eğitim algoritmalarına yine bu bölümde değinilmektedir. Dördüncü bölümde ise eğitim sürecinin sonrası olan görselleştirmeler üzerinde durulmaktadır. Görselleştirme teknikleri örnek veri kümeleri üzerinde gösterilmiştir. Son olarak beşinci bölümde, tez kapsamında oluşturulan veri kümesi üzerinde SOM teknikleri ve görselleştirme metotları uygulanmıştır. Çalışmanın sonuçları ve elde edilen bilgiler sunulmaktadır.

BÖLÜM 2

OECD VERİ KÜMESİ

Veri kümesi, OECD (Organization for Economic and Co-Operation Development, Ekonomik Kalkınma ve İşbirliği Örgütü) ülkeleri ve bazı gelişmekte olan ülkelere ait eğitim, enerji, küreselleşme, iş gücü, nüfus, fiyatlar, üretim ve gelirler, kamusal maliye, yaşam kalitesi ve bilim ve teknoloji alanlarındaki veriler ele alınarak oluşturulmuştur. Toplanan veriler, OECD'nin 2010 yılında yayımlamış olduğu verilerden elde edilmiş olup, bu veriler genel olarak 2008 yılına aittir. Bazı değişkenlerde ise sürecin daha önemli olduğu düşünülerek verilerin son beş yıldaki değişimi ele alınmıştır.

2.1 Ülkeler

Veri kümesi, 40 ülkeden (Ek-C) oluşmaktadır. Bu ülkeler, OECD (Ekonomik Kalkınma ve İşbirliği Örgütü) ile gelişmekte olan ve yakın gelecekte gelişmiş ülkelerin arasına katılması beklenen ülkelerdir. Veri kümesi, OECD ülkeleri başta olmak üzere, Avrupa Birliği (2007 yılı, beşinci katılımdan önceki 21 üyesi), G-8 ülkeleri, Akdeniz ülkeleri ve 2050 yılında dünyanın en güçlü ülkeleri olmaları beklenen BRIC (Brezilya, Rusya, Hindistan ve Çin), ve yine 2050 için N-11 (en gelişmiş 11 ülke) grubuna girmesi beklenen Türkiye, Meksika, Güney Afrika, Endonezya'yı değerlendirmemiz açısından önemlidir.

İlerleyen bölümlerde, gerçek veri kümesine SOM algoritmasının uygulanmasının ardından, yukarıda belirtilen kriterler açısından sonuçların incelenmesi ilginç olacaktır.

2.2 Değişkenler

Ülkelerin karşılaştırılmasında geniş bir yelpazeden (ekonomi, iş gücü, küreselleşme süreci, enerji, nüfus, fiyatlar, üretim, kamu maliyesi, yaşam kalitesi ve bilim ve teknoloji) seçilen 35 değişken kullanılmaktadır (Ek-C). Bu değişkenler toplam 10 kategoride toplanmıştır ve kategorilerine göre şu şekildedir:

Eğitim (3 değişken)

1. *Eğitimsel Kazanım:* Yeterince eğitilmiş ve iyi yetiştirilmiş bir toplum ülkenin sosyal ve ekonomik refahı açısından önemlidir. Aynı şekilde, bu altyapı bilimsel ve kültürel birikimin gelişmesine katkı sağlar. Burada, yüksek öğrenim görmüş ya da görmekte olan nüfusun toplam nüfustaki yüzdelik payı (25-64 yaş grubu için) ele alınmıştır.
2. *Öğrenci Harcamaları:* Siyasete yöne verenlerin eğitimde kaliteyi artırmaları, herkesin eğitim olanaklarından yararlanmasını sağlamaları gerekmektedir. Burada, yüksek öğrenim öğrencileri için devletin öğrenci başına yıllık ne kadarlık bir yatırım/harcama yaptığı ele alınmıştır.
3. *Yüksek Öğrenim Kayıt Oranları:* Yükseköğrenime yüksek oranda katılım olması o ülkenin gelişmekte olan/gelişmiş bir ülke olduğunun göstergesidir. Orta öğretimden mezun olup yüksek öğrenime kayıt yaptıran öğrencilerin oranları kullanılmıştır.

Enerji (3 değişken)

1. *Elektrik Üretimi:* Bir ülke tarafından üretilen elektrik miktarı, o ülkenin elektriklileşme oranını, nüfusun boyutunu ve ülkenin ekonomik gelişmesini kestirmemizde yardımcı olur. Ülkelerin yılda ürettikleri elektrik miktarı terawatt cinsinden ele alınmıştır.
2. *Enerji Üretimi:* Enerji üretimi de elektrik üretimine benzer şekilde ülkenin doğal kaynaklarını ne derece iyi kullandığının ve ekonomik teşviklerinin bir göstergesidir.
3. *Yenilenebilir Enerji:* Çoğu hükümet enerji politakalarını kurgularken sürdürülebilir kalkınma ve iklim değişiklikleri ile mücadeleyi de göz

önünde bulundurur. Yüksek enerji kullanımının sonucunda yüksek derecede atmosfere gaz salımı neticelendirmektedir. Bu noktada yenilenebilir enerjinin kullanımı önemlidir. Bunun yüzden ülkeler için yenilenebilir enerjinin toplam enerji tedarikindeki payı değerlendirilmiştir.

Küreselleşme Süreci (2 değişken)

1. *Ödeme Dengesi:* Ödeme dengesi, yurtdışından gelen ödemeler ile yurtdışına yapılan ödemelerin farkıdır. Bu farkın ülkeler için GSYİH'nın yüzde kaçına tekabül ettiği ele alınmıştır.
2. *Uluslararası Ticaret:* Uluslararası yapılan ticaret ekonomik büyümeyi ve yaşam standartlarının yükselişini beraberinde getirir. Tabii ki burada ticaret dengesinin (ihracat-ithalat farkının) pozitif bir değer olması önemlidir. Ülkelerin uluslararası ticaret dengeleri burada kullanılmıştır.

İşgücü (3 değişken)

1. *İstihdam Oranı:* Mevcut işgücünün en iyi şekilde değerlendirilmesine istihdam oranı denilebilir. Kısa süreli iniş ve çıkışlar ekonomik sirkülasyonun bir neticesi olarak görülebilir ancak uzun dönem istihdam oranları incelendiğinde ülkenin başarısı ya da başarısızlığı bu konuda değerlendirilebilir. Bunun için çalışma yaş grubunda yer alıp çalışan insanların oranları burada ele alınmıştır.
2. *Uzun Dönem İşsizlik:* Uzun dönem işsizlik oranlarının yüksek olması iş piyasasının efektif bir şekilde yürümediğinin göstergesidir. İşsiz kesim içinde bulunup 12 aydan daha uzun süre işsiz kalanların oranları değerlendirilmiştir.
3. *Serbest Meslek Oranları:* İşsizlik oranları değerlendirilirken göz önünde tutulması gereken bir faktördür. Toplumun bir kısmı kendi işini kurup bunun üzerinden gelir elde etmektedir.

Nüfus (1 değişken)

1. *Bağımlı Nüfus:* Bağımlı nüfus (Pasif konumdaki genç ve yaşlı nüfus), devletin üzerine ek bir sorumluluk yüklemektedir (Emeklilik maaşı

ödemesi, sağlık hizmetlerinin sunulması, eğitim vb.). Son yıllarda, özellikle gelişmiş ülkelerde bağımlı nüfusun artış gösterdiği gözlemlenmektedir. Burada, 65 yaş üzerinde olup bağımlı nüfus konumunda olan insanların aynı yaş grubu içindeki oranı ele alınmıştır.

Fiyatlar (3 değişken)

1. *Tüketici Fiyat Endeksi:* Tipik bir tüketicinin satın aldığı belli bir ürün ve hizmet grubunun fiyatlarındaki ortalama değişimleri gösteren ölçüttür. Yıllık enflasyon değerindeki değişimi ölçmek için kullanılmaktadır.
2. *Uzun Dönem Faiz Oranları:* Faiz oranları özellikle iş yatırımlarını etkilemektedir. Düşük faiz oranları yatırımcılar için bir teşvik niteliğine sahip iken faizlerin yükselmesi ise yatırımcıları korkutmaktadır.
3. *Üretici Fiyat Endeksi:* ÜFE, belirli bir referans döneminde ülke ekonomisinde üretimi yapılan ve yurtiçine satışa konu olan ürünlerin, üretici fiyatlarını zaman içinde karşılaştırarak fiyat değişikliklerini ölçen fiyat endeksidir.

Üretim ve Gelirler (4 değişken)

1. *Gayrisafi Yurtiçi Hâsıla Gelişimi:* Gayrisafi yurtiçi hâsılanın miktarından daha önemlisi yıllık büyüme/küçülme oranıdır. Burada (2004-2008) yılları için yıllık büyüme miktarı değerlendirilmiştir.
2. *Yatırım Oranları:* GSYİH'nın ne kadar bir kısmının yine GSYİH'nın gelişmesi ve büyümesi için yatırım olarak kullanıldığı burada önem arz etmektedir.
3. *Kişi Başına Düşen Millî Gelir:* Kişi başına milli gelir toplumlararası yaşam standartlarının karşılaştırılmasında birçok analist tarafından kullanılmaktadır.
4. *Gayrisafi Yurtiçi Hâsıla Miktarı:* GSYİH bir ülkenin belli bir zaman diliminde üretmiş olduğu mal ve servislerin toplam miktarını belirlemede standart ölçümdür.

Kamusal Finans (5 değişken)

1. *Eğitim Harcamaları:* Eğitim harcamaları bireysel ve sosyal gelişimlere katkı sağlamakta ve sosyal eşitsizliği düşürmektedir. GSYİH'nın yüzdesel olarak ne kadarının eğitime harcandığı ele alınmıştır.
2. *İç Borçlar:* Hükümet borcu kamu maliyesinin sürdürülebilirliği açısından önemlidir. Hükümet borcunun GSYİH'nın yüzde kaçına tekabül ettiği ele alınmıştır.
3. *Sağlık Harcamaları:* Yapılan sağlık harcamalarının GSYİH'nın ne kadarına tekabül ettiği ele alınmıştır.
4. *Maaş Ödemeleri:* Düzenli bir geliri olmayanlara verilen maddi yardımlardır. GSYİH'nın ne kadarının bu alana ayrıldığı ele alınmıştır.
5. *Sosyal Harcamalar:* Korunmasız ve ezilen gruplar için yapılan harcamaların GSYİH'nın yüzdesel tabanda ne kadarına tekabül ettiği ele alınmıştır.

Yaşam Kalitesi (5 değişken)

1. *Bebek Ölüm Oranı:* Bebek ölüm oranı henüz erken yaşlarında bulunan bireylerdeki önemli bir sağlık sorunudur. Her bin doğumda ölü olarak doğan bebeklerin sayısı kullanılmıştır.
2. *Ortalama Yaşam Süresi:* Yükselen yaşam standartları, ilerleyen yaşam tarzı, iyi eğitim, yeterli beslenmenin artışı, sanitasyon vb. faktörler ortalama yaşam süresini artıran faktörlerdir. Ülkelerdeki ortalama yaşam süresi ele alınmıştır.
3. *Hapishane Nüfusu:* Hapishane nüfusu bir ülkedeki suç oranlarını yorumlamamızı sağlar. Yerleşik halkın yüzde kaçının hapiste olduğu değerlendirilmiştir.
4. *Trafik Kazaları:* Trafik kazalarının sayısı o ülkenin ulaşım kalitesi hakkında bize bilgi verir. Her bir milyon kişiden kaçının yol kazalarına bulaştığı ele alınmıştır.
5. *Gençlerin Etkisizliği:* Çalışmayan, eğitim olanaklarından faydalanmayan ya da herhangi bir staj vb. etkinlikte bulunmayan bireyler sosyal olarak

adeta dışlanmış konumdadır. Ayrıca bu kişiler kuvvetle muhtemel fakirlik sınırının altında kalmaktadır. Bu bireylerin çokluğu devletin elindeki insan gücünü iyi değerlendiremediği ve yönetemediği anlamına gelir. 20-24 yaş grubunda olup bu yaş grubundaki etkisiz gençlerin yüzdesi ele alınmıştır.

Bilim ve Teknoloji (6 değişken)

1. *Bilgisayar Erişimi:* En az bir bilgisayara sahip hanelerin yüzdesi kullanılmıştır.
2. *İnternet Erişimi:* İnternete erişebilen hanelerin toplum içindeki yüzdesi ele alınmıştır.
3. *Haberleşme:* Kişi başına düşen haberleşme erişim yolu sayısı ele alınmıştır.
4. *Araştırma ve Geliştirmeye Yapılan Harcamalar:* Araştırma ve geliştirme çalışmaları için yapılan toplam harcama, o ülkedeki GSYİH'nın yüzdesel karşılığına göre değerlendirilmiştir.
5. *Bilgi ve İletişim Teknolojileri Ürünlerinin İhracatı:* Son on yıla bakıldığında uluslararası ticarette bilgi ve iletişim teknolojileri ürünlerinin ön plana çıktığı görülüyor. Ülkelerin bu alanda yaptığı ihracat tutarı ele alınmıştır.
6. *Araştırmacılar:* Araştırmacılar, araştırma ve geliştirme sisteminin merkez unsurudur. Çalışanların yüzde kaçı araştırmacı olarak çalışmakta olduğu ele alınmıştır.

BÖLÜM 3

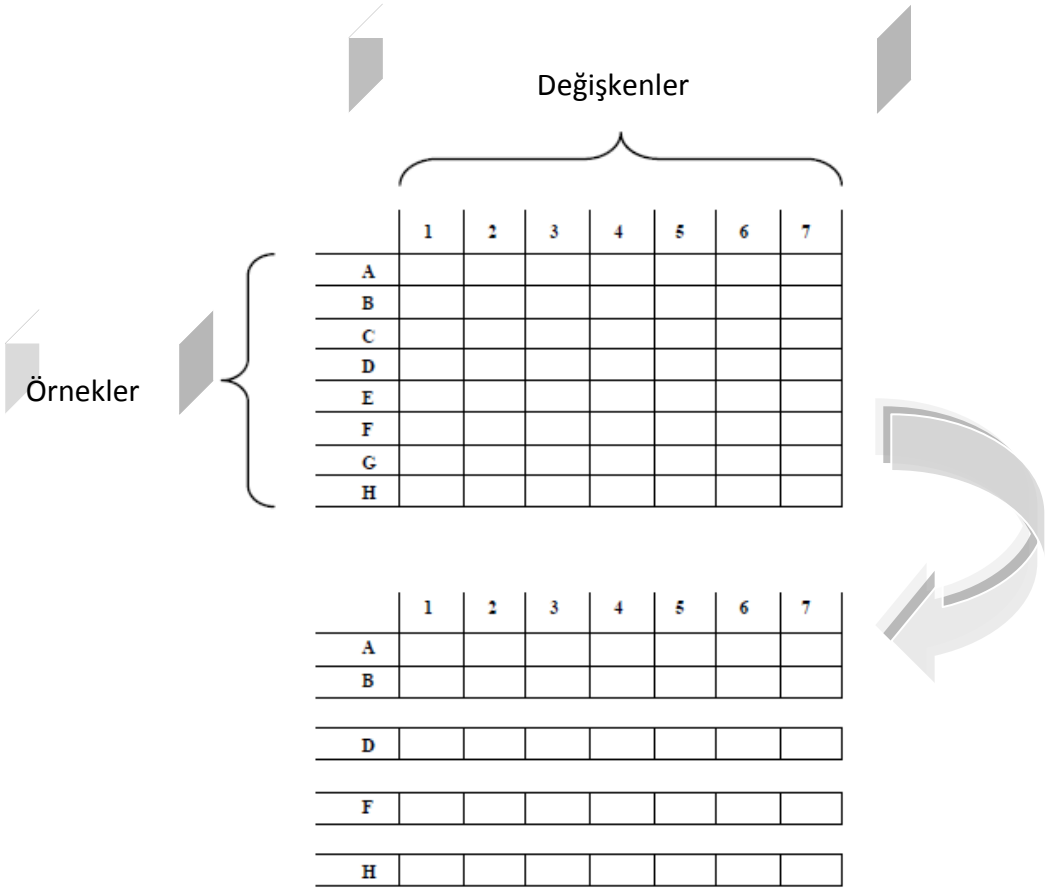
ÖZDÜZENLEYİCİ HARİTALAR

Özdüzenleyici harita tekniklerini anlatmadan önce, özdüzenleyici haritaların temel yapısını oluşturan vektör nicemleme ve vektör yansıtımı üzerinde durulması konunun anlaşılması bakımından yararlı olacaktır. Bu kavramlar anlaşılmadan özdüzenleyici haritalara değinmek bazı kavramların yetersiz ve eksik kalmasına neden olabilir.

3.1 Vektör Nicemleme

Vektör nicemleme (vector quantization), orijinal veri kümesini yeniden üretmeyi amaçlayıp, orijinal veriyi temsil edecek olan w_i , $i = 1...M$, vektörlerini elde etmektir. Bir başka ifade ile fazla miktardaki veri vektörlerini, nispeten daha az vektör ile tanımlama ve temsil etme işlemine vektör nicemleme (VN) denir. Burada yapılmak istenen, Şekil 3.1'deki gibi girdi vektörlerini başka vektörler ile temsil ederek girdi vektörlerinin sayısını azaltmaya çalışmaktır. VN sonrası elde edilen vektörler orijinal vektörlerden oluşmayabilir. Farklı yollarla, örneğin girdi vektörlerinin ortalaması alınarak oluşturulmuş olabilirler.

Peki bu niçin önemlidir? İşlemsel maliyetin minimize edilmesi veri madenciliği konularında önemli bir yer tutmaktadır. Vektör nicemleme sürecinde örnek sayısının azaltılmasından dolayı işlemsel maliyette bir azalış meydana gelir. Ayrıca nicemleme sonucunda temsîli vektörler, veri örneklerinin ortalamaları olarak şekillendiği için veride mevcut olan aykırı değerler büyük miktarda yok edilir [17].



Şekil 3. 1 Vektör nicemlemenin şematik gösterimi [5]

Bu temsîli vektörleri elde etmede en çok bilinen ve kullanılan yöntem k-ortalama (k-means) [18] algoritmasıdır. VN algoritmasının kalitesini ortaya koymak için w vektörüne ait nicemleme gürültüsü, $N_{gürültü}(w)$ şu şekilde hesaplanır;

$$N_{gürültü}(w) = \sum_{y \in Kw} \| y - w \| \tag{3.1}$$

y ; K_w kümesine ait eleman

K_w ; w vektörüne haritalanan örneklerin kümesi

Böylelikle, nicemleme gürültüsü ile verinin ne derece tutarlı bir şekilde temsil edildiği incelenebilir. Özdüzenleyici haritalar doğrusal fonksiyon (linear initialization) kullanılarak başlatıldığında eğitim süresince nicemleme gürültüsünün genellikle azalım gösterdiği görülür (Topoğrafik gürültü ise nicemleme gürültüsünün aksine artış gösterir).

Nicemleme gürültüsü ve topoğrafik gürültü özdüzenleyici haritaların kalite metrikleri olarak anılırlar. Bu metrikler gözlenerek ne derece kaliteli bir haritalama yapıldığı saptanır. Bu metriklere daha detaylı olarak ileride değinilecektir.

3.1.1 Bir Nicemleme Algoritması Olarak Özdüzenleyici Haritalar

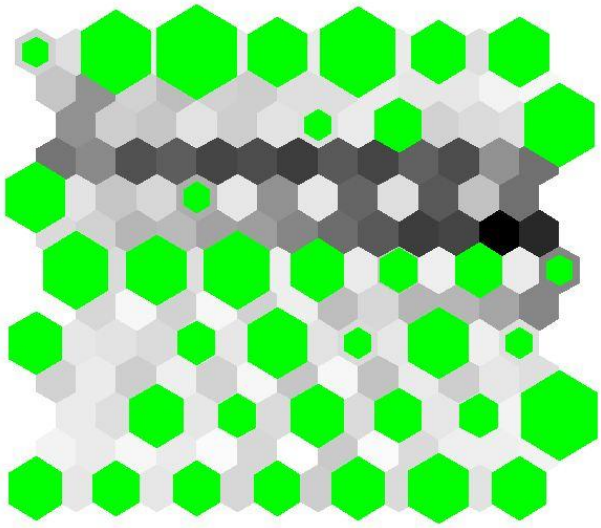
Özdüzenleyici harita algoritması (kısaca, SOM algoritması), k-ortalama algoritması ile yakın bir çalışma mantığına sahiptir. Klasik vektör nicemlemesi ile özdüzenleyici haritalar arasındaki fark, özdüzenleyici haritaların her harita birimine ait komşuluk ilişkilerinde yerel düzenlemeler yapmasıdır. Dolayısıyla özdüzenleyici harita uygulamalarında prototiplerin sıralı biçimde oluştuğu gözlenir. Fakat, σ komşuluk yarıçapı eğitim esnasında azaltılırsa öğrenme açısından daha dirençli bir nicemleme süreci elde edilmektedir [17].

Vektör nicemlemede iki tür sıkıntı ile karşılaşılmaktadır;

- *Sınır değerleri problemi:* VN açısından özdüzenleyici haritaların bazı dezavantajları bulunmaktadır. Bunlardan birisi, veri kümesinin sınır bölgelerinde bulunan birimlerin komşuluk ilişkileri ile iç bölgelerde bulunan birimlerdeki komşuluk kavramının birbirinden farklılık göstermesidir. Bunun nedeni, sınır birimlerindeki komşuluk ilişkisi simetrik olmadığı için veri yoğunluğu bakımından bölgeler arası bir farklılık ortaya çıkmaktadır. Sınır birimlerinde komşu sayısı, veri iç bölgelerindeki birimlere göre daha az olur. Bu durum aykırı değerlerin azalmasında önemli bir avantaj yaratıyor gibi gözükse de aslında özdüzenleyici haritaların bir eksikliğidir. Sınır bölgelerindeki birimler girdi uzayında daha büyük bir Voronoi bölgesine sahip olurlar ve doğal bir netice olarak da sınır bölgelerinin daha fazla BMU olarak seçildiği gözlenecektir (Şekil 3.2).
- *Enterpolasyon birimleri:* Girdi verisi kesikli olduğu zaman, özdüzenleyici haritalarda bulunan bazı nöronlar hiçbir girdi vektörü tarafından seçilmeyebilir. Yani, bazı nöronların hiçbir zaman BMU olmadığı durumlardır. Bunun neticesinde bazı harita birimleri veri kümesinin yoğun veriye sahip

bölgelerinden uzakta kalırlar. Bu birimler enterpolasyon birimleri olarak adlandırılmaktadır.

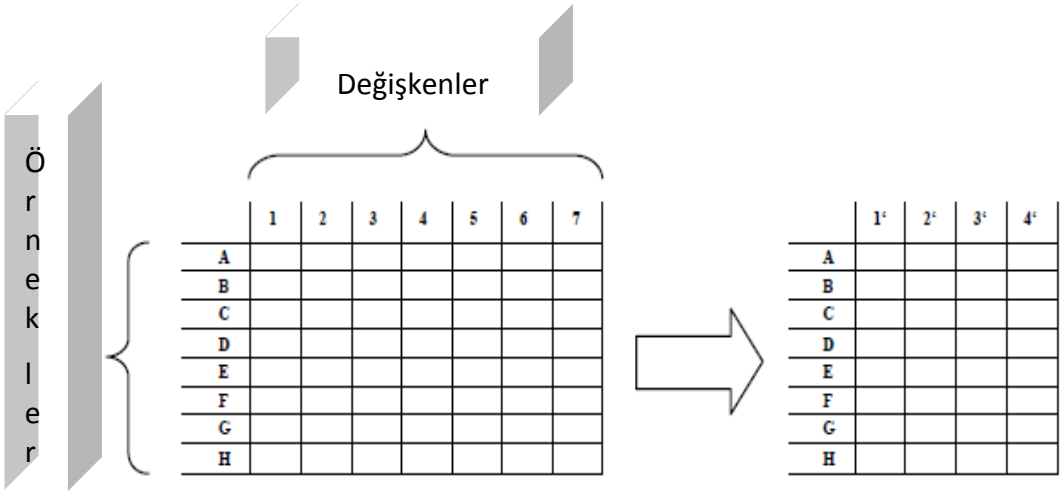
Özdüzenleyici haritaların kısıtlamaları adına daha fazla bilgi için Flexer'ın [19] çalışması incelenebilir.



Şekil 3. 2 Iris veri kümesi için 7x7 SOM. Kıyı hücrelerin daha fazla BMU seçilmesinin hit histogramı ile gösterimi

3.2 Vektör Yansıtımı

Vektör yansıtım (VY) metotları (visualization methods), yüksek boyuttaki verinin, veriler arası uzaklıkların ya da veri diziliminin korunması koşuluyla, düşük boyutlu düzeneklere aktarımını sağlamaya yöneliktir. Görselleştirme veri madenciliğinde, veri yapısının ortaya konulup bu veri hakkında detaylı bir bilgi sahibi olmak açısından çok önemlidir. Çok boyutlu verileri, renk grafikleri, şekil haritaları vb. görsel yöntemler ile iki boyutlu düzlemlerde temsil etmeye yönelik çalışmalar önceleri sıkça başvurulan yollardı. Fakat veri boyutu arttıkça veri analizi bu yöntemlerle zorlaşmaktadır. Daha sonraları ise çok-boyutlu ortamdan düşük-boyutlu (genellikle iki boyutlu düzlemelere) ortamlara veri aktarımı için farklı algoritmalar geliştirilmiştir. Çok boyutlu ölçekleme (multidimensional scaling, MDS), eğrisel bileşen analizi (curvilinear component analysis, CCA) [21], Sammon haritalaması (Sammon’s mapping) [22], GTM (generative topographic mapping) [23], temel bileşen analizi (principal component analysis, PCA) [3], VQ-P, eğrilik çizgisi (principal curves) yaygın olarak bilinen ve kullanılan yansıtım algoritmalarıdır.



Şekil 3. 3 Vektör Yansıtımı [5]

Vektör yansıtımı metotları doğrusal ve doğrusal-olmayan haritalama olarak iki grupta incelenmektedir. Doğrusal haritalamada genellikle iki-boyutlu düzlemlere geometrik yansıtımlar uygulanırken doğrusal olmayan haritalamada ise veri kümelerindeki daha karmaşık verilerin ortaya çıkartılması amaçlanır. Doğrusal olmayan haritalama [24] algoritmaları aykırı değerlere karşı, doğrusal algoritmalardakine oranla daha hassas olsalar da uygulanabilirlik ve geliştirilebilirlik yönünden zayıf kalırlar. Sammon haritalaması, CCA, MDS doğrusal-olmayan haritalama yöntemlerine birer örnektir. PCA ise doğrusal haritalama yöntemine göre çalışan bir yöntemdir.

Yukarıda belirtilen metotlardan PCA, CCA ve Sammon haritalaması aşağıda verilmiştir.

3.2.1 Temel Bileşen Analizi (PCA)

Temel bileşen analizi, veri analizi ve veri sıkıştırmada kullanılan klasik bir istatistiksel metottur. Rastgele bir değişkenin, varyans ve kovaryans ikilisini ele alarak yüksek boyutlu veri kümelerinin daha düşük boyuta indirgenmesini gerçekleştirir. Yani m-boyutlu girdi verisinin d-boyutlu yeni değişkenler ve örnekler topluluğuna dönüştürülmesidir (m > d). PCA, CCA ve Sammon haritalamasından farklı olarak doğrusal haritalama grubuna dahildir ve tekil değer ayrıştırma (SVD) yöntemine benzerlik gösterir. PCA tekniğinin bir avantajı da özvektörlerin, birkaç doğrusal eşitliğin çözülmesi ile kolayca hesaplanabilmesinden dolayı hızlı çalışmasıdır.

PCA'da yeni boyutlar, veriye ait değişkenleri en fazla kapsayacak şekilde olmalıdır. Bunun için öncelikle, veriye ait kovaryans matristen $e_1, e_2, e_3, \ldots, e_d$ özvektörleri ve $\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_d$ özdeğerleri hesaplanır. Daha sonra ise, özdeğerler azalan sırada ($\lambda_1 > \lambda_2 > \ldots > \lambda_d$) sıralanır. Genellikle e_1 özvektörünün değişkenlerin çoğunu kapsadığı görülür. Yine ikinci en büyük özvektör daha sonraki en büyük değeri kapsar ve bu işlem bu şekilde devam eder. En büyük birkaç özvektör tarafından verinin uzaya dağıtılması ile değişkenlerin hemen hepsi korunmuş olur ve bu özvektörlere karşılık gelen özdeğerlerin toplamı ile de yansıtımda korunma oranı hesaplanır. Böylece yansıtımın hata oranı da incelenebilir. Örneğin, 2-boyutlu bir ortama yansıtım şu şekilde tanımlanır;

$$y = \begin{bmatrix} e_1^T \\ e_2^T \end{bmatrix} . w \tag{3.2}$$

PCA'da yeni verinin kaç boyut ile temsil edileceği, yani temel bileşenlerin sayısı önemli bir husustur. Temel bileşen sayısının tespiti konusunda farklı yaklaşımlar kullanılarak uygun değer elde edilebilmektedir. Bu yaklaşımlardan bazıları, scree-plot [25], kırık çubuk (the broken stick) [24], değişim yüzdesi (percent variance) [25], sıralı testler (sequential tests) [25] ve açıklanmış değişme (variance explained) [7]'dir. Scree-plot yöntemi ile yeni değişkenlerin, verinin ne kadarını kapsadığı grafiksel olarak elde edilebilir ve bu sonuçlara göre bileşen sayısı seçilebilir. Şekil 3.4'te iris veri kümesinin scree-plot ile grafiksel sonuçları verilmiştir. Iris veri kümesi için bakıldığında en büyük bileşen, verinin %90'dan fazlasını örtmekte, iki bileşenin ise verinin hemen hemen tamamını örttüğü görülmektedir. Iris veri kümesi için iki bileşenli yeni yapının neredeyse veri kayıpsız bir şekilde ilk değişkenleri temsil ettiği söylenebilir.

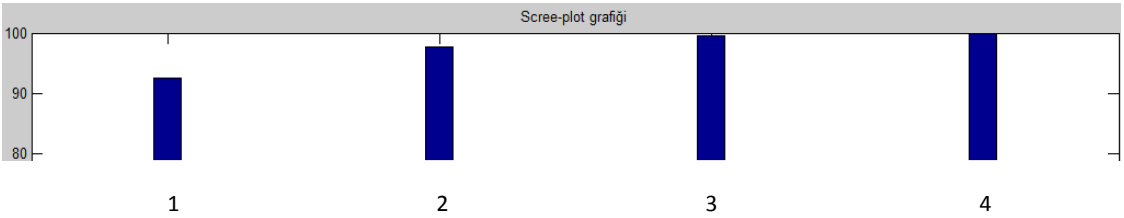
Değişim yüzdesi yöntemi, scree-plot gibi sık başvurulan bir yöntemdir. Seçilen bileşenlere ait özdeğerlerin toplamı, tüm bileşenlere ait özdeğerlerin toplamına bölünür ve bu sonuç belli bir eşik değerin üstünde ise seçilen bileşenlerle sağlıklı bir analiz yapılabilir.

$$\frac{\lambda_1 + \lambda_2 + \ldots + \lambda_b}{\lambda_1 + \lambda_2 + \ldots + \lambda_d} \geq \psi \tag{3.3}$$

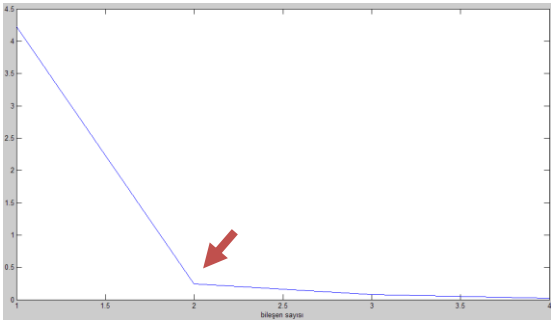
ψ : Önceden belirlenmiş eşik değeri (%90, % 80 gibi.)

b : Seçilen bileşen sayısı

d : Toplam bileşen sayısı



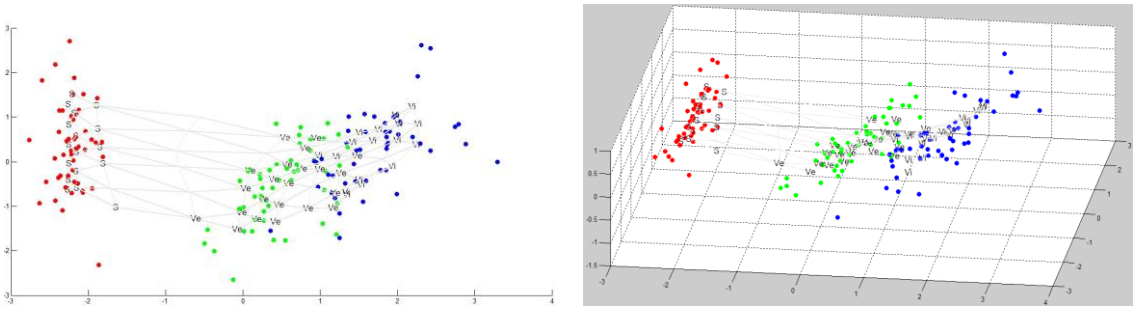
Şekil 3. 4 Scree-plot: İris veri kümesi



Şekil 3. 5 Iris: Varyans açıklamalı gösterim. Yatay eksen, özvektörleri (önem sırası soldan sağa); Dikey eksen, özvektörlere karaşılık gelen özdeğerleri göstermektedir.

Şekil 3.5'de verilen variance explained grafiği iris veri kümesi için oluşturulmuştur. Yatay eksen özvektörleri, dikey eksen karşılık gelen özdeğerleri göstermektedir. Bu yöntem ile kırılmanın en fazla olduğu noktaya bakılarak bu noktaya karşılık gelen özvektör sayısı en uygun değer olarak seçilebilir(Kırılmanın en yüksek olduğu noktalar şekilde "ok işareti" ile gösterilmiştir). Çünkü bu noktada, özdeğerde çok keskin bir azalım gerçekleşmektedir. Başka bir ifade ile bileşenlerin kapsadığı veri miktarında ani bir düşüş meydana gelmektedir. Grafiklerden de görüldüğü üzere, her iki veri için de bileşen sayısı 2 olduğu an büyük bir düşüş gerçekleşmektedir. Dolayısıyla iki bileşenli yapılar her iki veri kümesi için de en uygun sonuçları üretmektedir.

Şekil 3.6'da iris veri kümesine ait PCA uygulması verilmiştir.



Şekil 3. 6 PCA: İris (a) 2-boyutlu (b) 3-boyutlu

3.2.2 Eğrisel Bileşen Analizi (CCA)

CCA [21], Demartines tarafından yüksek boyutlu verilerin boyutlarının azaltılması ve analizi için ortaya atılan bir stratejidir. CCA, iki adımlı bir çalışma mantığına sahiptir. İlk olarak vektör nicemlemesi daha sonra da nicemlenmiş vektörlerin, bir çıktı sistemine doğrusal olmayan yansıtım yöntemleri ile yansıtımı gerçekleştirilir. Yansıtım süreci diğer doğrusal olmayan haritalama algoritmalarındaki (Sammon haritalaması, MDS, vb.) gibi gerçekleşirken, nicemleme tarafında CCA algoritması sabit bir örgü yerine, verinin şeklini alabilen elastiki bir örgü kullanır [21]. Ayrıca, birbirine yakın birimlerin ağırlıkları daha fazla verildiği için, küçük uzaklıklar daha iyi korunur.

CCA, Sammon haritalaması ve MDS (Multi-dimensional scaling) doğrusal olmayan yansıtım metotları olup, birimler arası uzaklıkların girdi ve çıktı uzaylarında oluşturdukları farka göre şekillenen bir "enerji fonksiyonu" oluştururlar ve bu fonksiyonu minimum seviyeye çekmeye çalışırlar [7]. CCA algoritması için bu fonksiyon şu şekildedir;

$$Gürültü_{CCA} = \sum_{i=1}^{N} \sum_{j=1}^{N} (d_{ij}^{'} - d_{ij})^2 / e^{-d_{ij}} \tag{3.4}$$

$d_{ij}^{'}$: girdi uzayında birimler arası öklid uzaklığı

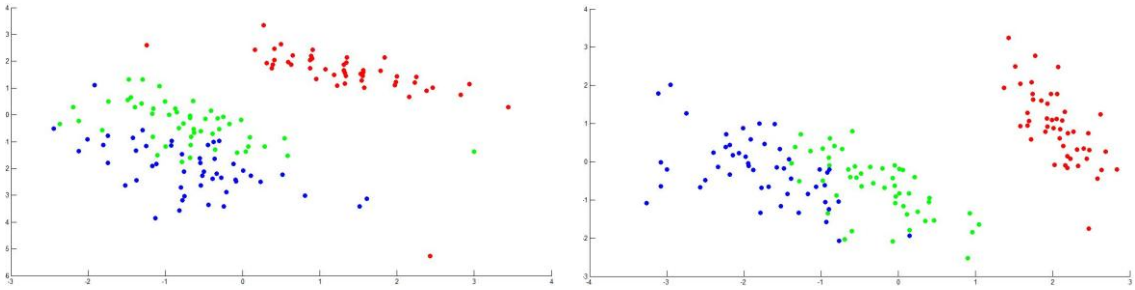
d_{ij} : Örgü üzerinde birimler arası öklid uzaklığı

CCA, çıktı örgüsündeki küçük uzaklıklara yoğunlaşırken Sammon haritalaması ise girdi uzayındaki küçük uzaklıklara vurgu yapar [17]. Şekil 3.7(a)'da CCA yöntemi uygulanan iris veri kümesinde verinin ortada yoğunlaştığı, Şekil 3.7(b)'de Sammon haritalaması uygulandığında ise daha dağınık bir sonuç elde edildiği görülmektedir.

3.2.3 Sammon Haritalaması

Sammon [22] tarafından geliştirilen bir diğer doğrusal olmayan haritalama yöntemidir. Sammon haritalaması, PCA yöntemine alternatif olarak geliştirilmiştir. Fakat PCA tekniğinde olduğu gibi doğrusal değil, doğrusal olmayan haritalama yöntemine göre çalışır. İşlemsel olarak ise en karmaşık olan VY yöntemidir[22].

$$Gürültü_{Sammon} = \sum_{i=1}^{N} \sum_{j=1}^{N} (d_{ij}^{'} - d_{ij})^2 / d_{ij}^{'} \tag{3.5}$$



Şekil 3. 7 İris (a) CCA (b) Sammon haritalaması

3.2.4 Yansıtım Algoritması Olarak Özdüzenleyici Haritalar

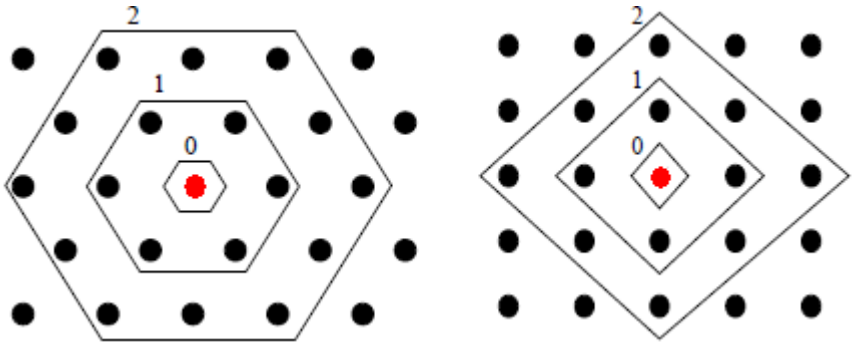
Bir VY algoritması olan SOM, doğrusal olmayan haritalama mantığındadır. SOM, veriler arası orjinal uzaklıkların korunmasından (CCA algoritmasında olduğu gibi) ziyade, prototip vektörleri önceden tanımlanmış olan bir örgü üzerine topolojik ilişkileri koruyarak aktarmaya çalışır. Eğer iki veri SOM örgüsü üzerinde birbirine yakın olarak konumlanmışsa bu veriler çok boyutlu ortamda da birbirlerine yakın bir komşuluk ilişkisi sergiler. Yani, SOM orjinal verideki komşuluk ilişkilerinin örgü üzerinde de korunmasının garantisini vermektedir.

3.3 SOM Metodu

Temel olarak özdüzenleyici haritalarda veri, her biri m-boyutlu $x = [x_1, x_2, x_3, ..., x_m]^T$ örneklerinden oluşmaktadır. Harita ağırlık vektörlerinden yararlanarak veri kümesine uyum sağlamaya çalışır. Süreç sonucunda ağırlık vektörlerin yardımıyla eldeki veri uzayının düşük-boyutlu bir haritası elde edilir. Veri uzayında topolojik olarak birbirine

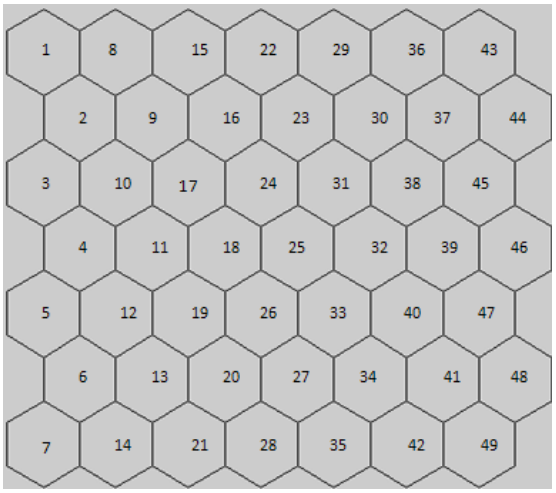
yakın bulunan birimler, oluşturulmuş olan harita üzerinde de topolojik açıdan yakın komşuluk ilişkisi sergiler.

Özdüzenleyici haritalar yapısal olarak iki farklı şekilde olabilmektedir; altıgenimsi ve dikdörtgenel örgü:



Şekil 3. 8 SOM örgüleri, 5x5 boyutunda (a) Altıgenimsi örgü (b) Dikdörtgenimsi örgü. Harita üzerinde bulunan çizgiler, kırmızı renk ile belirlenmiş olan nöronun komşuluk derecelerini belirtmekte; $K_i = \{ j \mid\mid\mid r_i - r_j \mid\mid \leq \kappa \}, \kappa = 1,2,3.$ [5]

Genellikle, aksi belirtilmedikçe "komşuluk" kavramı kullanıldığında $\kappa = 1$ olarak algılanmalıdır. Yani bir SOM örgüsünde yer alan düğüme ait, eğer altıgenimsi örgü kullanılmışsa altı, dikdörtgenel örgü kullanılmışsa dört adet komşu bulunmaktadır. Şekil 3.8-a'da altıgenimsi örgü kullanıldığında kırmızı nokta ile belirtilen nörona ait komşular "1" numaralı çizginin iç bölgesinde kalan altı birimdir. Yine benzer olarak Şekil 3.8-b'de dikdörtgenel örgü için komşuluk ilişkisi verilmiştir. Harita örgüsündeki birimlerin indisleri ise Şekil 3.9'da gösterilmiştir.



Şekil 3. 9 Altıgenimsi 7x7 örgü üzerinde harita birimlerinin indisleri

Özdüzenleyici haritalar “rekabete dayalı bir öğrenme” yöntemi ile çalışmaktadır. Bu yöntemde, ağa ait çıktı nöronları “etkinleşmek” (aktifleşmek), yani kazanan nöron olmak için birbirleriyle rekabet halindedirler ve neticede her bir zaman diliminde sadece bir nöron seçilebilmektedir [2]. Şekil 3.10’da görüldüğü gibi girdi uzayındaki veri ile düşük-boyutlu harita üzerinde yer alan her bir nöron arasında sinaptik bağlantılar oluşturulmakta ve bu süreç sonucunda, girdi uzayındaki örneğe (vektöre) en yakın olan ya da başka bir deyişle en fazla benzerliğe sahip olan çıktı düğümü etkin (aktif) hale gelmektedir. Bu etkin hale gelen nöron *Kazanan Nöron* (Best-Matching Unit, BMU) olarak anılır. Ağın ilk olarak kurulmasından itibaren gerçekleşen süreç, Haykin [2] tarafından üç ana süreçte incelenmektedir; *rekabet*, *ortak çalışma ve sinaptik uyum.* Rekabet sürecinde, her bir girdi vektörüne ait kazanan düğümler (nöronlar) belirlenmekte, daha sonra ortak çalışma sürecinde komşu nöronlar ile ilişkilerin ele alınması açısından her bir aktif düğüme ait komşu düğümlerin özdüzenleyici harita üzerindeki koordinatları belirlenmekte ve son olarak da sinaptik uyum sürecinde, komşu nöronların ağırlık vektörlerinde güncellemeler gerçekleştirilerek bir sonraki aşamada girdi kalıbı için daha uygun bir uygulama gerçekleştirilmektedir. Bu aşamalar, Haykin'in [2] çalışması baz alınarak açıklanmıştır.

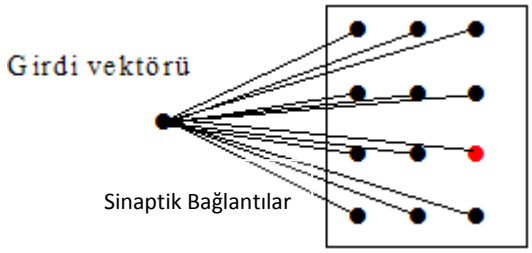
Rekabet

m, girdi uzayının boyutunu temsil etmek üzere, girdi uzayından rastgele olarak bir x vektörü, $x = [x_1, x_2, x_{3,}..., x_m]^T$ seçilmiş olsun. Harita yapısında bulunan her bir nörona ait sinaptik-ağırlık vektörlerinin boyutları da girdi uzayının boyutu ile aynı olmak durumundadır. Örneğin, ağ üzerinden seçilmiş olan bir j nöronuna ait sinaptik-ağırlık vektörü şu şekilde tanımlanabilir;

$$w_j = [w_{j1}, w_{j2}, w_{j3},..., w_{jm}]^T, \quad j = 1,2,3,..., l \tag{3.6}$$

Eşitlikte belirtilen l, SOM örgüsü üzerinde bulunan toplam nöron sayısını belirtmektedir. x girdi vektörüne ait kazanan nöronun (BMU) belirlenmesi için tüm ağırlık vektörleri sırasıyla x vektörü ile iç çarpım (skaler çarpım) $(w_j^T x)$ işlemine tabi tutulurlar. İç çarpım sonucu en büyük değeri hangi ağırlık vektörü oluşturuyor ise, kazanan nöron olarak seçilmektedir. $(w_j^T x)$ iç çarpım değerinin enbüyültülmesi, x

vektörü ile w_j ağırlık vektörleri arası öklid uzaklığının en küçük değere yaklaşması anlamına gelmektedir. İki vektör arası öklid uzaklığı azaldıkça da, iki vektör arası benzerliğin aynı derece artması anlamına gelir. Burada, vektörler arası iç çarpım uygulanmakta ve en yüksek iç çarpım değeri seçilerek girdi vektörü x'e özdüzenleyici harita üzerinde bulunan nöronlardan en yakın olanı, başka bir deyişle en benzer olan nöron seçilir.



Şekil 3. 10 Kohonen modelinde sinaptik bağlantılar ve temsili kazanan nöron

A, örgü üzerindeki bütün nöronları temsil eden küme ve $d(x)$ dizini, x vektörüne ait BMU olarak belirtildiğinde, rekabet aşağıdaki denklem ile tanımlanabilir;

$$d(x) = \arg\min_j \| x - w_j \|, \quad j \in \mathrm{A} \tag{3.7}$$

Ortak çalışma

Nörobiyolojik açıdan düşünüldüğünde, beyinde bulunan ve benzer özellikler taşıyan sinir hücrelerinin birbiriyle komşuluk ilişkisi içinde bulunduğu görülmektedir. İnsan beyni incelendiğinde, benzer duyulara ait sinir hücreleri öbeklenmiş durumda olduğu görülür. Özdüzenleyici haritalar üzerinde bulunan nöronlar arasında da bir komşuluk olduğu varsayılır ve nöronlar arası komşuluk ilişkileri için bir *komşuluk fonksiyonu* tanımlanır. Bu fonksiyona göre kazanan nöron ile bu nörona ait en yakın komşuları arasındaki topolojik komşuluklar ortaya koyulabilmektedir.

$h_{j,i}$, i kazanan nöronu ile i'ye ait komşuları (j) arasındaki topolojik komşuluk ilişkilerini ortaya koyan bir fonksiyon; $d_{j,i}$, kazanan nöron i ile komşu j nöronu arasındaki örgüsel uzaklığı gösterdiğinde;

Topolojik komşuluk $h_{j,i}$ örgüsel uzaklık olan $d_{j,i}$'ye ait unimodal fonksiyondur [2]. Bunun anlamı $d_{j,i}$ sıfır olduğu zaman, $h_{j,i}$ en büyük değerine ulaşmış olur. Benzer şekilde $d_{j,i}$ örgüsel uzaklığı artış gösterdiğinde, $h_{j,i}$ sıfıra yakınsamaya başlayacaktır.

$$h_{j,i(x)} = \exp(-\frac{d_{j,i}^2}{2\sigma^2}), j \in \mathrm{A} \tag{3.8}$$

σ ; topolojik komşulukta en etkin genişlik(komşuluk yarıçapı)

σ , eğitim süresince ($h_{j,i}$ de olduğu gibi) azalım göstermektedir.

Komşuluk Kerneli

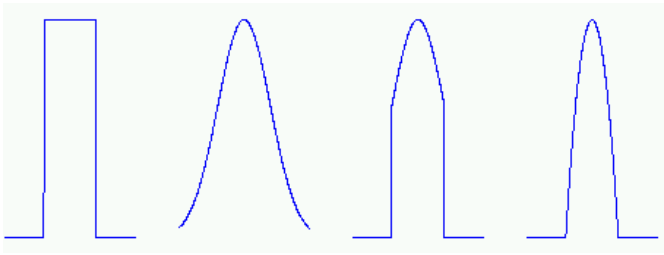
Eğitim ve eğitim sonrası özdüzenleyici haritalarda önemli kavramlardan birisi komşuluk kerneli'dir ve bu fonksiyon monoton bir şekilde azalım gösterir. Çalışma prensibi olarak; iki birim arasındaki mesafe değerine göre birimler arası yakınlığı ortaya koyar. Birbirinden uzak birimler için komşuluk değeri düşük olurken, birimler arası mesafe azaldıkça bu birimlere ait komşuluk değeri artış gösterir. Kazanan birimde kernel maksimum seviyeye ulaşır.

Kernel, aynı zamanda istatistiğin çoğu alanında (örn., olasılık yoğunluk tahminleri) kullanılan bir kavramdır ve parametrik olmayan tahmin tekniklerinde kullanılan bir ağırlıklandırma işlevidir. Kerneller, kernel yoğunluk tahminlerinde (kernel density estimation) rastgele değişkenlerin yoğunluk fonksiyonlarını ya da kernel regresyonlarında (kernel regression) rastgele bir değişkenin koşullu beklentisinin bulunmasında kullanılır [1]. Zaman serilerinde (time-series) de yine önemli bir yere sahiptir.

Bu noktada kullanılan gelen bir çok kernel fonksiyonu mevcuttur. Bunlardan en bilineni ve ayrıca bu tezde de kullanılan Gauss kerneli'dir.

SOM Toolbox [16] içerisinde mevcut olan komşuluk fonksiyonlarının sayısı dörttür ve bunlar; kabarcık (bubble), Gauss (Gaussian), kesik Gauss (cut Gaussian) ve epanechicov komşuluk fonksiyonları şeklindedir. Komşuluk fonksiyonları, nöronların birbirine ne kadar sıkı bir biçimde bağlı olduğunu ortaya koyar. Bu fonksiyonlar içinde en basit olanı

kabarcık fonksiyondur. Kabarcık fonksiyon, kazanan nöronun çevresinde belli bir değere sahipken diğer bölgelerde sıfır değerini alır.



Şekil 3. 11 SOM Toolbox'ta komşuluk fonksiyonları (a) Kabarcık (bubble) fonksiyon (b) Gauss (Gaussian) fonksiyonu (c) Kesik Gauss (Cut-Gaussian) fonksiyonu (d) Epanechicov fonksiyonu (aslında, maks(0, $1-y^2$) şeklinde olan bir fonksiyondur) [16]

Uyum süreci

Kazanan nöronların seçilmesi ve komşulukların ortaya koyulmasının sonrasında, komşu nöronlar bir "*Kendini Güncelleme Süreci*" (KGS)'ne dahil olurlar. Bu süreçte BMU'ya ait komşu nöronlar ve özdüzenleyici haritalarda bulunan diğer nöronlar girdi vektörüne yaklaşmak için hareket ederler. k zamanındaki j nöronuna ait sinaptik-ağırlık vektöründen, $k+1$ zamanındaki sinaptik-ağırlık vektörü elde edilmesi (güncelleme) şu şekildedir;

$$w_j(k+1) = w_j(k) + \eta(k)h_{j,i(x)}(k)(x(k) - w_j(k)) \tag{3.9}$$

Verilen KGS süreci özdüzenleyici haritalarda bulunan ve i nöronuna komşu olan tüm birimler tarafından uygulanır. Bu işlem belli bir sayıda tekrarlandıktan sonra, özdüzenleyici harita birimleri girdi uzayındaki vektörlerin topolojik oluşumunu yakalamaya başlayacaktır. Böylece girdi uzayında komşu olan birimler, örgü üzerinde de birbirlerine yakın pozisyonlara hareket edecek ve "topolojik korunma" gerçekleşmiş olacaktır.

$\eta(k)$, öğrenme parametresi zamanla azalım gösterir. Başlangıçta belirlenen bir değerden (η_0) başlayarak gitgide azalım gösterir.

Aynı zamanda, $h_{j,i}$ komşuluk fonksiyonu da zaman bağımlı bir değişkendir ve zamanla dinamik olarak güncellenir ve azalım gösterir.

Basit SOM Algoritması;

$x = [x_1, x_2, x_3, ..., x_m]^T$, m boyutlu girdi uzayından bir örnek ve j nöronuna ait ağırlık vektörü de, $w_j = [w_{j1}, w_{j2}, w_{j3}, ..., w_{jm}]^T$ olmak üzere;

1-boyutlu bir uzayda tanımlı olan özdüzenleyici haritalar için tanımlı algoritma şu şekildedir:

Adım 1: İlk atamaların yapılması

Sinaptik-ağırlık vektörlerine ilk değerleri ata.

Öğrenme katsayısını ($\alpha(t)$) ata.

Komşuluk derecesini(κ) ve komşuluk fonksiyonunu($h_{j,i}$) ata.

Adım 2: Öklid uzaklığını kullanarak girdi örneği ile her nöron arası uzaklığı hesapla.

$$d(j) = \sum_{i=1}^{N} (w_{i,j} - x_i)$$

Adım 3: Girdi verisine en yakın nöronu (BMU) bul.

(Adım 2'de en küçük değere sahip j indisi bu değere sahiptir.)

Adım 4: Verilmiş olan parametrelere göre vektör güncelleştirmesini,

$w_j(k+1) = w_j(k) + \eta(n) h_{j,i(x)}(k)(x(k) - w_j(k))$ denklemine göre yap.

Adım 5: Her girdi verisi için Adım 2-4 ü gerçekleştir.

Adım 6: Öğrenme katsayısını güncelle.

Adım 7: Topolojik komşuluk katsayısını güncelle.

Adım 8: Çalışmanın sonlandırılmasını kontrol et.

Adım 9: Sonlandırma olmadığı sürece Adım 2-8 gerçekleştir.

3.3.1 İlk Değerlerin Atanması

Eğitim sürecinin öncesinde, prototip vektörlere ilk değerlerin atanması gerekir. Bu işlem gerçekleştirilmesi için üç farklı yoldan birisi tercih edilebilir;

- *Rastgele Değer Atama (Random Initialization)*: Ağırlık vektörleri, rassal olarak küçük değerlerle yüklenir.

- *Girdi Referanslı Değer Atama (Sample Initialization)*: Ağırlık vektörleri, girdi verilerden rassal olarak çekilen değerlerle yüklenir.
- *Doğrusal Değer Atama (Linear Initialization):* Ağırlık vektörleri, girdi verisine ait en büyük iki özdeğere (eigenvalue) karşılık gelen iki özvektörün (eigenvector) yine girdi uzayına dağıtılması sonucu yüklenir.

3.3.2 Özdüzenleyici Haritalarda Eğitim Algoritmaları

Temel olarak Özdüzenleyici haritaların eğitiminde iki tip eğitim algoritması kullanılmaktadır. Bunlar; *Sıralı (ya da artımlı) öğrenen algoritma* ve diğeri de *Batch yöntemi'*dir. Yukarıda SOM metodu anlatılırken sıralı-öğrenen algoritma üzerinden gidilmiştir. Özdüzenleyici haritalarda eğitim algoritması denilince de sıralı-öğrenen algoritma akla gelmektedir. Fakat bu çalışmadaki uygulamalarda sıralı algoritma yerine, Matlab'de daha efektif çalışan Batch eğitim algoritması tercih edilmiştir.

3.3.2.1 Sıralı Öğrenen Algoritma

Özdüzenleyici haritalar yinelemeli çalışan bir mekanizmaya sahiptir ve ağırlık vektörlerinin iteratif olarak güncellenmesi bakımından da k-ortalama algoritmasına büyük benzerlik göstermektedir [7]. Her adımda, x veri örneği, eğitim kümesinden rastgele olarak seçilir ve x ile diğer tüm ağırlık vektörleri arasındaki uzaklıklar ölçülür. Uzaklıkların ölçümünde genellikle öklid uzaklığı tercih edilmektedir. $d(x)$, x girdi vektörüne en yakın harita elemanı olmak üzere,

$$d(x) = \arg\min_{j} \| x - w_j \| \tag{3.10}$$

w_j : x vektörünün j komşusuna ait vektör

denklemiyle bulunur. Eğer x vektörü kayıp değer ya da değerlere sahipse, bu değişkenler uzaklık hesaplamada göz ardı edilir [17]. Daha sonra ise, biraz önce de değinilen uyum sürecine göre vektörlerin güncellenmesi gerekmektedir. j vektörüne ait güncelleme denklemi,

$$w_j(k+1) = w_j(k) + \eta(k) h_{j,i(x)}(k)(x(k) - w_j(k)) \tag{3.11}$$

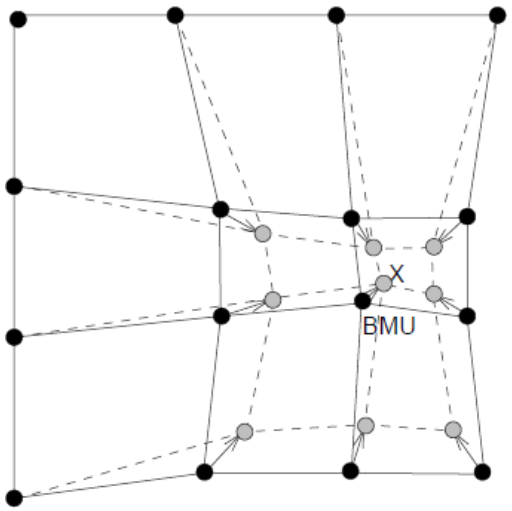
k : eğitim adımı

$\eta(k)$: k anındaki öğrenme katsayısı

$h_{j,i(x)}(k)$: k anında komşuluk katsayısı

şeklindedir. Bu işlem, her x girdisi için tüm w_j komşu vektörlerini kapsayacak şekilde belli sayıda (epok) yapılmalıdır. Komşuluk katsayısı ($h_{j,i(x)}(k)$), kazanan nöronda en yüksek değerine ulaşırken, harita örgüsü üzerindeki uzaklık arttıkça monoton bir şekilde azalım gösterir. Yine, öğrenme katsayısı ($\eta(k)$) da zamanla monoton olarak düşmektedir. Komşuluk katsayısı, farklı dağılımlara sahip olabilir. Fakat genellikle, Gauss dağılımı tercih edilmektedir.

Eğitim süresince, SOM esnek bir ağ gibi hareket eder [17]. Kazanan nörona (BMU) ait komşu birimler, kazanan noktaya doğru hareket ederler (Şekil 3.12). Böylece, BMU ile benzer vektörlere sahip olmaya başlarlar.



Şekil 3. 12 Kazanan Nöron (BMU) ve komşuların güncellenmesi. Güncelleme x girdi örneğinin doğrultusunda gerçekleştirilmektedir. Siyah ve gri renk sırasıyla, güncelleme öncesi ve sonrasını göstermektedir [17]

3.3.2.2 Batch Algoritması

Bir diğer önemli öğrenme kuralı "Batch Map" öğrenme algoritmasıdır. Bu yöntem, sıralı-öğrenme yöntemine göre işlemsel zaman bakımından daha hızlıdır ve "sabit nokta yinelemesi" (fixed point iteration) mantığı ile çalışmaktadır [7]. Bu mantık şu şekilde açıklanabilir; her adımda bütün girdi örneklerine ait BMU'lar bulunur ve ağırlık vektörleri,

$$w_j(k+1) = \frac{\sum_{i=1}^{N} h_{j,i}(k).x_i}{\sum_{i=1}^{N} h_{j,i}(k)} \quad (3.12)$$

N : girdi vektörlerin sayısı

x_i : i'inci girdi vektörü

denklemi ile hesaplanır. Batch yöntemi, sıralı-öğrenme algoritmasından farklı olarak, herhangi bir öğrenme katsayısı ($\eta(k)$) kullanmamaktadır. Yine, sıralı-öğrenme yöntemindeki, örneklerin sırayla alınması kuralı, batch yönteminde bulunmamaktadır ve veri örneklerinin sırası önemli değildir. Eğitim bittikten sonra, özdüzenleyici harita eğitilmiş veriye göre oluşturulur. Bu noktada, komşu birimlere ait vektörlerin birbirine benzerlik göstermesi beklenir.

Komşuluk ilişkilerini konusunda değinilmesi gereken bir başka kavram "Voronoi bölgeleri" (Voronoi regions)'dir. Ağırlık vektörleri, bir "mozaik" topluluğu gibi girdi uzayını kaplarlar. Bu alanların tamamı da Voronoi kümeleri, $V_j = \{x_i \mid\mid\mid x_i - w_j \mid\mid < \mid\mid x_i - w_r \mid\mid, \ \forall r \neq j\}$ olarak anılır ve her biri bir harita birimine tekabül eder. Ağırlık vektörlerinin güncellenmesinde kullanılan bir diğer seçenek de, Voronoi kümesinin kütle merkezlerinin ağırlıklı ortalamalarının hesaplanmasıdır (3.14). Bu eklenti kullanılarak özdüzenleyici haritaların (3.9)'den daha etkin bir uygulamasının gerçeklenmesi mümkündür [5]. Voronoi kümesinin kütle merkezleri;

$$n_k = \frac{1}{N_k} \sum_{w_i \in V_k} w_i \quad (3.13)$$

N_k: Vk'daki örnek sayısı (ya da başka bir ifade ile i ağırlık vektörünün BMU olduğu örnek sayısı)

denklemiyle hesaplanır.

$$w_j(k+1) = \frac{\sum_{i=1}^{M} N_i.h_{j,i}.n_i}{\sum_{i=1}^{M} N_i.h_{j,i}} \quad (3.14)$$

M : prototip vektörlerin sayısı

3.4 Özdüzenleyici Haritalarda Kalite Metrikleri

Özdüzenleyici haritarın eğitiminin ardından elde edilen verilerin ne derece tutarlı ve doğru olduğu, topolojik bozulmaların olup olmadığı, varsa da bozulmanın büyüklüğü, Özdüzenleyici haritaların kalitesinin belirlenmesinde önemlidir. Burada genel olarak iki metrik kalite değerlendirilmesinde göz önünde tutulmaktadır.

Nicemleme Hatası ($N_{gürültü}$)

Nicemleme hatasına (nicemleme gürültüsü) aslında bölüm 3.1'de değinmiştik. Nicemleme hatası, girdi verisinin ağırlık vektörleri tarafından ne derece sağlıklı ve kesin olarak temsil edildiğini göstermektedir. Bir w ağırlık vektörüne ait nicemleme hatası($N_{gürültü}(w)$) eşitlik 3.1'de verilmişti. Özdüzenleyici haritaların kalitesini değerlendirmek adına, ortalama nicemleme hatası şu şekilde hesaplanmaktadır;

$$N_{gürültü} = \frac{1}{M}\sum_{i=1}^{M} N_{gürültü}(w_i) \tag{3.15}$$

M ; ağırlık vektörlerinin toplam sayısı

Özdüzenleyici haritalarda doğrusal değer atama yöntemine göre ilk değerler atandığında, nicemleme gürültüsünün genellikle eğitim süresi boyunca azalım gösterdiği görülür.

Topolojik Hata ($T_{gürültü}$)

Topolojik hata ise vektör yansıtımı sırasında oluşabilecek topolojik bozulmalara yoğunlaşmaktadır. Vektör yansıtımı mantığına göre seçilmiş olan kazanan nöronların (BMU, 2'nci BMU, 3'üncü BMU vb.) harita üzerinde komşu olması beklenir. Eğer bu şekilde bir sonuç oluşmuyorsa vektör yansıtımında topolojik bozulmalar olduğu anlamına gelir. Vektör yansıtımında oluşabilecek olan topolojik bozulma,

$$T_{gürültü} = \frac{1}{N}\sum_{k=1}^{N} u(w_k) \tag{3.16}$$

eşitliği ile hesaplanır. Burada $u(w_k)$;

$$u(w_k) = \begin{cases} 0, \text{Eğer } w_k \text{ ya ait 1. ve 2. BMU SOM örgüsünde yanyana ise} \\ 1, \text{Değilse} \end{cases} \tag{3.17}$$

Özdüzenleyici haritalarda doğrusal değer atama yöntemine göre ilk değerler atandığında, topolojik gürültü genellikle eğitim süresince artış gösterir, fakat ortalama topolojik gürültü ise azalım gösterir.

BÖLÜM 4

GÖRSELLEŞTİRME

Görselleştirme evresi, eğitim sürecinin tamamlanmasının ardından başlar. Veri analizi ve sonuçların elde edilmesi açısından bu evre önemlidir. Yapay Sinir Ağları'nın (YSA) görselleştirilmesi açısından diğer metotların aksine özdüzenleyici haritalar son derece esnek ve kolay uygulanabilir bir yöntemdir. Zaten bu nedenle özdüzenleyici haritalar yaygın olarak kullanılagelmiştir.

Bu bölümde verinin SOM örgüsüne ne şekilde yansıtılabileceği, yansıtma yöntemleri, uzaklık matrisleri, tekli değer ve çoklu değerlerin gösterilmesi, ağırlık vektörlerinin görselleştirilmesi, SOM örgüsünün görselleştirmede ne şekilde kullanılabileceği üzerinde durulacaktır.

Verinin görselleştirmesi konusunda bir çok çalışma yapılmış ve bu çalışmalarda çeşitli tasnif yöntemleri kullanılarak verinin kategorilere ayrılması ve incelenmesi sağlanmıştır. Vesanto [17], görselleştirme tekniklerini genel olarak, *SOM vektörlerinin görselleştirilmesi* ve *SOM örgüsünün görselleştirme aracı olarak kullanılması* olmak üzere iki kısımda ele almıştır. Bunun yanısıra, değişkenlerin görselleştirilmesi ve örgü üzerindeki kümelerin görselleştirilmesi üzerine çalışmalarda bulunmuştur.

Bu alanda bir başka çalışma Pölzlbauer [7] tarafından yapılmıştır. Pözlbauer, Vesanto tarafından yapılan tasnifin genel olarak ikinci kısmında yani SOM örgüsünün bir görselleştirme aracı olarak kullanılması üzerinde durmuştur ve tekil değer görselleştirme ve çoklu değer görselleştirme olarak iki grupta ele almıştır. Çoğu görselleme çalışması tekil değerli olarak yapılsa da çoklu değer görselleştirmeler detaylı bir şekilde veri analizlerinin yapılması açısından önemlidir.

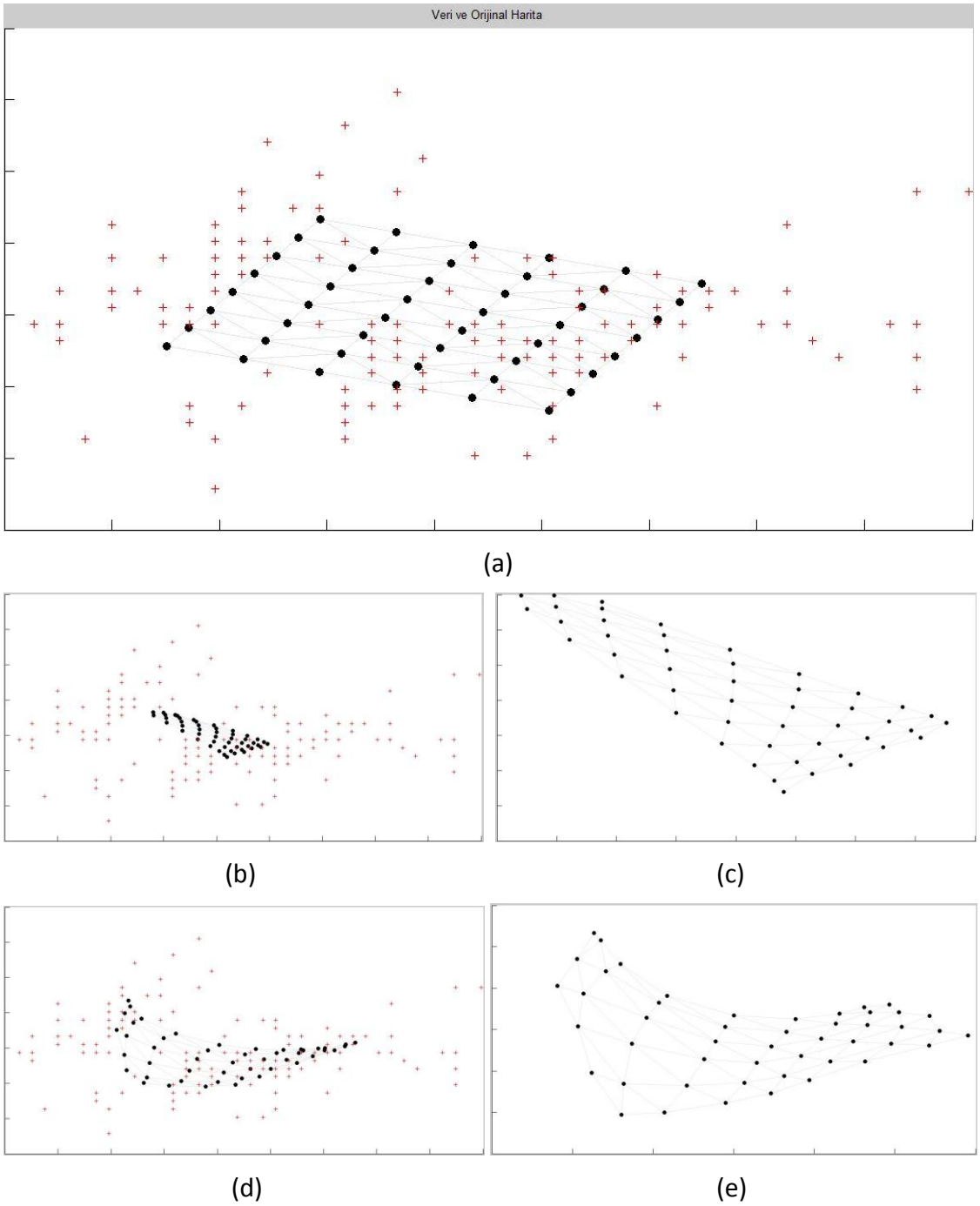
Bölüm 4.1'de SOM vektörlerinin, bölüm 4.2'de SOM örgüsünün ve bölüm 4.3'te kümelerin ve değişkenlerin görselleştirilmesi ele alınmıştır.

4.1 SOM Vektörlerinin Görselleştirilmesi

Özellikle eğitim sürecinde ağırlık vektörlerinin konumlarındaki değişimin ve öğrenimlerinin incelenmesi açısından bu bölüm önemlidir. Eğitim sürecinin her devresinde, örgünün veri üzerindeki hareketleri ve izlediği yol takip edilebilmektedir. Eğitim sürecinin sonunda ise örgünün veriye uyum sağladığı gözlemlenir. Bu tür görselleştirmelerde, SOM örgüsü esnek bir yapıya dönüşmekte olup eğitim süresince BMU'ya olan uzaklığına bağlı olarak 2-boyutlu örgü üzerindeki her birim konumunu günceller. Şekil 4.1'de iris veri kümesi ve 7x7 örgü kullanılarak örnekleme yapılmıştır. Şekil 4.1(a)'da örgü ve verilerin, ilk değerlerin atanmasından sonraki ve eğitime henüz başlanmadan önceki durumu verilmektedir. Şekil 4.1(b), eğitim sürecinin ortasında örgünün durumu verilmiştir. Bu aşamadan itibaren örgü, veri dağılımına benzerlik göstermeye başlamaktadır. Şekil 4.1(c)'de eğitimin orta noktasındaki haritanın durumu (veri örnekleri gösterilmeden) verilmiştir. $\eta(k)$ öğrenme parametresi de bir önceki durumu göre azalım göstermiştir. Şekil 4.1(d), eğitim sürecinin sonunda örgü ve veri durumunu göstermektedir. Şekil 4.2(e) de eğitimin sonunda haritanın durumu (veri örnekleri gösterilmeden) verilmiştir. Süreç sonunda ağırlık vektörleri son şeklini almıştır ve bir sonraki sürece hazır hale gelmiştir.

4.2 Görselleştirmede SOM Örgüsünden Yararlanma

Bu bölümde, veri analizi için 2-boyutlu bir örgüsel sistem kullanılmaktadır. Bunun için örgü üzerindeki her parsel, boyutlandırma, renklendirme, işaretleme, grafikleme vb. teknikleri kullanılarak bilgi edinimi sağlanmaya çalışılır. Bir önceki bölümden farklı olarak örgü eksenleri esnek bir yapıya sahip değildir.



Şekil 4. 1 İris Veri kümesi, 6x8 SOM. Eğitimin farklı zamanlarında veri ve haritaya ait PCA yansıtımları: (a) Süreç başlamadan önce (b)(d) Sırasıyla sürecin ortasında ve sonundaki harita birimleri ve vektörler (c)(e) Sırasıyla sürecin ortasında ve sonunda sadece harita birimleri

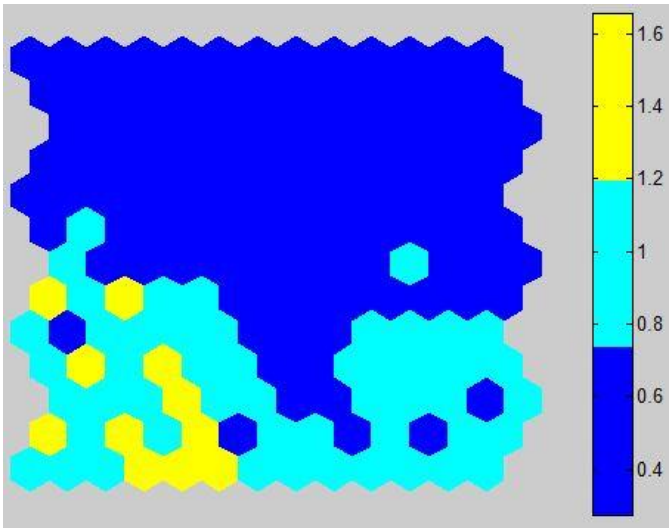
4.2.1 Tekli Değer Görselleştirme

Bu teknikte, her bir harita parçasında sadece tek bir değerin gösterilmesi söz konusudur. Parçalar için, sayısal değerin yanı sıra parsel boyutlama ve renklendirme

gibi farklı teknikler uygulanabilir. Tekli değer görselleştirmelerde uygulanan tekniklerden bazıları şu şekildedir;

Renk Kodlaması

Renklendirmede gri-skala ya da Matlab'in "jet" renklendirmesi yöntemleri seçilebilir. Jet renklendirme tekniğinde en büyük değer, koyu kırmızı ile; en küçük değer de mavi renk ile temsil edilirken, ara değerler de ara renkler ile temsil edilmektedir. İsteğe bağlı olarak belli renk gruplar oluşturularak seçilmiş olan renkler ile kodlamalar gerçekleştirilebilir (Şekil 4. 2).



Şekil 4. 2 Matlab jet renklendirmesi (3 seviyeli renklendirme)

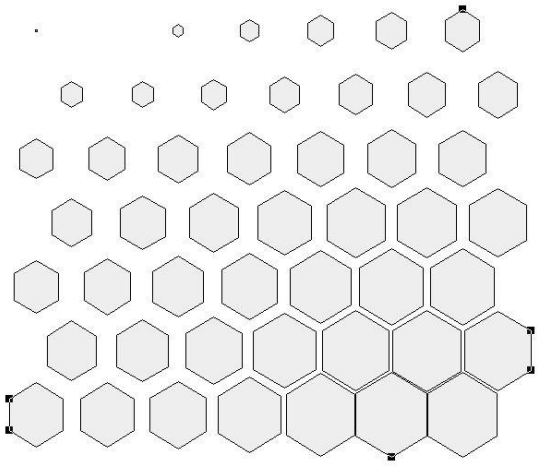
Parsel Boyutlandırma

Bu teknikte parsel boyutu değerlere göre değişim gösterir. Büyük değerlere karşılık büyük parsel tekabül ederken, küçük değerler için de nispeten daha küçük parseller kullanılır. Bu yöntem özellikle, uzaklık matrislerinin görselleştirilmesinde ve hit histogramlarında kullanılır (Şekil 4. 3).

İşaretleme

Az da olsa görselleştirme çalışmalarında karşılaşılan bir tekniktir. Harita biriminin sahip olduğu değere göre işaretlemenin boyutu değişim gösterir ve bu işaretler harita birimlerinin üzerine yerleştirilir. İşaretlemede baz alınan nokta ise en büyük değere karşılık harita biriminin (parselin) tamamı işaretlenir. Harita biriminin dikdörtgenimsi ya da altıgenimsi olması önemli değildir. Önemli olan nokta en büyük değerin karşılık

gelen parseli tamamıyla kaplamasıdır. Bu yöntem parsel boyutlandırma ile benzerlik gösterir. Bölgelere ait yoğunlukların incelenmesi için bu yöntem tercih edilebilir.



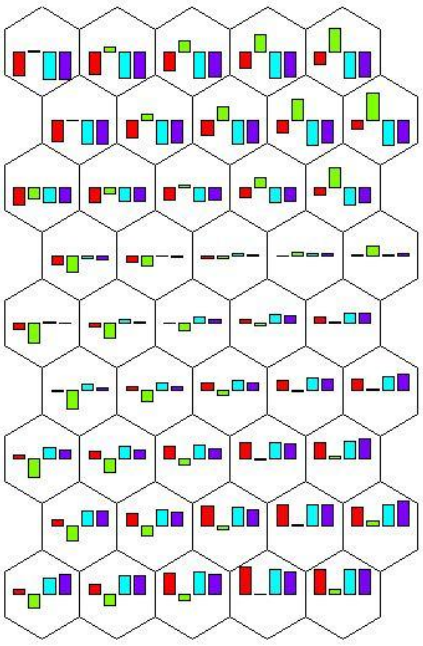
Şekil 4. 3 Parsel boyutlandırma, 7x7 SOM

4.2.2 Çoklu Değer Görselleştirme

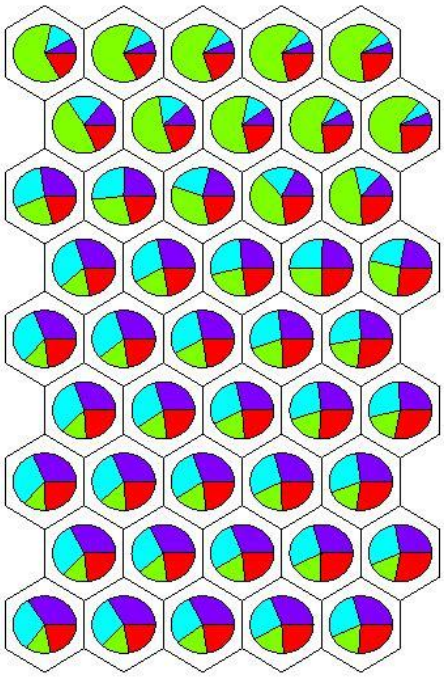
Çoklu değerlerin görselleştirmesinde, bir harita biriminde birden fazla değerin gösterilmesi söz konusudur. Bu işlem sütun grafikleri, dairesel grafikler, nümerik dökümler, renk düzlemleri, vb. yollarla gerçekleştirilebilir. Bu gruba dahil olan teknikler şu şekildedir;

Sütunlu Grafikler

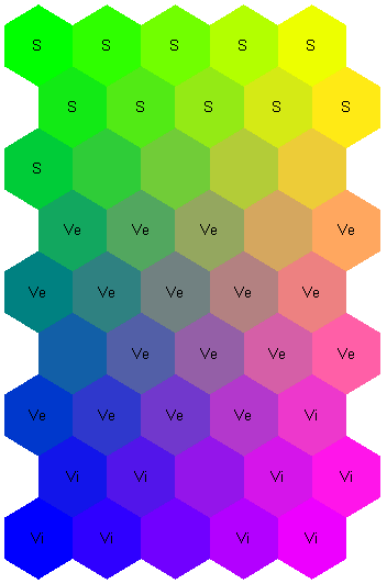
Sütunlu grafiklerde her bir bileşen bir sütun ile temsil edilmektedir ve bu şekilde sütun diyagramları oluşturulur. Bileşen sayısına göre harita parsellerine sütun grafikleri çizilir ve bileşenlerin sahip olduğu değerle doğru orantılı olarak da sütunların boyu değişim gösterir. Bu teknikte genel olarak, her bileşen harita parsellerinde belli bir renk ile temsil edilir (Şekil 4. 4). Tabiki gri-skalanın kullanıldığı yerlerde böyle bir durumdan bahsedilemez. Sütun grafiklerinde değişken sayısı arttıkça verinin gösterilmesi zorlaşmaktadır. Aslına bakılırsa, bileşen sayısında teorik olarak herhangi bir sınırlandırma olmasa da tüm çoklu değer görselleştirmeleri için bileşen sayısının artması bir dezavantajdır. Şekil 4.4'te haritanın üst bölümünün diğer bölgelerden farklı olduğu görülmektedir. Yine Sağ-alt ve sol-orta bölgelerde benzer verilerin toplandığı şekilden anlaşılmaktadır.



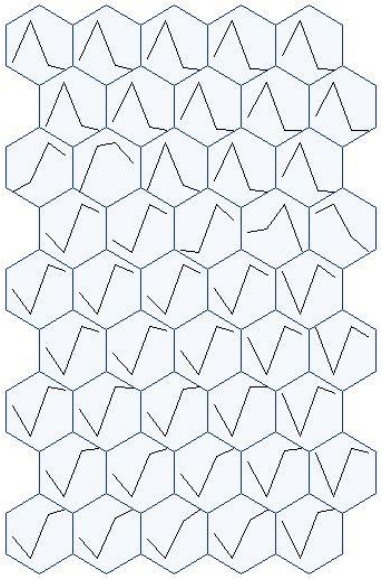
Şekil 4. 4 Sütun grafiği, 9x5 SOM, İris veri kümesi



Şekil 4. 5 Dairesel grafik, 9x5 SOM, İris veri kümesi



Şekil 4. 6 Renk düzlemi, 9x5 SOM, İris veri kümesi



Şekil 4. 7 Sinyal grafiği, 9x5 SOM, İris veri kümesi

Dairesel Grafikler

Dairesel grafikler, her harita birimi için bir daire grafiği çizilerek oluşturulur. Üzerinde durulması gereken bir nokta, dairesel grafiklerin negatif değer içeren vektörlere uygulanamamasıdır. Buna çözüm olarak veri düzenlemelerine (örneğin, verinin [0,1] aralığına normalize edilmesi) gidilebilir. Bu teknikte, her bileşen dairenin bir dilimi ile temsil edilir. Daire dilimleri bir parsele ait bileşenlerin değerlendirilmesi içindir;

parseller arası kıyaslama ise dairelerin boyutlandırılması yoluna gidilmektedir (Şekil 4.5). Dairesel grafiğe bakıldığında üst bölgedeki harita birimlerinin yeşil renginin diğer bölgelere göre daha fazla olduğu görülüyor. Dolayısıyla burada benzer verilerin kümelendiği söylenebilir. Şekil 4.6'ya bakıldığında bu bölgede setosa iris türlerinin olduğu anlaşılmaktadır.

Renk Düzlemleri

PCA'da olduğu gibi bu teknikte de vektör yansıtımı uygulanır ve yüksek boyutlu vektör renk düzlemine yansıtılır (Şekil 4.6). Renk düzlemlerinde birbirine yakın renk tonlarına sahip veriler benzer verilerdir. Örneğin setosa iris türüne bakıldığında bu birimlerin yeşil renk tonuna sahip olduğu görülmektedir.

Sinyal Grafiği

Sinyal grafiğinde prototip vektörler, basit çizgi grafikleri halinde gösterilirler (Şekil 4.7).

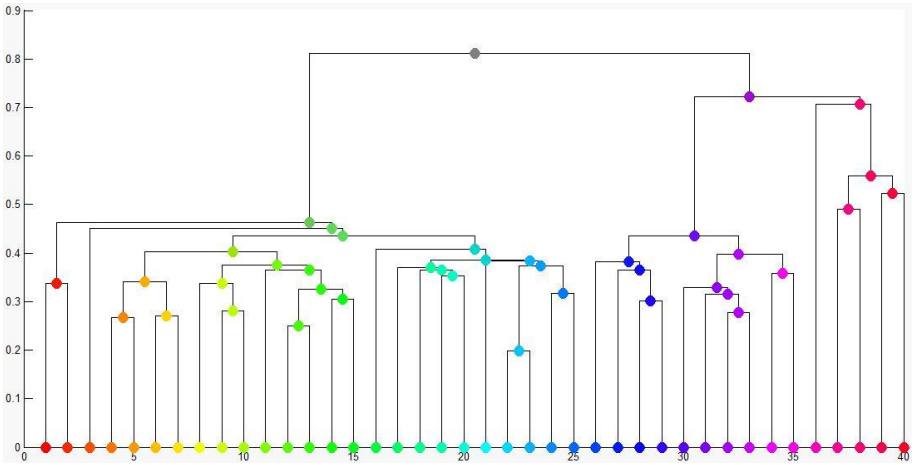
4.3 Kümelerin ve Değişkenlerin Görselleştirilmesi

Veri içi kümelenmelerin ve veri yoğunluklarının analizinde birçok metot kullanılmaktadır. Kümeleme metotları genel olarak araştırmacılar tarafından iki ana başlık altında ele alınmaktadır. Bunlar; parçalı (partioning) metotlar ve hiyerarşik metotlardır.

Hiyerarşik yöntemde, dendrogramlar kullanılarak kümeleme hiyerarşisi gözlenmektedir. Başlangıçta her veri örneği bir küme olarak değerlendirilir ve iterasyonlar ilerledikçe birbirine en yakın olan kümeler birleştirilir. Bu işlem, tüm birimleri içine alan tek küme kalana değin devam eder (Şekil 4.8).

Parçalı yöntemler ise, [7] tarafından üç ana başlık altında işlenmiştir. Bunlar; k-ortalama, örgü temelli yöntemler ve yoğunluk temelli örnekler'dir.

Bu çalışmada üzerinde durulan kümeleme metotlarının yanısıra farklı birçok teknik bulunmaktadır. Bunun için [30] ve [31]'den yararlanılabilir.



Şekil 4. 8 İris: Hiyerarşik kümeleme

4.3.1 K-Ortalama

K-ortalama, girdi vektörlerini k kümede toplayan iteratif kümeleme metotudur. Küme merkezleri kümeye ait örneklerin ortalaması (c_j) ile belirlenir. Kümeye ait hata fonksiyonu,

$$E(C) = \sum_{j=1}^{k} \sum_{x_i \in C_j} \| x_i - c_j \| \tag{4.1}$$

k ; küme sayısı

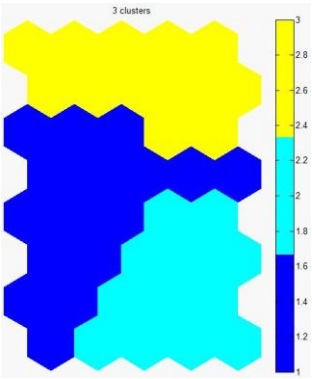
x_i ; örnekler

c_j ; kümeler

eşitliğiyle hesaplanır. K-ortalama algoritması ise şu şekildedir:

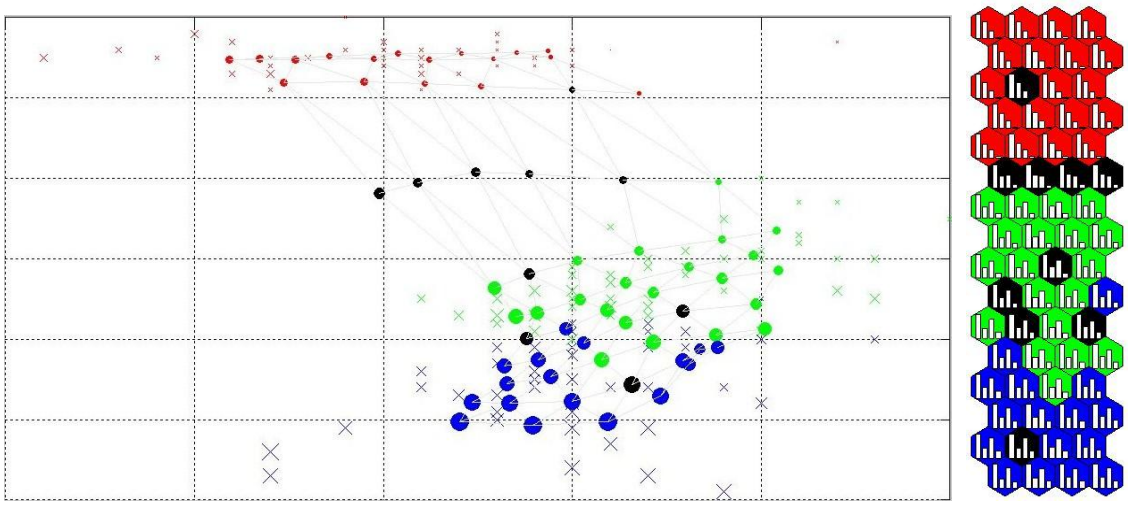
1. Küme merkezlerinin rassal olarak belirlenmesi
2. Örneklerin kümelere atanması
3. Örneklerin aritmetik ortalamlarının hesaplanması ve küme merkezlerinin belirlenmesi
4. Kümelere ait örnekler aynı kalana dek 3. işlemin tekrarlanması

K-ortalama tekniğinin bir dezavantajı küme sayısının(k) önceden atanmış olmasıdır. Bu şekildeki bir davranış en uygun küme sayısının tespitine engel olmaktadır. Şekil 4.9'da Iris ve BloodTransfusion veri kümelerine ait örnekler verilmiştir.



Şekil 4. 9 K-ortalama: iris veri kümesi

PCA da parçalı yöntemler grubuna dâhildir. Şekil 4.10'da iris veri kümesinin PCA uygulaması verilmiştir.



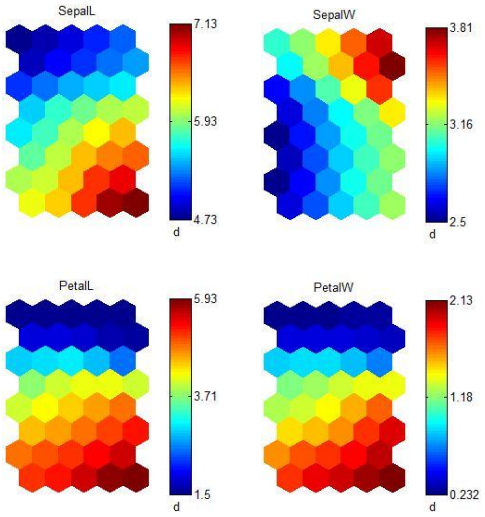
Şekil 4. 10 İris veri kümesine ait PCA grafiği. Sağ grafik: ağırlık vektörlerinin 16x4 SOM örgüsünde gösterilmesi ve sütun grafikleri (Renkler: Kırmızı(Setosa), Yeşil(Versicolor), Mavi(Virginica)

4.4 Ağırlık Vektörleri Tabanlı Görselleştirme Teknikleri

Ağırlık vektörü tabanlı görselleştirmelerde girdi verileri dikkate alınmamakta sadece ağırlık vektörleri kullanılarak çalışmalar yapılmaktadır. Bilindiği üzere, eğitim süreci sonunda ağırlık vektörleri girdi uzayındaki verilerin temsilcisi haline gelmektedir. Bundan dolayı harita, girdi verisindeki karakteristiklere sahiptir. Bileşen Düzlemleri ve U-matrisler bu gruba dâhil görselleştirme teknikleridir. Ayrıca k-means veya hiyerarşik kümeleme metotları kullanılarak ağırlık vektörlerinin görselleştirilmesi mümkün olabilmektedir.

Bileşen Düzlemleri

Bileşen düzlemleri, girdi verisinin her bir değişkenini ayrı ayrı temsil etmek için oluşturulur. Tabii ki, girdi uzayının boyutundaki artışa paralel olarak bileşen düzlemlerinin sayısında artış meydana gelir ve yüksek boyutlu verilerde görsel analizlerin yapılması zorlaşır. Bu gibi durumlarda, bölüm 1.4'te de bahsedildiği üzere, birbiriyle ilişkili ya da birbirine benzer değişkenlerin gruplanması yoluna gidilerek daha iyi analizlerin yapılması sağlanmaktadır. Şekil 4.11'de iris veri kümesine ait 4 adet bileşen düzlemi verilmektedir. SepalL, PetalL ve PetalW bileşenlerinin birbirine benzer yapıda olduğu görülmektedir. Bu bileşenlerde yüksek değerler grafiğin sağ alt bölgesine, düşük değerler ise genel olarak grafiğin yukarı kısımlarına toplanmıştır. SepalW bileşeni ise bizlere farklı bilgiler sunmaktadır; Yüksek ve düşük değerlerin birbirine yakın olması burada kümeler arası bir sınır olduğunu ifade eder. Diğer bileşenlerin aksine SepalW bileşeninde Koyu kırmızı ve koyu mavi birimlerinin birbirine yakın olduğu görülüyor.



Şekil 4. 11 İris veri kümesine ait bileşen düzlemleri. SepalL, PetalL ve PetalW bileşenlerinin birbirine benzer olduğu grafiklerden anlaşılmaktadır

U-Matris

Ağırlık vektörü tabanlı görselleştirme teknikleri içinde en çok başvurulan teknik U-matris'tir (Birleşik uzaklık matrisi) ([4], [6], [11]). U-matris iki farklı şekilde oluşturulabilmektedir. Bunlardan ilki, düğümler arası uzaklıkları göstererek (birimler arası uzaklıklar toplamının doğrudan U-matrisin oluşumu için kullanılması); ikincisi,

ortalama uzaklıkların kullanılması (birimler arası uzaklıklar toplamının nörona ait komşu sayısına bölünüp ortalama uzaklığın elde edilmesi) ile yapılmasıdır (Şekil 4. 12).

U-matrisler SOM'un üzerine inşa edilirler. SOM üzerindeki her nörona ait yükseklik değerleri eşitlik 4.3'te belirtildiği şekliyle hesaplanır.

$$d(x) = \arg\min_{j} \| x - w_j \| \tag{4.2}$$

w_j : x vektörünün j komşusuna ait vektör

$d(x)$: x girdi vektörüne en yakın harita elemanı

$$U - yükseklik(r) = \sum_{i=1}^{l} d(w(r), w(i)) \tag{4.3}$$

$d(x, y)$: SOM algoritmasındaki uzaklık

r : SOM üzerinde bir nöron

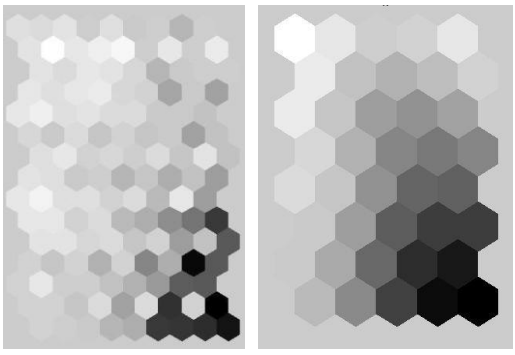
$BK(r)$: r'ye ait birimcil komşular

$w(r)$: r nöronu ile ilişkili ağırlık vektörü

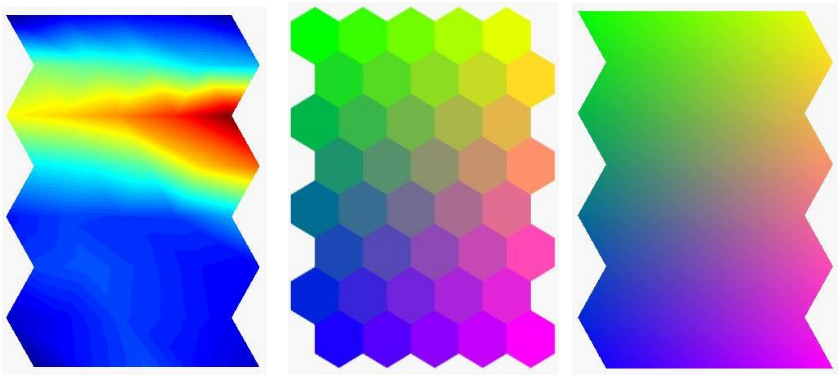
l : SOM örgüsündeki toplam nöron sayısı

$$U - ortalama(w_i) = \frac{1}{\zeta_i}.U - yükseklik(w_i) \tag{4.4}$$

ζ_i : w_i birimine ait komşu sayısı



Şekil 4. 12 U-matris görünümleri (a) Düğümler arası uzaklıklar (b) Ortalama uzaklıklar



Şekil 4. 13 İris (8x5 SOM) (a) Ara değerler eklenmiş U-matris (b) Renk düzlemi (c) Ara değerler eklenmiş renk düzlemi

Yüksek U-yükseklik değerine sahip nöronlar girdi uzayında diğer vektörlerden daha uzakta olurken; daha düşük U-yükseklik değerine sahip nöronlar ise girdi uzayında birbirine daha yakın durumdadırlar. Dolayısıyla, girdi uzayında diğer nöronlardan daha uzakta bulunan vektörlerin "gürültü" olması olasıdır. Bu birimler, genellikle U-matrisi üzerindeki "sınır" olarak nitelenen bölgelerde yer alır. Bir başka ifade ile bu bölgeler e*nterpolasyon* bölgeleri olarak adlandırılır.

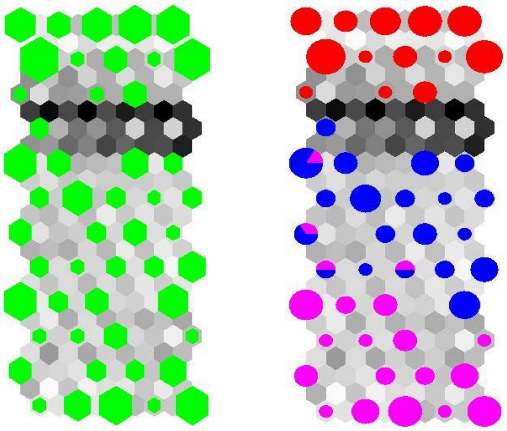
U-matris ile yapılmak istenen, haritanın hangi bölgelerinde verilerin yoğunlaşma gösterdiği ve benzer verilerin ortaya koyulmasıdır. Bunun neticesinde kümeleme çalışmalarında da kullanılmaktadır. U-matris, SOM ile kaç adet küme yapısının oluştuğunu ortaya koyan bir anahtardır denilebilir.

4.5 Veri Örnekleri Tabanlı Görselleştirme Teknikleri

SOM haritası ile veri kümesi arası ilişkilerin gösterilmesi bu bölümde ele alınacaktır. Bu bölümde ele alınacak metotlar şunlardır; Hit histogram, P-matris, U*-matris.

Hit Histogram

Hit histogramları, her harita birimin kaç kez kazanan nöron olduğuna bağlı olarak girdi verisinin dağılımını sunar. Yalnız, harita birimi sayısının girdi vektörlerinin sayısından fazla olduğu durumlarda her veri örneğinin bir harita birimine karşılık gelmesi söz konusu olabilir. Bundan dolayı, bu tür hit histogramlarında küme sınırlarının belirlenmesi zor olabilmektedir. Şekil 4.14 iris veri kümesine ait hit histogramlarını göstermektedir. Sol grafik genel olarak verinin hit haritasını, sağ grafik ise kümelere özgü hit histogramları ayrı ayrı göstermektedir.



Şekil 4. 14 İris veri kümesine ait hit histogramları (a) Tüm örnekler (b) Kırmızı: setosa, mavi: versicolor, mor: virginica

P-Matris

SOM haritalarındaki veri-bazlı bir diğer görselleştirme tekniği P-matristir ve bir haritanın öznitelik uzayının yoğunluğunu betimlemeye yarar [7]. P-matris U-matrise benzer bir çalışma prensibine sahiptir. Fakat bu kez yükseklik, U-matriste kullanılan yerel uzaklıklar yerine veri uzayındaki veri yoğunluğuna bağlı olarak hesaplanmaktadır [3]. r nöronuna ait P-yükseklik değeri şu şekilde hesaplanır;

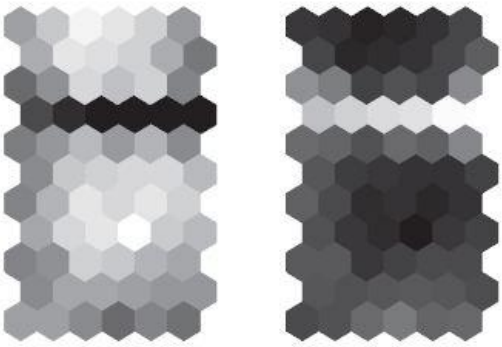
$$P - yükseklik(r) = p(w(r), X) \tag{4.5}$$

$w(r)$: İlişkili ağırlık vektörü

$p(x, X)$: X veri uzayının x noktasındaki deneysel yoğunluk tahmini

P-matris için, SOM üzerindeki veri yoğunluğunun topoloji korumalı olarak yansıtılması denilebilir. P-matrislerin özellikleri [4] tarafından şu şekilde tanımlanmaktadır;

- ❖Kazanan nöronların konumları girdi uzayının topolojisini yansıtır.
- ❖Yüksek P-yükseklik değerleri veri uzayının yoğun bölgelerini göstermektedir.
- ❖Düşük P-yükseklik değerlerine sahip birimler veri uzayında yalnız kalmış birimlerdir.
- ❖Girdi uzayındaki gürültüler P-matris üzerindeki *bacalarda* ortaya çıkmaktadır.
- ❖P-matris üzerindeki *hendekler* küme sınırlarını gösterir.
- ❖*Platolar* ise küme merkezlerini işaret eder.



Şekil 4. 15 İris, 11x6 SOM. (a) P-matris (b) U*-matris

Genel olarak P-matrise ait özellikler U-matris özelliklerinin tersidir. U-matrisler veriler arası öklid uzaklığına dayanırken, P-matrisler ise veriler arası yoğunluğa bağlı olarak çalışır. Aşağıda, yoğunluk tahminin nasıl hesaplandığı verilmektedir.

Yoğunluk tahmini

Yoğunluk tahmini, eldeki verinin olasılık yoğunluk fonksiyonunun doğru şekilde tahmini sonucu oluşturulması anlamına gelmektedir. Veri yoğunluğu bugüne değin farklı yollarla hesaplanagelmiştir. Silvermann [13], X rastgele sayıdaki veri noktası ve f , X'e ait olasılık yoğunluk fonksiyonu olduğu varsayıldığında;

$$P(a < X < b) = \int_a^b f(d)d_x \text{ ,(} a < b \text{ olmak koşuluyla)} \quad (4.6)$$

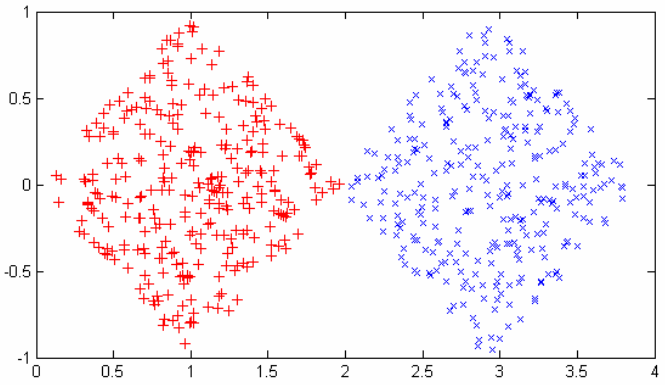
f yoğunluk fonksiyonuna bağlı X dağılımı eşitlik 4.6'daki gibi tanımlamıştır.

Ultsch [4], yoğunluk tahminini, belli yarıçaptaki bir hiperkürenin içerisinde kalan noktaların o hiperkürenin hacmine bölünmesi ile elde edildiğini belirtmiştir. Fakat, hacim hesaplaması yüksek-boyutlu veriler için sorun oluşturmaktadır. Hiperkürenin çapı sabit tutulduğunda, yoğunluk hiperkürenin içerisinde kalan veri noktalarının sayısı ile doğru orantılı olur.

U-Matris*

U-matris kullanımları verinin kümelenmesi için fevkalade yararlı olmaktadır. Fakat, bazı kümeleme çalışmalarında bu durum olumsuz sonuçlar doğurabilmektedir [3]. Bunun nedeni de küme içi yoğunluk kavramının, kümeler arası uzaklık kavramı ile aynı şekilde değerlendirilmesidir. Ultsch [3], seçtiği bir veri kümesine hem hiyerarşik kümeleme algoritmasını hem de U-matris yöntemini sınamış, sonuç olarak da hiyerarşik

kümeleme algoritması mevcut kümeleri ortaya çıkarırken, U-matris yönteminin kümeleri ortaya çıkarmada başarısız olduğu gözlenmiştir (Şekil 4.16). Bunun üzerine, U-matris görselleştirmelerini iyileştirmek amacıyla veri kümesindeki veri yoğunluğunu dikkate alan P-matris yöntemini U-matris ile birleştirilerek yeni bir metot olan U*-matris yöntemi ortaya koyulmuştur. U*-matris yöntemi hiyerarşik kümeleme algoritması ile bağdaşan bir metottur ve veri kümesi içindeki küme yapıları bu yöntem ile elde edilebilir.



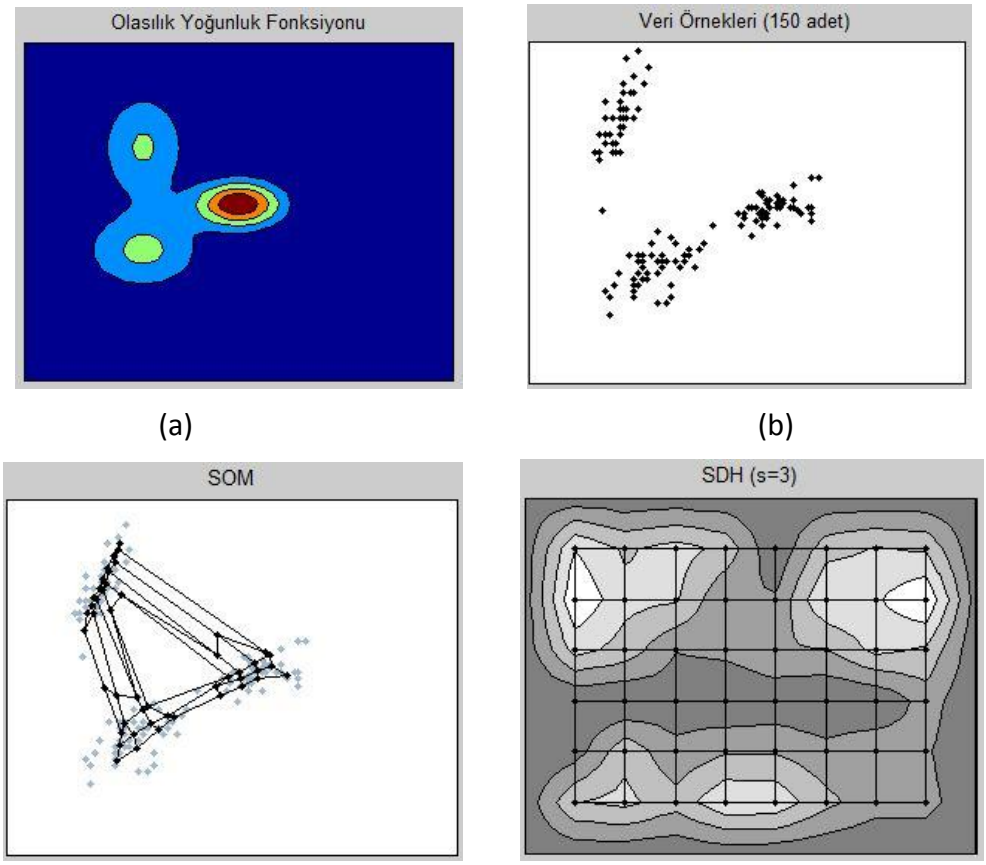
Şekil 4. 16 TwoDiamond veri kümesi [3]

İşlenmiş Veri Histogramları (SDH)

Pampalk [27] tarafından geliştirilmiş olan bir metottur. SDH, veri kümesinde barınan kümelenmeleri olasılık dağılım tahminlerine göre görselleştirir. Bu işlem gerçekleştirilirken SOM bir işlenmiş veri histogramı olarak kullanılır. Veri uzayındaki küme merkezleri ağırlık vektörleri tarafından belirlenmekte ve aynı şekilde küme çaplarının büyüklüğü vektörler arası uzaklığa göre şekillenmektedir.

SDH temelde hit histogramlarına dayanır. Fakat harita birimlerinin hit değerleri, hit histogramlarındakinden farklı bir şekilde hesaplanır. Her veri örneği için harita üzerindeki kazanan nöronlar BMU, ikinci BMU, üçüncü BMU vb. şeklinde belirlenir. (Bu işlemin nasıl yapıldığı daha önceki bölümlerde anlatılmıştı). Belirlenen s değerine göre hit değerleri oluşturulur ve harita birimlerinin almış olduğu değerler atanır. s, "yayılma derecesi", veri örneklerinin harita üzerindeki BMU izlerini algılamaktadır. Bunun için de SDH yönteminde, hit histogramlarındaki gibi sınır değerleri problemi yaşanmaz.

SDH tekniğindeki topolojik bozulmalar da kolaylıkla burada görülebilmektedir. s-BMU'nun harita üzerinde yanyana gelmesi beklenir. Eğer bir veriye ait BMU'lar harita üzerinde komşu nöronlar üzerinde oluşmuyorsa burada topolojik bozulma var demektir.



Şekil 4. 17 İris veri kümesi (a) Olasılık yoğunluk fonksiyonu (b) Veri dağılımı (c) SOM vektörleri ve verinin birlikte gösterimi (d) SDH [16]

Şekil 4.17'de SDH tekniği uygulanan iris veri kümesinin 3 merkezde toplandığı görülüyor.

BÖLÜM 5

SONUÇ VE ÖNERİLER

5.1 Genel Bakış

Bu bölümde, SOM görselleştirme tekniklerinin ve kümeleme yöntemlerinin Uygulama veri kümesi üzerinde denenmesi ve bu çalışmadan elde edilen sonuçlar ele alınmıştır.

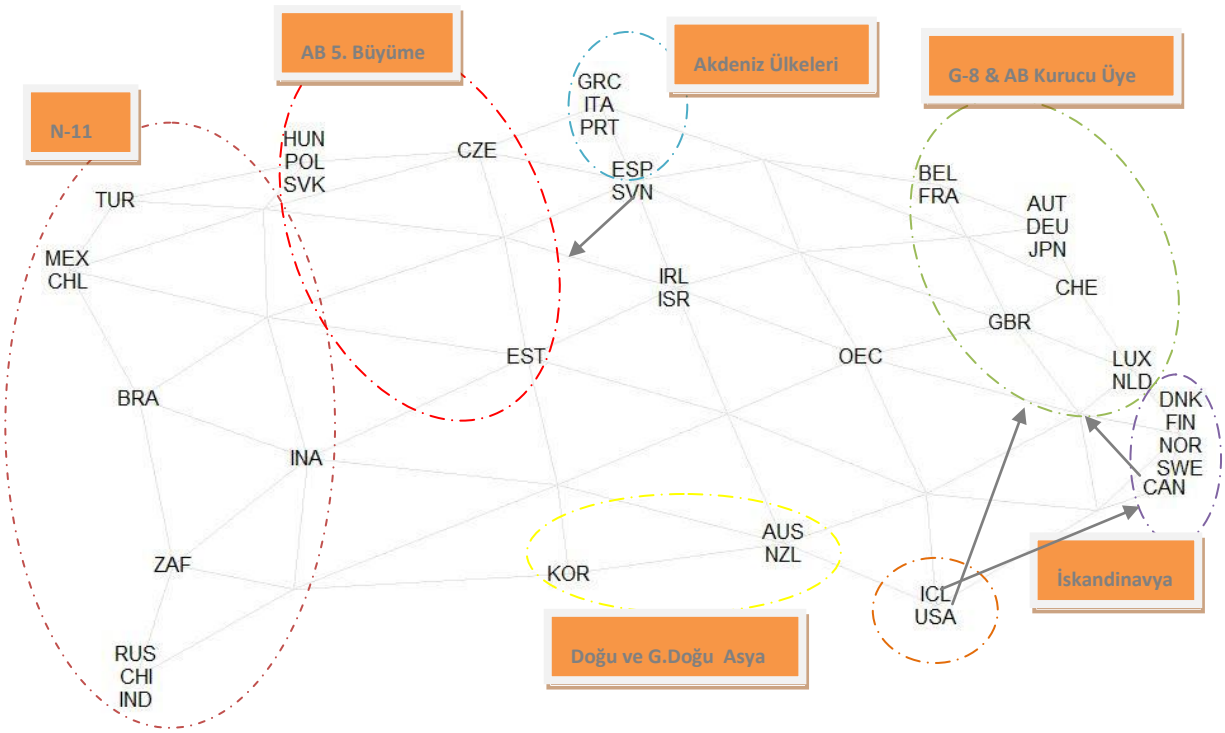
Bu esnada, ön bilgi olarak sunulan ülkelerin farklı alanlardaki konumları (AB üyesi ülkeler, G-8 ülkeleri, Akdeniz ülkeleri, BRIC ve N11 ülkeleri gibi) ile SOM uygulamasının sonuçları karşılaştırılmıştır. Karşılaştırma işlemi yapılırken 4. bölümde verilen görselleştirme teknikleri burada kullanılmıştır. Aşağıda, veri kümesine ait u-matris gösterimleri, etiketlemeler, renk düzlemleri, hit histogramları, bileşen düzlemleri verilmiştir. Bunun yanısıra, kıyaslamalar yapabilmek adına veri kümesine k-ortalama ve PCA teknikleri uygulanmıştır. Uygulama veri kümesine ait nicemleme gürültüsü ve topolojik hata değerleri (SOM kalite metrikleri) ne derece kaliteli bir SOM uygulaması olduğunu görmek adına hesaplanmıştır.

Veri kümesinin derinlemesine analizi için kategorisel temelli (eğitim, bilim ve teknoloji, yaşam kalitesi vb.) yapılan çalışmalar verilmiştir. Bu bilgiler ışında ülkelerin, örneğin eğitim alanında, SOM üzerinde aldıkları konumlar ve ülkeler arasındaki gelişmişlik seviyeleri irdelenmiştir.

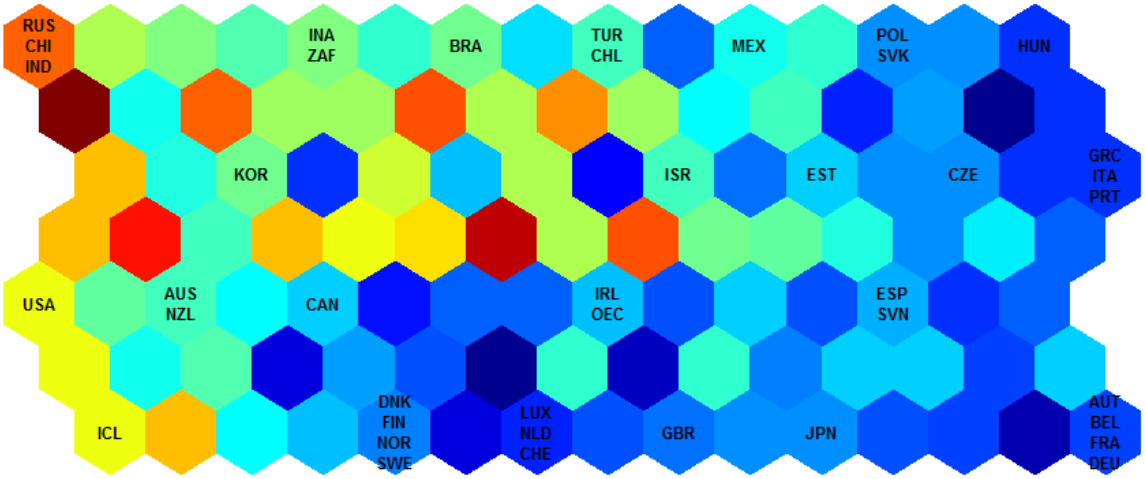
5.2 Uygulama Analizi

PCA Analizi

OECD veri kümesine PCA tekniği uygulandığında veri örneklerinin harita üzerindeki dağılımı Şekil 5.1'deki gibidir. Önceki bölümlerde ön bilgi olarak verilen ülkeler arası bazı birlik ve gruplanmaların harita üzerinde ortaya çıktığı görülmektedir. Haritanın sol bölgesinde ilerleyen yıllarda dünya siyasetinde söz sahibi olacak olan ülkelerin toplandığı görülmektedir. Başlangıçta biz bu ülkeleri N-11 grubu içerisinde vermiştik. Bu bölgenin hemen sağ üst bölgesinde AB'ye beşinci büyümede dahil olan ülkelerin olduğu görülmektedir. Bu bölgenin sağ tarafında ise genel olarak Akdeniz ülkeleri olarak tanımlanan ülkelerin öbeklendiği, haritanın en sağ bölgesinde ise G-8 ve AB kurucu üyeleri ile İskandinavya bölgesi ülkeleri olan Danimarka, Norveç, Finlandiya, İsveç ve İzlanda bulunmaktadır. Haritanın alt bölgesinde ise Doğu ve Uzak Doğu ülkeleri olarak nitelenen Güney Kore, Yeni Zelanda ve Avusturalya bulunmaktadır. Haritanın Orta bölgesinde diğer kümelere uzak kalan İsrail, İrlanda bulunmaktadır. Şekilde bulunan ok işaretleri tez kapsamında verilen ön bilgilere göre ülkelerin ait olduğu kümeleri işaret etmektedir.



Şekil 5. 1 OECD veri kümesi: Temel bileşen analizi (PCA)



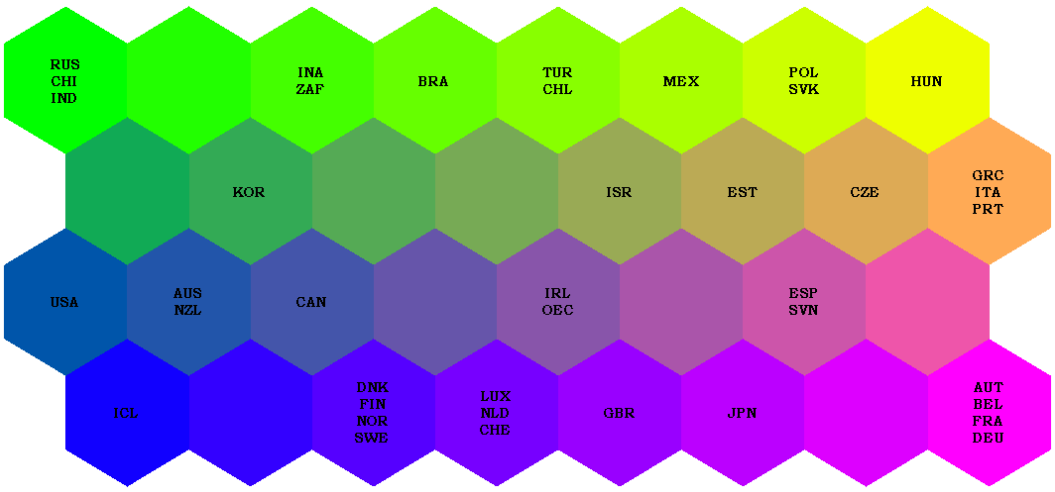
Şekil 5. 2 OECD veri kümesi: U-matris ve etiketler (4x8 SOM)

U-matris

OECD veri kümesinin U-matris uygulamasına bakıldığında (Şekil 5.2) haritanın sol-orta bölgesinde bir sınır bölgesi (koyu kırmızı tonları) oluştuğu görülmektedir. Gelişmekte olan ülkeler ile gelişmiş ülkelerin net bir şekilde birbirinden ayrıldığı açıktır. Haritanın sağ-alt ve sağ-üst bölgelerinde kümelerin oluştuğunu söyleyebiliriz. Yine haritanın orta-alt bölgesinde kalan ülkeler de bir küme oluşturmuştur. PCA uygulamasında ortaya çıkan gruplaşmalar U-matris uygulaması ile benzer şekildedir.

Renk Haritası

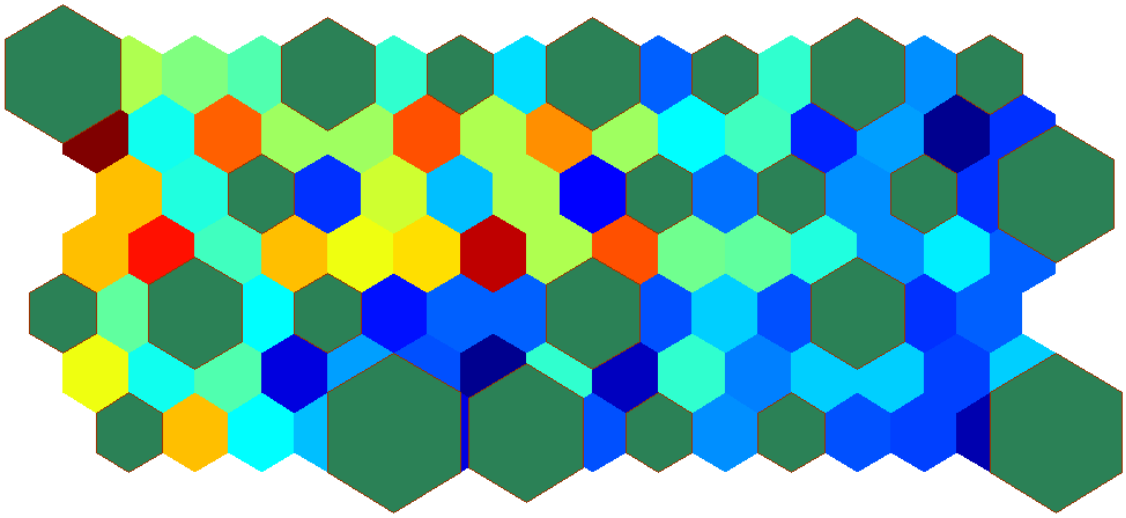
Şekil 5.3'de yukarıdaki haritaya ait renk haritası verilmiştir. Renk haritası ile hangi ülkelerin birbirine benzer yapıda olduğunu daha rahat görülebilir.



Şekil 5. 3 OECD veri kümesi: Renk haritası

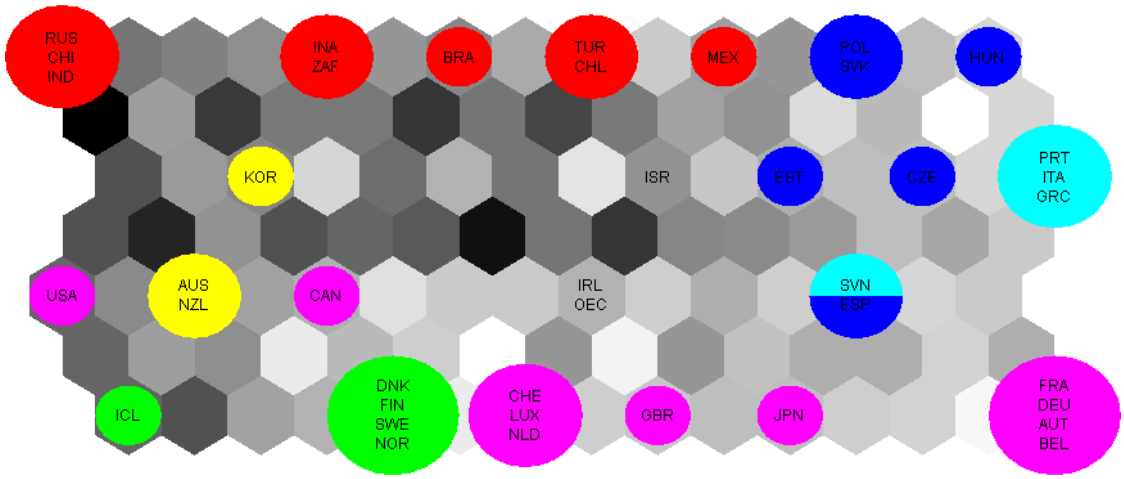
Haritanın yukarı bölgesindeki ülkelerin genel olarak yeşil renk tonuna hâkim olduğu görülmektedir. Aşağı bölgenin sol ve orta kısımlarında ise genel olarak mavi renk hâkimdir. Sağ-alt bölgede ise farklı bir renk tonu belirginleşmiştir. Bu sonuçlar, daha önce giriş bölümünde değinilen birlik ülkelerinin genel olarak harita üzerinde yakın olduğunu göstermektedir.

Hit Histogramları



Şekil 5. 4 OECD veri kümesi: hit histogramı (4x8 SOM)

OECD veri kümesine ait hit haritasına bakıldığında kazanan nöronların haritanın yukarı, aşağı-sağ ve aşağı-orta bölümlerinde yoğunlaştığı görülmektedir. Şekil 5.5'te OECD veri kümesine ait hit histogramı kategorisel temelli yeniden verilmiştir.



Şekil 5. 5 OECD veri kümesi, grup olarak hit histogramları

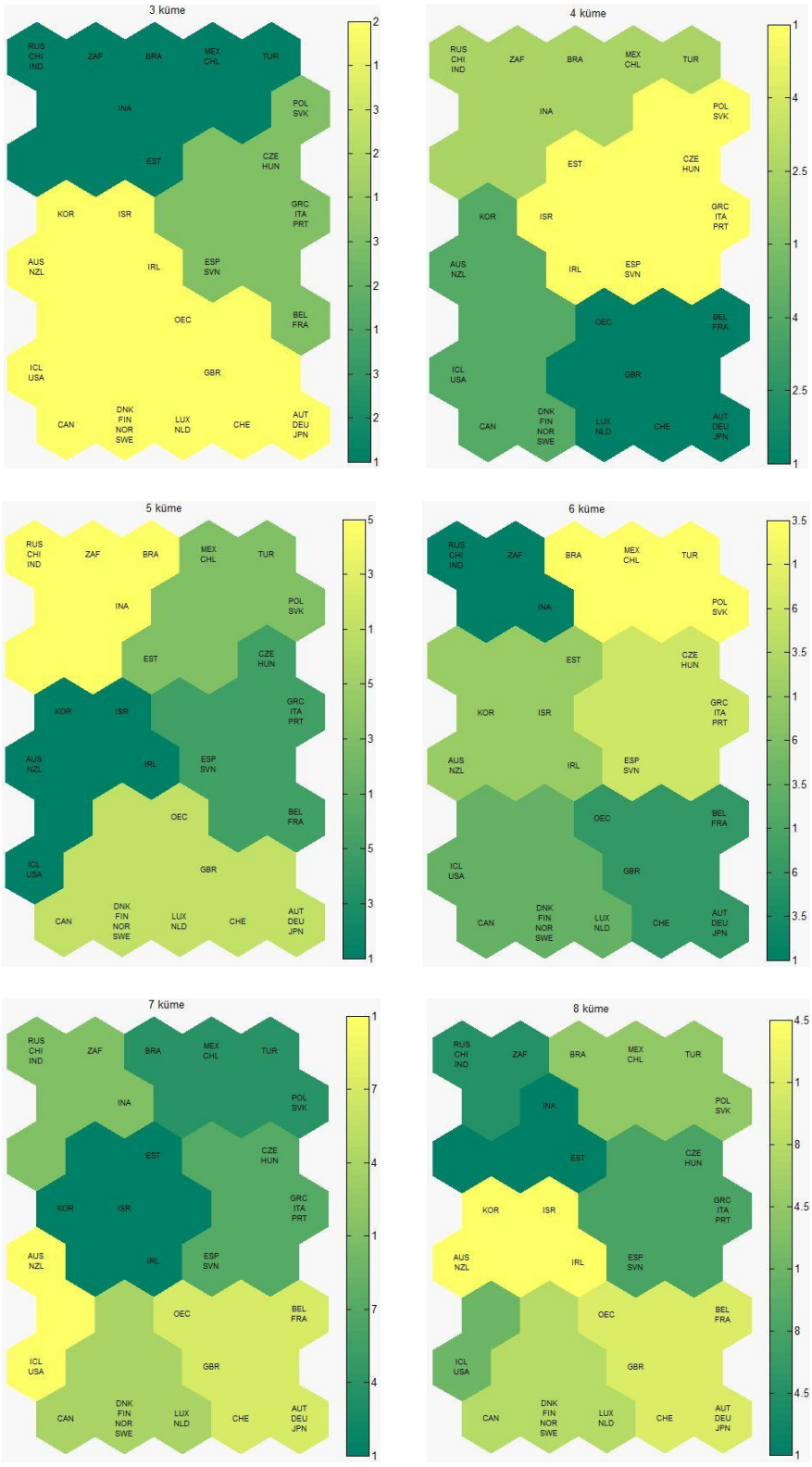
Çizelge 5. 1 Hit histogramındaki (Şekil 5.5) renk bilgileri

Renk	**Açıklama**
Kırmızı	N-11 Ülkeleri
Mavi	AB-Beşinci Büyüme
Turkuaz	Akdeniz Ülkeleri
Mor	G-8 ve AB kurucu ülkeleri
Yeşil	İskandinavya Ülkeleri
Sarı	Uzak Doğu Ülkeleri

Renklendirme işlemi bu çalışmada verilen ön bilgilere göre yapılmıştır. Aynı kümeye ait ülkelere ait hit histogramlarının yanyana olduğu görülmektedir. Renk tonlarına eşliğinde haritaya bakıldığında gruplanmaların daha net bir şekilde görüldüğü açıktır.

K-ortalama

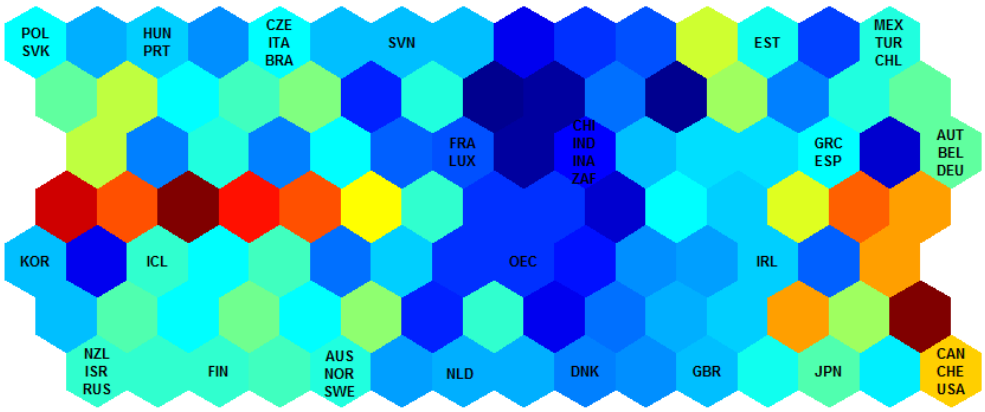
Şekil 5.6'da OECD veri kümesinin k-ortalama metoduna ait sonuçları verilmiştir. Küme sayısı 3 ilâ 8 arasında olan yapılara bakıldığında siyasi, ekonomik, sosyal vb. açılardan birbirine yakın ilişki içerisinde olan ülkelerin yine aynı kümeler içerisinde yer aldığı görülmektedir. K-ortalamanın sonuçları PCA ve U-matris yöntemlerinin sonuçları ile büyük yakınlık içerisindedir. K-ortalamanın eksikliklerinden birisi küme sayısının önceden atanması olduğu için, biz burada farklı durumlardaki k-ortalama sonuçlarını vermeyi daha uygun bulduk. Şekil 5.6'da 8x5 SOM kullanılarak oluşturulmuş sırasıyla 3, 4, 5, 6, 7 ve 8 küme sayısına sahip haritalar verilmiştir.



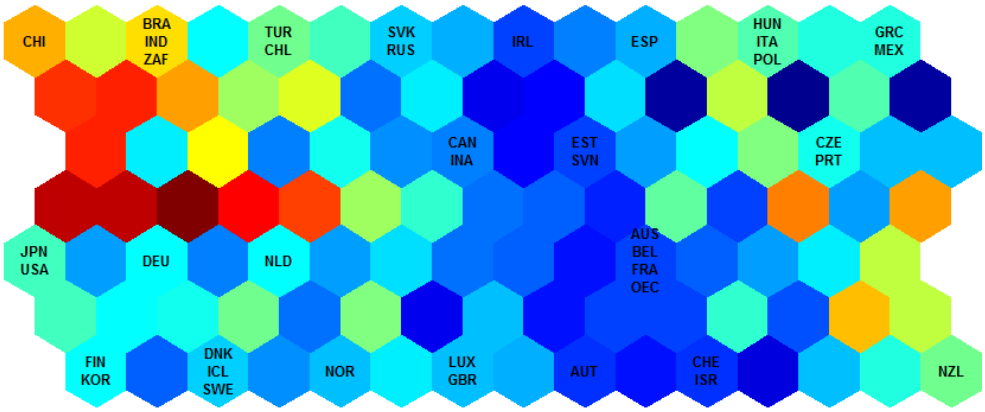
Şekil 5. 6 OECD veri kümesine k-ortalamanın uygulanması, 8x5 SOM: Sırasıyla küme sayısı 3, 4, 5, 6, 7, 8 olarak alınmıştır

Kategorisel Temelli Analiz

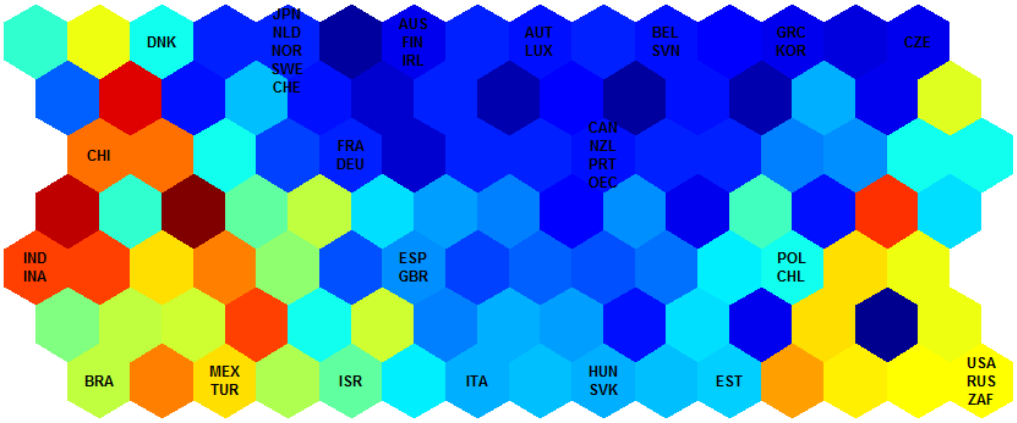
OECD veri kümesini kategorisel temelli olarak incelemek de mümkündür. Bu çalışmadaki değişkenler daha önce açıklandığı üzere 10 farklı kategoride toplanmıştır. İstenilen kategori seçilerek kategorisel temelli olarak ülkeler arası değerlendirmelerin yapılması SOM haritalarında mümkündür. Örneğin, eğitim kategorisinde ülkelerin konumu Şekil 5.7'de verilmiştir. Şekle bakıldığında haritanın sol-üst ve sol-alt bölümlerinin birbirinden ayrıldığı görülüyor. Yine eğitimsel temelli en ileri seviyelerdeki ülkeler olan Kanada, ABD, İsviçre ve Japonya'nın haritanın sağ alt köşesinde diğer ülkelerden bağımsız bir şekilde yerlerini aldığı görülmektedir. Eğitim kategorisinin yanısıra, Şekil 5.8'de bilim ve teknoloji, Şekil 5.9'de yaşam kalitesi ve Şekil 5.10'da üretim ve gelirler kategorisinin sonuçları sunulmuştur.



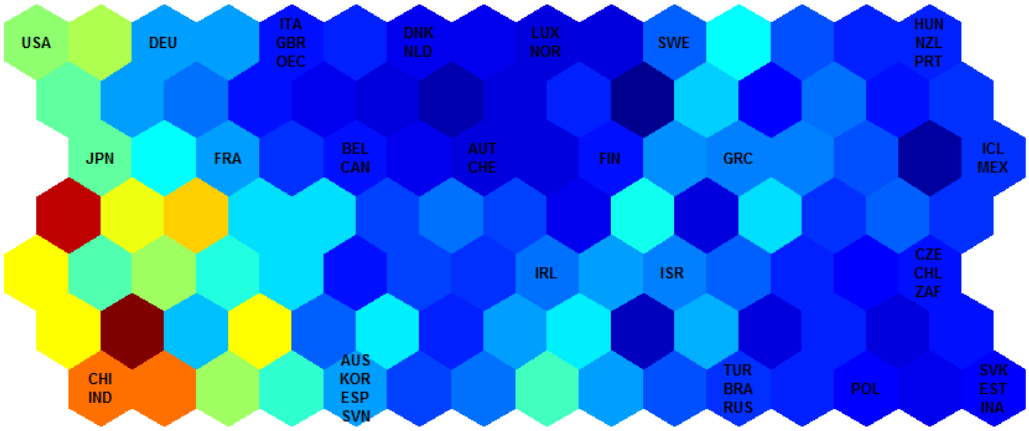
Şekil 5. 7 Eğitim kategorisi



Şekil 5. 8 Bilim ve teknoloji kategorisi



Şekil 5. 9 Yaşam kalitesi kategorisi

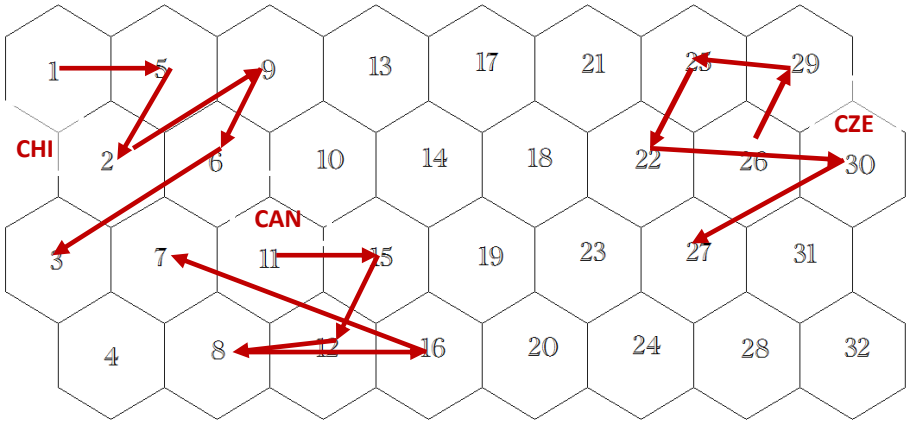


Şekil 5. 10 Üretim ve gelirler kategorisi

Kalite Metrikleri

Bilindiği üzere, SOM tekniğini diğer tekniklerden ayıran belki de en önemli farkı yüksek boyutlu veri kümelerinden daha düşük boyutlu sistemlere veri aktarımında topolojik korunmayı sağlamasıdır. Topolojik korunmanın ne derece başarılı bir şekilde gerçekleştirildiğini anlamak için verilere ait kazanan nöronlara bakılabilir. Şöyle ki; bir girdi verisine ait birincil, ikincil, vb. kazanan nöronların harita üzerinde birbirine komşu olması beklenir. Bu şekilde bir durum söz konusu olduğunda o girdi verisine ait topolojik korunmanın olduğunu söyleriz, aksi durumda girdi verisinin harita üzerine aktarım sırasında topolojik bozulma gerçekleşmiştir. Şekil 5.11'de veri kümesinden seçilen üç ülkenin (Çin, İsviçre ve Kanada) kazanan nöronları ayrı ayrı olarak gösterilmiştir. "Ok"lar sırasıyla verilerin birincil, ikincil,...,altıncıl kazanan nöronlarını takip etmektedir. Şekil incelendiğinde ülkelere ait kazanan nöronların birbirine komşu olduğu ve dolayısıyla çalışmada topolojik korunmanın sağlandığı söylenebilir.

Çalışmanın kalitesi adına topolojik korunmanın sağlanması önemlidir. Harita üzerindeki rakamlar harita indislerini göstermektedir.



Şekil 5. 11 Topolojik bozulma/korunma (Harita üzerindeki sayılar harita indislerini göstermektedir)

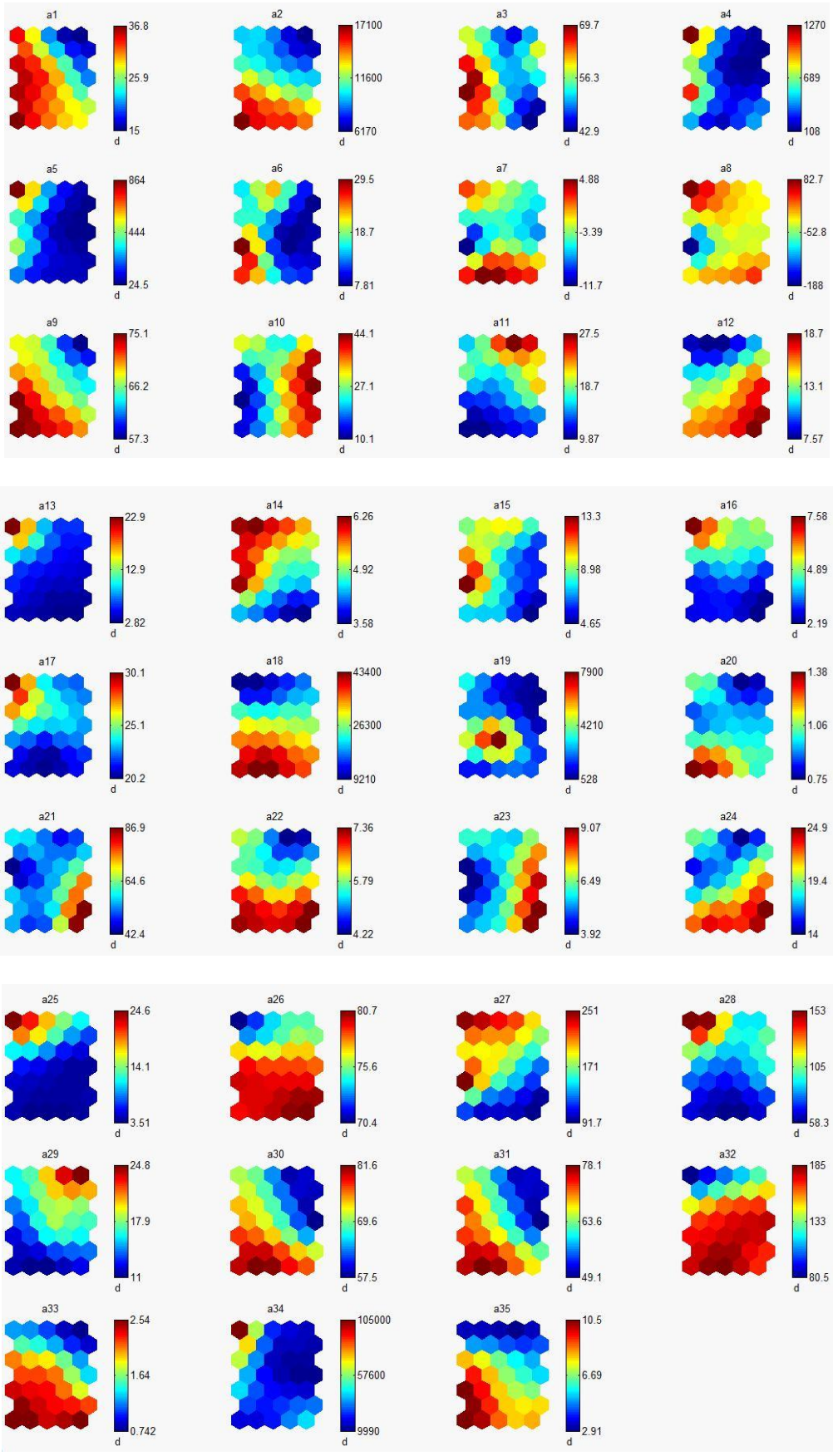
Çizelge 5.2'de SOM'un kalite metrikleri olan nicemleme ve topolojik hatalarının OECD veri kümesi için sonuçları gösterilmektedir (Nicemleme hatası ve topolojik hataların nasıl hesaplandığı bölüm 3.4'te açıklanmıştı). OECD veri kümesinin sonuçlarına bakıldığında, topolojik bozulma sıfır, nicemleme hatası ise 3.712 olarak neticelenmiştir. Topolojik bozulmanın sıfır ya da sıfıra yakın olması SOM'un kalitesini ortaya koymaktadır.

Çizelge 5. 2 SOM kalite metrikleri

	Nicemleme Hatası($N_{gürültü}$)	**Topolojik Bozulma($T_{gürültü}$)**
Uygulama Veri Kümesi	3.712	0

Bileşen Düzlemleri

Şekil 5.12'de OECD veri kümesine ait 35 bileşen düzlemi verilmiştir. Her bir değişkenin veri kümesinin SOM üzerinde şekillenmesinde nasıl bir katkısı olduğu bu şekillerden çıkartılabilir. Birbirine benzer olan ve gruplama yoluyla birleştirilebilecek olan değişkenlerin hangileri olacağına karar vermek için yine bu şekillerden yararlanılabilir.



Şekil 5. 12 Bileşen Düzlemleri

Örneğin, 1, 2, 9, 30, 31 ve 33 numaralı değişkenler birbirine benzer niteliktedirler. Bu değişkenlerin sağ üst köşelerinde düşük değerlerin (mavi ile temsil edilen bölge), sol alt köşede ise yüksek değerlerin (kırmızı ile temsil edilen bölge) bulunduğu görülmektedir. Özellikle bileşen azaltmalarında bu grafiklerden yararlanılmaktadır.

5.3 Sonuç

Bu çalışmada, OECD ülkelerinin verileri kullanılarak seçilen 40 ülke ve 35 değişkenden oluşan veri kümesine SOM algoritması ve görselleştirme metotları uygulanmıştır.

İkinci bölümde, oluşturulmuş olan OECD veri kümesine ait değişken bilgileri ve bu değişkenlerin kategorileştirilmiş hali verilmiştir.

Üçüncü bölümde, SOM algoritması, vektör nicemleme ve vektör yansıtım teknikleri, SOM metodunun detaylı analizi ve özdüzenleyici haritalarda kullanılan kalite metrikleri anlatılmıştır.

Dördüncü bölümde, SOM görselleştirme metotlarının detaylı analizi ve bunların örnek veri kümeleri üzerindeki uygulamaları sunulmuştur. Burada, SOM vektörlerinin görselleştirilmesi, görselleştirmede SOM örgüsünden yararlanma, küme ve değişkenlerin görselleştirilmesi, ağırlık vektörü tabanlı görselleştirmeler ve veri örnekleri tabanlı görselleştirme alt başlıkları ile konu detaylandırılmıştır.

Son olarak beşinci bölümde ise, dördüncü bölümde değinilen tekniklerin uygulama veri kümesi üzerindeki sonuçları verilmiş ve bu sonuçların değerlendirmesi yapılmıştır.

Sonuçlar incelendiğinde, SOM/U-matris haritasının üst ve alt bölgesinin yüksek uzaklık değerleri ile ayrıldığı, bir sınır bölgesi oluştuğu görülmüştür. Burada, üst bölgelerde genel olarak mevcut ülkeler içerisinde daha alt gelişmişlik seviyesinde olan ülkelerin bulunduğu, haritanın alt bölgesinde ise gelişmiş ülkelerin yer aldığı ortaya çıkmıştır. Harita daha ayrıntılı incelendiğinde bu iki bölgenin de kendi içinde alt gruplara parçalandığı tez kapsamında verilen ön bilgilere (ülkelerin çeşitli gruplara tasnifi) göre şekillendiği görülmüştür.

Yine beşinci bölümde, bazı kümeleme tekniklerinin de uygulama veri kümesi üzerindeki uygulamaları verilmiş ve bu uygulamaların sonuçlarının kıyaslanması yoluna gidilmiştir.

Kümeleme sonuçlarına bakıldığında, bağdaşık ülkelerin aynı bölgeler içerisinde kaldığı görülmektedir.

Çalışma bir de kategorisel temelli olarak ele alınmış, mevcut 10 kategoriden bazıları (eğitim, yaşam kalitesi, bilim ve teknoloji ve üretim ve gelirler) seçilerek bu kategorilerde ülkelerin durumları incelenmiştir. Kategorisel temelli incelemeler, özellikli bir alanda yorum ve değerlendirmelerde bulunmamızı sağlar ve o alanda ülkelerin diğer ülkelere göre konumunu verir. Bu açıdan kategorisel temelli görselleştirmeler yararlı olmaktadır.

Görselleştirme kavramına giren tüm kavramlar vasıtasıyla, oluşturulan veri kümesinin sorgulanması sonucunda ülkelere ait gerçek dünya kategorileri ile çalışmanın kıyaslanması mümkün kılınmıştır. Gerek coğrafi açıdan gerek çeşitli alanlarda veriler incelendiğinde sonuçların gerçek dünya verilerine uygun olduğu sonucu elde edilmiştir. SOM/U-matris haritasına bakıldığında, yenidünya düzeninde ön sıralara doğru ilerlemekte olan Rusya, Çin, Hindistan, Güney Afrika, Brezilya, Endonezya, Türkiye ve Meksika gibi ülkelerin (N-11 ülkeleri) haritanın üst bölgesinde yer aldıkları görülmektedir. İskandinavya bölgesi ülkeleri olan Norveç, Danimarka, İsveç, İzlanda, Finlandiya gibi ülkelerin de haritanın alt sol bölgesinde birbirine yakın olarak yer almaktadır. Haritanın sağ alt bölgesinde ise genel olarak G-8 ve AB kurucu ülkeleri bulunmaktadır. Akdeniz ülkeleri ve AB'ye sonradan dâhil olan ülkeler ise genel olarak haritanın sağ üst bölgesinde yoğunlaşmıştır.

Bu çalışmada, daha önceki çalışmaların aksine ülkelerin tek bir açıdan değil; eğitim, enerji, çevre, küreselleşme, işgücü, nüfus, fiyatlar, üretim ve tüketim, kamu maliyesi, yaşam kalitesi ve bilim ve teknoloji gibi birçok alanı içine alan geniş bir yelpazeden değerlendirilmesine olanak sağlanmıştır.

Bilhassa 4. bölümde değinilen görselleştirme teknikleri ve uygulamalar, özdüzenleyici ağlar konusunda yeterince bilgi sahibi olmayan insanların bile bu konunun bilincine varmalarını ve yapılan uygulamaları anlamalarını sağlayacak seviyededir. SOM görselleştirmeleri veri hakkında bilgi sahibi olunması açısından önemlidir.

Bu tez yine göstermiştir ki, herhangi bir istatistiksel test kabulüne dayanmayan SOM metotu yüksek-boyutlu verilerle başa çıkmada önemli ve dikkate değer bir başarıya

sahiptir. Kümeleme açısından bakıldığında ise SOM'un bu alandaki esnekliği ve yararlılığı ortaya koyulmuştur. SOM metotu yüksek boyutlu verilerdeki kümelenmelerin görünürlüğünü sağlamaktadır. Bu işlemi gerçekleştirirken de topolojik korunmaların en üst seviyede garanti altına alındığı yine bu çalışmada görülmüştür. Bizim çalışmamızdaki bir başka önemli nokta da oluşturulan veri kümesinin herhangi bir küme etiketine sahip olmamasıdır. Burada SOM metotunun eğiticisiz öğrenme yolu ile veri analizi ve görselleştirme konusundaki başarısı sergilenmiştir.

KAYNAKLAR

[1] Kohonen, T. (2001), Self-Organizing Maps, Springer Series in Information Sciences. Springer 3^{rd} edition, Berlin, Heidelberg.

[2] Haykin, S. (2009), Neural Networks and Learning Machines, 3^{rd} edition. Pearson International Edition, Chapter 9, Self-Organizing Maps. New Jersey.

[3] Ultsch, A.,(2003a),"U*Matrix: A Tool to Visualize Clusters in High Dimensional"

[4] Ultsch, A., (2003b), “Maps for the Visualization of High-dimensional Data Spaces”.

[5] Pölzlbauer, G., (2004), Application of Self-organizing Maps to a Political Dataset, Wien.

[6] Pölzlbauer, G., Dittenbach, M. ve Rauber, A., (2005), “Gradient Visualization of Grouped Component Planes on the Lattice”.

[7] Pölzlbauer, G., (2008), Advanced Data Exploration Methods Based on Self-organizing Maps, Phd. Thesis, Wien.

[8] Fernandez, E. A. ve Balzarini, M., (2007), “Improving Cluster Visualization in Self-Organizing Maps: Application in Gene Expression Data Analysis. Computers in Biology and Medicine”, 37(12): 1677 – 1689.

[9] Taşdemir, K. ve Merenyi, E., (2006), “Data Topology Visualization for the Self-Organizing Map”, Proc. 14^{th} European Symposium on Artificial Neural Networks, ESANN’2006, sayfalar: 125-130, Bruges, Belgium.

[10] Taşdemir, K. ve Merenyi, E., (2009), “Exploiting the Data Topology in Visualizing and Clustering of Self-Organizing Maps”, IEEE Trans. Neural Networks, 20(4): 549 – 562.

[11] Silva, S.D., Monteiro, A.M.V. ve Medeiros, J.S., (2004), “Visualization of Geospatial Data by Component Planes and U-Matrix”.

[12] Vellido, A., Lisboa, P.J.G., ve Meehan, K., (1999), “Segmentation of the On-line Shopping Market Using Neural Networks”, Expert Systems with Applications, sayfalar: 303-314.

[13] Silvermann, BW., (1986), “Density Estimation for Statistics and Data Analysis”, Monographs on Statistics and Applied Probability, Chapman and Hall, Londra.

[14] Little, R.J.A. ve Rubin, D.B., (2002), “Statistical Analysis with Missing Data”, 2^{nd} Edition: John Wiley, New York.

[15] Wang, H. ve Wang, S., (2007), “Visualization of the Critical Patterns of Missing Values in Classification Data”, LNCS 4781, sayfalar: 267-274, Springer-Verlag, Berlin, Heidelberg.

[16] Vesanto, J., Himberg, J., Alhoniemi, E. ve Parhankangas, J., (2000), “SOM Toolbox for Matlab 5”, Helsinki University of Technology.

[17] Vesanto, J. (2002), Data Exploration Process Based on the Self-Organizing Map. PhD Thesis, Helsinki University of Technology, Finland.

[18] MacQueen, J.B., (1967), “Some Methods for Classification and Analysis of Mutivariate Observations”, Proceedings of 5 th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley.

[19] Flexer, A., (1997), “Limitations of Self-organizing Maps for Vector Quantization and Multidimensional Scaling”, Viyana.

[20] Lehtimaki, P. Ve Raivio, K., (2005), “A SOM Based Approach for Visualization of GSM Network Performance Data”.

[21] Demartines, P. ve Herault, J., (1997), “Curvilinear Component Analysis: A Self-organizing Neural Network for Non-linear Mappings of Data Sets”, IEEE Transactions on Neural Networks, Vol. 8.

[22] Sammon, J. W., (1969), “A Nonlinear Mapping for Data Structure Analysis”, IEEE Transactions on Computers, 18(5): 401-409.

[23] Bishop, CM., Svensen, M. ve Williams, CKI., (1998), “GTM: Generative Topographic Mapping”, Neural Computation, 10: 215-234.

[24] Martinez, W.L. ve Martinez, A.R., (2005), Exploratory Data Analysis with MATLAB, Computer Science and Data Analysis Series, Chapman & Hall/CRC Press, USA.

[25] Zhu, M. ve Ghodsi, A., (2006), “Automatic Dimensionality Selection from the Scree-Plot via the Use of Profile Likelihood”, Computational Statistics & Data Analysis, 51: 918-930.

[26] Kohonen, T. (1998), “The Self-organizing Map”, Neurocomputing 21, Finland.

[27] Pampalk, E., Rauber, A. ve Merkl, D. (2002), "Using Smoothed Data Histograms for Cluster Visualization in Self-organizing Maps." In Proceedings of the International Conference on Artificial Neural Networks (ICANN'02), Madrid, Spain.

[28] Vikipedya, Kernel(statistics), http://en.wikipedia.org/wiki/Kernel, 20 Mayıs 2011

[29] Laboratory of Computer and Information Science, Implementation in SOMToolbox,http://www.cis.hut.fi/somtoolbox/documentation/somalg.shtml , 21 Mayıs 2011

[30] Berkhin, P. (2002), "Survey of Clustering Data Mining Techniques.", Technical Report, Accrue Software, San Rose, CA, USA.

[31] Jain, A., Murty, M. ve Flynn, P. (1999), "Data Clustering: A Review". ACM Clustering Surveys, 31(3): 264-323, 1999.

EK-A

YARDIMCI VERİ KÜMELERİ

Bu bölümde, tezde kullanılmış olan veya değinilen veri kümeleri verilmiştir.

A-1 Iris

Kaynak	R. A. Fisher
Veri Kümesi Karakteristiği	Çok değişkenli
Özellik	Reel
Örnek Sayısı	150
Değişkenler	4 tane (1. Çanak yaprak uzunluğu 2. Çanak yaprak genişliği 3. Taç yaprak uzunluğu 4. Taç yaprak genişliği)
Etiketler	3 etiket: 1. Iris Setosa 2. Iris Versicolor 3. Iris Virginica
Kayıp Değer	Yok
Açıklama	Çeşitli taç ve çanak yaprak uzunluk ve genişliklerine sahip Iris bitkilerinin hangi türe ait olduğunu bulmak

EK-B

KOD DÖKÜMÜ

Bu bölümde, tez kapsamında değinilen eğitim algoritmaların C programlama dili ile gerçeklemeleri verilmiştir.

B-1 Sıralı Eğitim Algoritması

Sıralı eğitim algoritmasının bir turluk çalışması;

```
for (j = 0; j < n; j++)
{
   bmu = -1;
   min = 100000;
   for (i = 0; i < m; i++)
   { /* Kazanan nöronu belirle */
      mesafe = 0;
      for (k = 0; k < d; k++)
      {
         fark = X[j][k] - M[i][k];
         mesafe += fark * fark;
      }
      if (mesafe < min)
      {
         min = mesafe;
         bmu = i;
      }
   }
```

```
   for (i = 0; i < m; i++) /* Güncelleme */
   {
      h = alpha * exp(U(bmu,i)/r);
      for (k = 0; k < d; k++){
         M[i][k] -= h * (M[i][k] – X[j][k]);
      }
 }
}
```

Burada Gauss dağılımına göre gerçekleme verilmektedir. Burada belirtilen değişkenler;

`X[j][k]`; i'inci girdi verisinin k'ıncı bileşeni

`M[i][k]`; Örgü elemanı j'ye ait k'ıncı bileşen

`U`; Harita birimleri arası uzaklıkların tutulduğu tablo.

`r`; Komşuluk yarıçapı

B-2 Batch Algoritması

```
for (i = 0; i < m; i++)/* ilk deger atama */
{
   vn[i] = 0;
   for (k = 0; k < d; k++){
      S[i][k] = 0;
   }
}
for (j = 0; j < n; j++)
{
   bmu = -1;
   min = 100000;
   for (i = 0; i < m; i++) /* kazanan nöronun bulunmasi */
   {
      mesafe = 0;
      for (k = 0; k < d; k++)
```

```
      {
         fark = X[j][k] - M[i][k];
         mesafe += fark * fark;
      }
      if ( mesafe < min){
         min = mesafe;
         bmu = i;
         vn[bmu]++;
      }
   }
   for (k = 0; k < d; k++){
      S[bmu][k] += X[j][k]; /* Voronoi Bolgeleri Toplami */
   }
}
for (i = 0; i < m; i++){
   for (k = 0; k < d; k++){
      M[i][k] = 0;
   }
}
for (i1 = 0; i1 < m; i1++){
   htot = 0;
   for (i2 = 0; i2 < m; i2++){
      h = exp(U[i1][i2]/r);
      for (k = 0; k < d; k++){
         M[i1][k] += h * S[i2][k];
      }
      htot += h * vn[i2];
   }
   for (k = 0; k < d; k++){
     M[i1][k] /= htot;
     }
}
```

EK-C

ÜLKELER VE VERİ GÖSTERGELERİ

C-1 Ülkelerin Listesi

Ülke Adı	Kısaltması	Ülke Adı	Kısaltması
Almanya	DEU	İsrail	ISR
Avustralya	AUS	İtalya	ITA
Avusturya	AUT	İzlanda	ICL
Belçika	BEL	Japonya	JPN
Birleşik Devletler	USA	Kanada	CAN
Brezilya	BRA	Kore (Güney)	KOR
Çek Cumhuriyeti	CZE	Lüksemburg	LUX
Çin	CHI	Macaristan	HUN
Danimarka	DNK	Meksika	MEX
Endonezya	INA	Norveç	NOR
Estonya	EST	Polonya	POL
Finlandiya	FIN	Portekiz	PRT

Fransa	FRA	Rusya Federasyonu	RUS
Güney Afrika	ZAF	Slovakya	SVK
Hindistan	IND	Slovenya	SVN
Hollanda	NLD	Şili	CHL
İrlanda	IRL	Türkiye	TUR
İngiltere	GBR	Yeni Zellanda	NZL
İspanya	ESP	Yunanistan	GRC
İsveç	SWE	OECD Ortalama	OEC
İsviçre	CHE		

C-2 Veri Göstergeleri (Değişkenler)

İmge	**Açıklama**	**Grup**
E1	Eğitimsel Kazanım	Eğitim
E2	Eğitimsel Harcama	Eğitim
E3	Yüksek Öğretim Öğrenci Sayısı	Eğitim
P1	Elektrik Üretimi	Enerji
P2	Enerji Üretimi	Enerji
P3	Yenilenebilir Enerji	Enerji
K1	Ödemeler Dengesi	Küreselleşme
K2	Uluslararası Ticaret	Küreselleşme

I1	İstihdam Oranı	İş Gücü
I2	Uzun Dönem İşsizlik	İş Gücü
I3	Serbest Meslek	İş Gücü
N1	Bağımlı Nüfus	Nüfus
F1	Tüketici Fiyat Endeksi	Fiyatlar
F2	Uzun Dönem Faiz Oranları	Fiyatlar
F3	Üretici Fiyat Endeksi	Fiyatlar
G1	Gayrısâfi Yurtiçi Hâsıla Gelişimi	Üretim/Gelir
G2	Yatırım Oranları	Üretim/Gelir
G3	Kişi Başına Düşen Milli Gelir	Üretim/Gelir
G4	Gayrısâfi Yurtiçi Hâsıla Toplam Miktarı	Üretim/Gelir
H1	Eğitim Harcamaları	Kamusal Finans
H2	İç Borçlar	Kamusal Finans
H3	Sağlık Harcamaları	Kamusal Finans
H4	Maaş Ödemeleri	Kamusal Finans
H5	Sosyal Harcamalar	Kamusal Finans
Y1	Bebek Ölüm Oranı	Yaşam Kalitesi
Y2	Ortalama Yaşam Süresi	Yaşam Kalitesi
Y3	Hapishane Nüfusu	Yaşam Kalitesi

Y4	Trafik Kazaları	Yaşam Kalitesi
Y5	Gençlerin Etkisizliği	Yaşam Kalitesi
T1	Bilgisayar Erişimi	Bilim ve Teknoloji
T2	İnternet Erişimi	Bilim ve Teknoloji
T3	Haberleşme	Bilim ve Teknoloji
T4	AR-GE Harcamaları	Bilim ve Teknoloji
T5	Bilgi ve İletişim Ürünü Ticareti	Bilim ve Teknoloji
T6	Araştırmacı Sayısı	Bilim ve Teknoloji

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı : Ekrem Öncel KORKMAZ

Doğum Tarihi ve Yeri : 03.01.1986 - Aydın

Yabancı Dili : İngilizce

E-posta : eoncelkorkmaz@gmail.com

ÖĞRENİM DURUMU

Derece	**Alan**	**Okul/Üniversite**	**Mezuniyet Yılı**
Lisans	Bilgisayar Müh.	İstanbul Üniversitesi	2008
Lise	Fen Bilimleri	Aydın Adnan Menderes An. L.	2004

İŞ TECRÜBESİ

Yıl	**Firma/Kurum**	**Görevi**
2010-2011	Sbt Analiz	Yazılım Mühendisi
2009	Tübitak-MAM	Araştırmacı