Afyon Kocatepe Üniversitesi Fen ve Mühendislik Bilimleri Dergisi
Afyon Kocatepe University – Journal of Science and Engineering
https://dergipark.org.tr/tr/pub/akufemubid

Araştırma Makalesi / Research Article
e-ISSN: 2149-3367
DOI: https://doi.org/10.35414/akufemubid.1594932
AKÜ FEMÜBİD **25** (2025) 065103 (1348-1358)
AKU J. Sci. Eng. **25** (2025) 065103 (1348-1358)

Manifold Öğrenme Tekniklerinin Uygulaması ve Karşılaştırmalı Analizi

Makale Bilgisi / Article Info
Alındı/Received: 02.12.2024
Kabul/Accepted: 18.07.2025
Yayımlandı/Published: 03.12.2025

Application and Comparative Analysis of Manifold Learning Techniques

Hatice ÇOBAN EROĞLU*

Doğuş Üniversitesi, Mühendislik Fakültesi, Yazılım Mühendisliği Bölümü, İstanbul, Türkiye

© 2025 The Authors | Creative Commons Attribution-Noncommercial 4.0 (CC BY-NC) International License

Öz

Bu makalede, İzomap, Yerel Doğrusal Gömme (LLE), Hessian LLE (HLLE), t-dağıtımlı Stokastik Komşu Gömme (t-SNE) ve Boyut Azaltma için Düzgün Manifold Yaklaşımı ve Projeksiyonu (UMAP) manifold öğrenme tekniklerinin parametre uzayı konveks ve konveks olmayan manifoldlar üzerindeki veri setlerine uygulamaları yapılmıştır. Ayrıca bu tekniklerin performansları sayısal büyüklükler ve gömme sonuçlarının görselleri açısından incelenerek bir analiz sunulmaktadır. Bu çalışmada, kullanılan parametre uzayı konveks ve konveks olmayan manifoldlar Swiss Roll ve delikli Swiss Roll olarak adlandırılan manifoldlardır.

Anahtar Kelimeler: *Manifold öğrenmesi; Konveks ve konveks olmayan manifoldlar; Boyut azaltma teknikleri; Makine öğrenmesi.*

Abstract

In this paper, Isomap, Local Linear Embedding (LLE), Hessian LLE (HLLE), t-distributed Stochastic Neighbour Embedding (t-SNE) and Uniform Manifold Approximation and Projection for Dimensionality Reduction (UMAP) manifold learning techniques are applied to data sets on manifolds with convex and non-convex parameter spaces. In addition, an analysis of the performance of these techniques in terms of numerical quantities and visualizations of the embedding results are presented. The parameter space convex and non-convex manifolds used in this study are the so-called Swiss Roll and Swiss Roll with a rectangular hole.

Keywords: *Manifold learning; Convex and nonconvex manifolds; Dimensionality reduction; Machine learning.*

1. Giriş

Boyut azaltma (BA), bir veri kümesindeki özellik sayısının, mümkün olduğunca fazla bilgi korunarak azaltılması olarak tanımlanmaktadır. Yüksek boyutlu verilerin anlaşılması için önemli bir yaklaşım olarak görülen BA teknikleri, yüksek boyutlu verilerden doğrudan düşük boyutlu bir görselleştirme oluşturmayı amaçlamaktadır. Son yıllarda, çok sayıda BA tekniği geliştirilmiştir. Bu teknikler, genel olarak doğrusal ve doğrusal olmayan boyut azaltma (DOBA) olmak üzere iki temel kategoriye ayrılmaktadır. Doğrusal boyut azaltma yöntemleri arasında en yaygın olarak kullanılanı, temel bileşen analizidir (PCA: Principal Component Analysis). Bu tekniğin amacı yüksek boyutlu veri kümesindeki değişkenlerin doğrusal kombinasyonları aracılığıyla daha düşük boyutlu bir uzayda karşılık gelen temsillerini bulmaktır. PCA tekniğinin sıklıkla kullanılmasının nedenleri arasında matematiksel temelinin güçlü ve maliyetinin düşük olması yer almakta olup, bu özellikler yüksek boyutlu verilerin analizinde etkili ve verimli sonuçlar elde edilmesine olanak sağlamaktadır (Jolliffe ve Cadima 2016). Ancak, doğrusal tekniklerin, verideki doğrusal olmayan yapısal unsurları koruyamaması nedeniyle doğrusal olmayan yöntemlere olan ilgi artmıştır. Bu doğrultuda, İzomap (Tenenbaum vd. 2000), yerel doğrusal gömme (LLE: Locally Linear Embedding) (Roweis ve Saul 2000), Hessian yerel doğrusal gömme (HLLE: Hessian Locally Linear Embedding) (Donoho ve Grimes 2003), t-dağıtımlı stokastik komşu gömme (t-SNE: t-distributed Stochastic Neighbor Embedding) (van der Maaten ve Hinton 2008) ve boyut azaltma için düzgün manifold yaklaşımı ve projeksiyonu (UMAP:Uniform Manifold Approximation and Projection for dimenson reduction) (McInnes vd. 2018) gibi yöntemler geliştirilmiştir (Lee ve Verleysen 2007). Bu yöntemler, verinin yapısında doğrusal olmayan ilişkilerin bulunması durumunda tercih edilmekte ve bu tür ilişkilerin daha uygun biçimde temsil edilmesini mümkün kılmaktadır.

DOBA tekniklerinde, düşük boyutlu uzaydaki temsillerin, veri kümesindeki yerel yapıları, özellikle de veri noktaları arasındaki mesafeleri ve komşuluk ilişkilerini mümkün olduğunca koruyacak biçimde elde edilmesi hedeflenmektedir. Bu bağlamda, manifold kavramı temel bir rol oynamaktadır. Manifoldlar, her bir noktanın yakın çevresinde Öklid uzayına benzerlik gösteren topolojik uzaylar olarak tanımlanmaktadır (Lee 2010). DOBA

*Sorumlu Yazar/Corresponding Author: Hatice Çoban Eroğlu e-posta/e-mail: hcoban@dogus.edu.tr

yöntemleri, bu tür doğrusal olmayan manifoldların düşük boyutlu uzaylara eşlenmesi ya da gömülmesi süreciyle doğrudan ilişkilidir ve bu nedenle literatürde sıklıkla "manifold tabanlı boyut azaltma yöntemleri" olarak anılmaktadır. Bu tekniklerde, veri kümesindeki elemanların bir manifold üzerinde yer aldığı varsayılmakta; dolayısıyla, söz konusu manifoldun geometrik ve topolojik özelliklerinin korunması esas alınarak, anlamlı ve yapıyı koruyan düşük boyutlu temsillerin elde edilmesi amaçlanmaktadır.

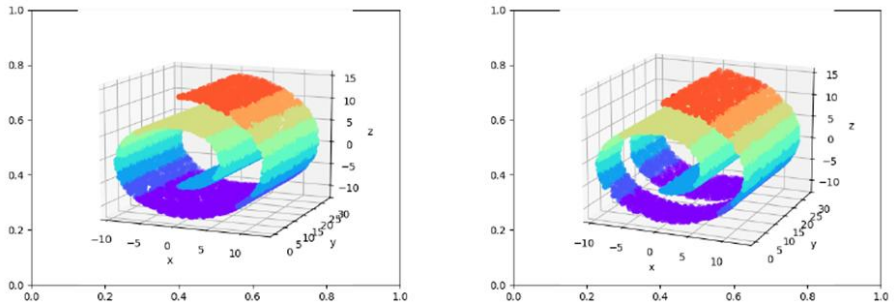
Girdi olarak, $\{m_i\}_{i=1}^{N} \subset \mathbb{R}^n, n \gg 1$ ile bir dizi yüksek boyutlu veri noktası verildiğinde, bu veri noktalarının $d \ll n$ olmak üzere d boyutlu bir $M \subset \mathbb{R}^n$ manifoldunun üzerinde olduğu kabul edilir ve $M \subset \mathbb{R}^n$ veri manifoldu olarak adlandırılır. Amaç, $M \subset \mathbb{R}^n$ kümesinin d parametreli bir parametrizasyonunu bulmaktır; yani bir $\theta \subset \mathbb{R}^d$ kümesinden M'ye giden bir $\varphi: \theta \to M$ fonksiyonu inşa etmektir. Yapısal özellikler, manifoldların topolojik ve/veya geometrik özelliklerini içermektedir.

Bu çalışmanın temel amacı, çeşitli manifold öğrenme tekniklerinin sayısal performanslarının ve elde edilen düşük boyutlu gömülme temsillerinin sistematik biçimde değerlendirilmesi ve karşılaştırılmasıdır. Bu doğrultuda, farklı topolojik ve geometrik özellikler taşıyan manifold yapıları üzerinde gerçekleştirilen deneysel analizler aracılığıyla, ilgili yöntemlerin yapısal özellikleri koruma performansları incelenecektir. Parametre uzayı konveks yapıya sahip Swiss Roll manifoldu ve daha karmaşık bir topolojik yapıya sahip, parametre uzayı konveks olmayan; delikli Swiss Roll varyantı kullanılarak, manifold öğrenme algoritmalarının düşük boyutlu uzayda yapısal bilgiyi ne ölçüde koruyabildiği değerlendirilecektir.

Bu bağlamda, gerçekleştirilecek olan karşılaştırmalar; algoritmaların yerel ve global yapı koruma performansları, komşuluk ilişkilerinin korunması, gömülme kalitesi ve topolojik yapının korunması gibi ölçütler çerçevesinde sayısal olarak analiz edilmesini içermektedir. Böylece, manifold öğrenme tekniklerinin farklı veri kümeleri üzerindeki güçlü ve zayıf yönlerinin belirlenmesi amaçlanmaktadır. Elde edilen bulguların, boyut azaltma yöntemlerinin seçiminde ve uygun algoritmaların belirlenmesinde rehberlik edici olması hedeflenmektedir.

2. Materyal ve Metot

Yöntemlerin etkinliğini ve gömülme kalitesini nesnel biçimde değerlendirebilmek amacıyla, biri konveks, diğeri ise konveks olmayan yapıda iki farklı manifold üzerinde deneysel analizler gerçekleştirilmiştir. Çalışmada kullanılan ilk manifold, literatürde yaygın olarak başvurulan ve konveks kümede parametrize edilmiş Swiss Roll manifoldudur. Swiss Roll, matematiksel olarak, $\mathbb{R}^3$ uzayında gömülü, karmaşık geometrik yapıya sahip, yüksek derecede kıvrımlı, iki boyutlu bir manifold olarak tanımlanır (Cayton 2005). Bu manifold, genellikle veri kümelerinin düşük boyutlu, içsel yapısını ortaya çıkarmak amacıyla veri madenciliği, makine öğrenmesi ve yüksek boyutlu veri analizlerinde sıklıkla kullanılan standart bir test verisidir. Swiss Roll, başlangıçta düz ve iki boyutlu bir manifoldun, üç boyutlu uzayda spiral şeklinde sarılması sonucu oluşur ve veri kümelerinin doğal geometri ve topolojik yapılarının incelenmesinde önemli bir model teşkil eder. İkinci manifold ise, Swiss Roll yapısına benzemekle birlikte, yüzeyinde dikdörtgen biçiminde bir boşluk (delik) bulunan ve bu çalışmada "delikli Swiss Roll" olarak adlandırılan parametre kümesi konveks olmayan manifolddur. Her ne kadar bu ikinci manifold genel olarak konveks olmayan bir yapıya sahip olsa da, üzerinde yer alan boşluğun geometrik sınırlarının konveks olduğu göz önünde bulundurulmuştur.



(a) Swiss Roll (b) Delikli Swiss Roll

Şekil 1. 5000 veri noktalı manifoldlar

Bu iki farklı manifoldun seçilmesindeki amaç, manifold öğrenme algoritmalarının hem düzgün (konveks) hem de karmaşık ve topolojik olarak bozulmuş (konveks olmayan, delikli) yapılar üzerindeki performanslarının karşılaştırmalı olarak analiz edilmesini sağlamaktır. Böylelikle, yöntemlerin topolojik duyarlılıkları, yapısal bilginin korunmasına ilişkin yetkinlikleri ve farklı manifold karakteristikleri karşısındaki tutarlılıkları derinlemesine

değerlendirilebilecektir. Bu çalışmada gerçekleştirilen deneysel analizlerde, ilgili manifoldlar üzerinde 5.000 adet veri noktasının mevcut olduğu varsayılmıştır (*Şekil 1*). Söz konusu veri noktalarının, manifoldların yapısal özelliklerini temsil edecek şekilde konumlanarak, boyut azaltma yöntemlerinin performanslarının tutarlı ve karşılaştırılabilir biçimde değerlendirilmesi amaçlanmaktadır.

Aşağıda sunulan alt bölümler, karşılaştırmalı değerlendirme kapsamında incelenen manifold öğrenme algoritmalarına ilişkin genel bir bilgilendirme sağlamak amacıyla hazırlanmıştır. Her bir yönteme dair temel ilkeler, varsayımlar ve uygulama alanları özetlenerek, ilgili algoritmaların değerlendirme sürecindeki bağlamı daha açık hâle getirilmiştir. Bu sayede, yöntemler arası farkların ve benzerliklerin sistematik biçimde ortaya konulması hedeflenmektedir.

2.1 İzomap

İzomap, PCA ve klasik çok boyutlu ölçekleme (MDS: Multidimensional Scaling) yöntemlerinin bir kombinasyonu olarak ele alınabilir. Bu yöntem, yüksek boyutlu verilerin altında yatan düşük boyutlu yapıyı ortaya çıkarmayı amaçlayan, doğrusal olmayan bir boyut indirgeme tekniği olarak geliştirilmiştir. MDS yönteminin jeodezik mesafelerle genelleştirilmiş bir versiyonu olarak kabul edilebilir ve manifold öğrenme problemlerinde etkin sonuçlar üretmesiyle tanınır. İzomap algoritması, yüksek boyutlu veri uzayında bulunan noktalar arasındaki yerel geometrik ilişkileri koruyarak ve veri kümesinin yapısal bütünlüğüne sadık kalarak düşük boyutlu bir gömme elde edilmesini sağlar (Tenenbaum vd. 2000). Bu teknik, yüz tanıma (Tenenbaum vd. 2000), genetik veri analizi (de Ridder ve Duin 2002), robotik kontrol (de Silva ve Tenenbaum 2004), tıbbi görüntüleme (Bode vd. 2011) ve doğal dil işleme (Mikolov vd. 2013) gibi birçok alanda uygulanmıştır. Özellikle eğrisel (1-boyutlu) manifoldlar üzerinde yer alan verilerde anlamlı temsiller oluşturmakta etkilidir. Örneğin, yüz ifadelerinin düşük boyutlu temsilleri, sınıflandırma ve tanıma işlemlerini kolaylaştırırken, genetik veri analizinde biyobelirteçlerin tespitine yardımcı olur. Benzer şekilde, robotik sistemlerde sensör verilerinin düşük boyutlu haritalanması, hareket planlaması ve yer bulma problemlerinin çözümüne katkı sağlar.

Algoritma adımları:

Komşuluk grafiğinin oluşturulması: Bu adımda, yüksek boyutlu uzaydaki veri noktaları bir düğüm olarak temsil edilerek bir komşuluk grafı $G = (V, E)$ oluşturulmaktadır. Her veri noktası, Öklid mesafesine göre en yakın k komşusu ile bağlanır.

En kısa yol mesafelerinin hesaplanması: Bu aşamada, komşuluk grafında bulunan tüm düğüm çiftleri arasındaki en kısa yol mesafeleri hesaplanmaktadır. Bu mesafelerin belirlenmesinde, Dijkstra veya Floyd-Warshall algoritmaları (Dijkstra 2022, Floyd 1962) gibi en kısa yol algoritmaları kullanılabilmektedir. Elde edilen mesafe matrisi DG, manifold üzerindeki yaklaşık jeodezik mesafeleri temsil etmektedir.

MDS'nin uygulanması: Klasik MDS yöntemiyle boyut indirgeme sürecinde, jeodezik mesafeleri içeren DG mesafe matrisi, bu adımda uygun bir merkezleme işlemi uygulanarak karşılık gelen Gram matrisi B'ye dönüştürülmektedir:

$B = -\frac{1}{2}H(DG)^2H$, burada $H = I_n - \frac{1}{n}e_n e_n^T$ şeklinde tanımlanır ve $e_n = [1, 1, \dots, 1]^T \in \mathbb{R}^n$ şeklindedir. Daha sonra, B matrisi üzerinde özdeğer ayrışımı yapılarak, en büyük pozitif özdeğerlere karşılık gelen d özvektörleri seçilmekte ve bu özvektörler, karşılık gelen özdeğerlerin karekökleriyle çarpılarak düşük boyutlu bir temsil oluşturulmaktadır.

Düşük boyutlu gömmenin oluşturulması: Bu adımda, düşük boyutlu koordinatlar $\{y_1, y_2, \dots, y_d\} \subset \mathbb{R}^d$, orijinal verinin manifold yapısını koruyan düşük boyutlu temsili olarak elde edilmektedir.

2.2 Yerel Doğrusal Gömme (LLE)

LLE, yerel ilişkilerin global yapıdan daha önemli olduğu veriler için etkilidir (Roweis ve Saul 2000). LLE, her veri noktasının yerel komşularıyla olan doğrusal ilişkilerini koruyarak verinin yapısını düşük boyutlu uzaya yansıtır ve böylece karmaşık veri kümelerinin daha anlamlı ve sade temsillerinin elde edilmesini sağlar. Bu özellikleri sayesinde LLE; görüntü işleme, yüz tanıma, tıbbi görüntü analizi, doğal dil işleme, biyoinformatik, robotik, finansal analizler ve anomali tespiti gibi çok çeşitli uygulama alanlarında yaygın olarak kullanılmaktadır (Roweis ve Saul 2000, Zhang 2003). Yöntem, yüksek boyutlu verilerde doğrusal olmayan yapıları açığa çıkarma ve yerel komşuluk ilişkilerini koruma yeteneği ile veri görselleştirme ve kümeleme gibi görevlerde etkin sonuçlar vermektedir (Belkin ve Niyogi 2003). LLE, modern makine öğrenimi ve veri bilimi uygulamalarında önemli bir araç olarak kabul edilmektedir (Charte vd. 2021)

Algoritma adımları:

Komşuluk grafiğinin oluşturulması: Veri kümesindeki her bir veri noktası x_i için k en yakın komşular kümesi N_i belirlenir. Bu komşuluk seçimi genellikle Öklid mesafesine dayalı olarak yapılır ve her veri noktasının yerel geometrisini yakalamayı amaçlar.

Ağırlık matrisinin hesaplanması: Belirlenen k komşu kullanılarak, her veri noktası için yerel doğrusal yaklaşım ağırlıkları W_{ij} hesaplanır. Bu ağırlıklar, aşağıdaki optimizasyon problemi çözülerek elde edilir:

$$\min \sum_i ||x_i - \sum_{j \in N_i} W_{ij} x_j \; ||^2.$$

Bu optimizasyon, her veri noktasının komşularının doğrusal birleşimi ile en iyi şekilde temsil edilmesini sağlar. Ağırlıklar aşağıdaki kısıtlara uymak zorundadır:

$\sum_{j \in N_i} W_{ij} = 1, \quad W_{ij} = 0$ eğer $j \notin N_i$.

Düşük boyutlu gömmenin oluşturulması: Ağırlık matrisinin belirlenmesinin ardından, veri noktalarının düşük boyutlu temsilini bulmak için aşağıdaki global optimizasyon problemi çözülür:

$$\min \sum_i ||y_i - \sum_{j \in N_i} W_{ij} y_j \; ||^2.$$

Bu problem, düşük boyutlu koordinatların çıkarılmasını içerir. Çözüm, ağırlık matrisinin Laplace benzeri bir matrisine dayanır ve en küçük özdeğerleri ve buna karşılık gelen özvektörleri kullanır.

2.3 Hessian Yerel Doğrusal Gömme (HLLE)

Hessian LLE, temel olarak LLE algoritmasının bir modifikasyonu olarak değerlendirilebileceği gibi, aynı zamanda Laplacian eigenmap (LE) yöntemi (Belkin ve Niyogi, 2003) çerçevesinde Laplacian operatörünün Hessian operatörü ile değiştirilmesi suretiyle geliştirilmiş, boyut indirgeme problemine diferansiyel geometrik bir yaklaşım getiren, manifold öğrenme temelli bir algoritmadır. HLLE, konveks kümelerde parametrelendirilmiş manifoldları irdeleyen ilk manifold öğrenme tekniği olma özelliğine sahiptir. Bu yöntem, belirli ölçüde konveks olmayan veri yapılarının modellenmesine olanak tanıyan nadir tekniklerden biri olarak öne çıkar. Bu algoritma, veri noktalarının yerel yapısını koruyarak veriyi daha kompakt ve anlamlı bir şekilde temsil etmeyi amaçlar. Hessian LLE, özellikle verinin yerel eğriselliği ve ikinci türev bilgilerini dikkate almasıyla klasik LLE'den farklılık gösterir (Donoho ve Grimes 2003). Bu yöntem, görüntü işleme ve biyoinformatik gibi alanlarda, özellikle yüz tanıma ve gen ekspresyonu analizinde veri yapısının karmaşıklığını anlamak için tercih edilmektedir (Donoho ve Grimes 2003). Ayrıca ses tanıma ve finansal veri analizi gibi çok değişkenli sinyallerde temel yapıları ortaya çıkarmada başarılı sonuçlar vermektedir. Robotik alanında ise hareket analizlerinde yüksek boyutlu sensör verilerinin anlamlı temsilini sağlamak için kullanılmaktadır (Saul ve Roweis 2003). Bu uygulamalar HLLE'nin, verilerin lokal geometri bilgisini Hessian matrisleri ile yakalayarak, doğrusal olmayan yapıları daha hassas bir şekilde modelleyebilmesinden kaynaklanmaktadır.

Algoritma adımları:

Komşuluk grafiğinin oluşturulması: Her bir veri noktası için, Öklid mesafesindeki en yakın k komşuları bulunur. Bu komşular, yüksek boyutlu veri uzayında yerel manifold yapısının bir tahminini sağlar. Her bir N_i komşuluğu için yeniden merkezlenmiş noktaların satırlardan oluşan bir $k \times n$ boyutlu M_i matrisi oluşturulur. Bu noktalar şu şekilde oluşturulur:

$x_j - \overline{x_\iota}, j \in N_i$ öyle ki $\overline{x_\iota} = \frac{1}{|N_i|} \sum_{j \in N_i} x_j$.

Tanjant koordinatlarının elde edilmesi: M_i matrisine tekil değer ayrıştırması (SVD: Singular Value Decomposition) uygulanır ve sonuç olarak U, D ve V matrisleri elde edilir, burada U, boyutu $n \times \min(k, n)$ olan bir matristir. U matrisinin ilk d sütunu, N_i komşuluğundaki noktaların tanjant koordinatlarını verir.

Hessian tahmincisinin geliştirilmesi: Hessian'ın en küçük kareler yöntemiyle kestirimi için gerekli altyapı geliştirilir.

Eğer $f: M \to \mathbb{R}$ şeklinde düzgün (smooth) bir fonksiyon ise ve $f_j = f(x_j)$ olarak tanımlanırsa, burada x_j noktaları N_i komşuluğundaki noktaları ifade eder, o halde N_i komşuluğundaki noktalara karşılık gelen değerleri içeren bir v_i vektörü tanımlanabilir. Bu durumda, matris-vektör çarpımı $H_i v_i$ boyutu $\frac{d(d+1)}{2}$ olan bir vektör verir. Bu vektörün bileşenleri f fonksiyonunun yerel tanjant koordinatları üzerindeki ikinci mertebeden türevleri, yani Hessian matrisinin elemanları olan $\frac{\partial f^2}{\partial U_i \partial U_j}$ türevlerinin yaklaşık değerlerine karşılık gelir.

Simetrik matrisin oluşturulması: Her bir koordinat çifti (i, j) için aşağıdaki biçimde tanımlanan bir simetrik $\mathcal{H}_{ij}$ matrisi oluşturulur:

$$\mathcal{H}_{ij} = \sum_l \sum_r (H^l)_{ri} (H^l)_{rj}.$$

Burada $(H^l)_{ri}$, l-yinci komşuluk olan N_l üzerindeki Hessian fonksiyonlarının r-yinci satır i-yinci sütununu temsil eder. Bu tanım, Hessian yaklaşımıyla elde edilen ikinci türevlerin karelerinin ve çapraz çarpımlarının toplamını içerir; böylece manifold üzerindeki eğrisellik bilgisi küresel bir simetrik matris olarak modellenmiş olur.

Düşük boyutlu gömmenin oluşturulması: Hessian matrisinin spektral özelliklerini incelemek amacıyla bir özdeğer çözümlemesi gerçekleştirilir. Bu analiz sonucunda, en küçük $d + 1$ özdeğere karşılık gelen alt uzay belirlenir. Bu alt uzay, verinin düşük boyutlu

temsiline karşılık gelir. Bu özdeğerlerden biri sıfırdır ve sabit fonksiyonları temsil eden özvektörlere karşılık gelir. Bu nedenle, sıfır özdeğere karşılık gelen özvektör göz ardı edilerek, kalan d boyutlu alt uzay, gömülü manifoldun düşük boyutlu temsilini elde etmek için kullanılır.

2.4 t-Dağıtımlı Stokastik Komşu Gömme (t-SNE)

t-SNE, yerel yapıları koruyarak ve verilerdeki kümeleri ortaya çıkararak yüksek boyutlu verileri görselleştirmek için kullanılan istatistiksel bir yöntemdir. Düşük boyutlu gömme ile yüksek boyutlu verilerin ortak olasılıkları arasındaki Kullback-Leibler sapmasını (Kullback ve Leibler 1951) en aza indirgemek amacıyla veri noktaları arasındaki benzerlikleri ortak olasılıklara dönüştürür (van der Maaten ve Hinton 2008). Algoritma, veri görselleştirme amacıyla yaygın olarak kullanılmakta olup, kümelenme yapılarının ve yerel ilişkilerin düşük boyutlu temsillerde sezgisel olarak analiz edilmesini mümkün kılar. Bu teknik biyoinformatik, görüntü işleme, doğal dil işleme, pazarlama analitiği ve medikal veri analizinde yaygın olarak uygulanmaktadır (van der Maaten ve Hinton 2008). Örneğin, tek hücre RNA dizileme verilerinde hücre tiplerinin keşfi ve gen ekspresyon verilerinin analizi için etkili bir yöntem olarak tercih edilir (Kobak ve Berens 2019). Ayrıca, derin öğrenme tabanlı görüntü özelliklerinin görselleştirilmesi, kelime gömme modellerindeki anlamsal ilişkilerin incelenmesi ve müşteri segmentasyonunda da kullanımı artmaktadır (Mikolov vd. 2013). Sağlık alanında ise t-SNE, kanser hücre alt tiplerinin tanımlanması ve hasta verilerinin analizi gibi uygulamalarda başarıyla kullanılmaktadır (Amir vd. 2013).

Algoritma adımları:

Yüksek boyutlu uzayda ikili benzerliklerin hesaplanması:

Yüksek boyutlu bir veri seti $\{x_1, x_2, \dots, x_n\}$ verildiğinde, veri noktaları arasındaki benzerliklere orantılı olan p_{ij} olasılıkları hesaplanır. Veri noktası x_j'nin x_i veri noktasına olan benzerliği, x_i'nin komşularını, x_i merkezli bir Gauss dağılımı altında olasılık yoğunluklarına orantılı olarak seçeceği varsayımıyla, x_j'yi komşu olarak seçme koşullu olasılığı $p_{j|i}$ olarak tanımlanır (van der Maaten ve Hinton 2008).

$i \neq j$ için $p_{j|i} = \frac{\exp\left(-\left||x_i - x_j\right|\right|^2/2\sigma_i^2)}{\sum_{k\neq i}\exp\left(-||x_i - x_k||^2/2\sigma_i^2\right)}$, $\sum_j p_{j|i} = 1$ ve $p_{i|i} = 0$. Burada σ_i her veri noktasına özgü yerel bant genişliği parametresidir ve perpleksite kavramı çerçevesinde otomatik olarak belirlenir.

Simetrik Ortak Olasılıkların Elde Edilmesi: Koşullu olasılıkların tek yönlü tanımlı olmaları nedeniyle, simetrik bir benzerlik ölçüsü elde edebilmek için aşağıda tanımlanan ortak olasılık tanımı kullanılır:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}.$$

Bu ifade, yüksek boyutlu uzaydaki tüm gözlem çiftleri arasındaki karşılıklı benzerliği normalize edilmiş şekilde temsil eder.

Düşük Boyutlu Temsilde Olasılıkların Modellenmesi: Düşük boyutlu uzaydaki $\{y_1, y_2, \dots, y_d\}$ noktaları arasındaki benzerlikler, t-dağılımı kullanılarak tanımlanır:

$$q_{ij} = \frac{\left(1 + \left|\left|y_i - y_j\right|\right|^2\right)^{-1}}{\sum_{k\neq l}\left(1 + \left|\left|y_k - y_l\right|\right|^2\right)^{-1}}, \quad q_{ii} = 0.$$

Bu dağılım, düşük boyutlu uzayda daha geniş kuyruklara sahip olması sayesinde uzak gözlemlerin etkisinin korunmasına olanak tanır, bu da kümeler arasındaki ayrımın daha belirgin hale gelmesini sağlar.

Kullback-Leibler (KL) Farklılığının minimize edilmesi: Yüksek boyutlu ve düşük boyutlu olasılık dağılımları arasındaki fark, Kullback-Leibler (KL) ayrıklığı ile ölçülür:

$$KL(P||Q) = \sum_{i\neq j} p_{ij}\log\left(\frac{p_{ij}}{q_{ij}}\right).$$

Bu maliyet fonksiyonu, düşük boyutlu temsillerin öğrenilmesi sürecinde optimizasyon hedefi olarak kullanılır ve genellikle stokastik gradyan inişi (SGD: Stochastic Gradient Descent) veya onun momentum ve adaptif varyantları aracılığıyla minimize edilir.

Düşük Boyutlu Temsillerin Elde Edilmesi ve Görselleştirme: Optimizasyon süreci sonucunda elde edilen düşük boyutlu temsiller olan y_i noktaları, veri kümesinin yapısal özelliklerini koruyarak 2 veya 3 boyutlu uzayda görselleştirilmeye hazır hale gelir. Bu temsiller, özellikle kümeleme yapılarının ve aykırı değerlerin sezgisel olarak analiz edilmesini kolaylaştırır.

2.5 Boyut Azaltma için Düzgün Manifold Yaklaşımı ve Projeksiyonu (UMAP)

UMAP, son yıllarda geliştirilmiş olan ve manifold öğrenme prensiplerine dayanan bir boyut indirgeme yöntemidir. Algoritma, manifold öğrenme ile topolojik veri analizinden türetilen teorik temeller üzerine inşa edilmekte olup, hem genel bir boyut indirgeme çerçevesi sunmakta hem de bu çerçeve kapsamında belirli somut uygulamalara olanak tanımaktadır. UMAP, görselleştirme performansı açısından t-SNE ile rekabet edebilmekle birlikte, işlem süresi açısından belirgin bir üstünlük sergilemekte ve veri setinin global yapısını daha yüksek düzeyde muhafaza etmektedir. Ayrıca, UMAP algoritmasının gömme boyutuna ilişkin herhangi bir hesaplama kısıtlaması bulunmaması, onu makine

öğrenmesi uygulamalarında genel amaçlı ve esnek bir boyut indirgeme aracı haline getirmektedir (McInnes vd. 2018). UMAP, temelde manifold öğrenme paradigması çerçevesinde değerlendirilmekte olup kategori teorisi, Riemann geometrisi ve fuzzy topoloji kavramları üzerine inşa edilmiştir. Algoritmanın matematiksel çerçevesi, veri kümelerinin üzerinde bulunduğu varsayılan manifoldun düşük boyutlu izdüşümlerle, mümkün olduğunca bozulmadan temsil edilmesini amaçlamaktadır. Özellikle tek hücre RNA dizimi verilerinin görselleştirilmesinde ve biyoinformatik alanında başarılı sonuçlar göstermiştir (Becht vd. 2019). Ayrıca, genetik veri kümelerindeki yapısal farklılıkların ortaya çıkarılmasında (Diaz-Papkovich vd. 2019) ve doğal dil işleme uygulamalarında kelime ve cümle gömme vektörlerinin görselleştirilmesinde tercih edilmektedir (Kobak ve Berens 2019). UMAP, anomali tespiti gibi karmaşık veri setlerinde yapısal bilgiyi koruyarak anlamlı niteliklerin belirlenmesine de olanak sağlamaktadır (Kobak ve Linderman 2021).

Algoritma adımları:
Grafik Yapımı: Her bir veri noktası için, bir k en yakın komşuluk yapısı kurulur. Bu yapı, her noktanın çevresindeki lokal topolojik yapının çıkarılmasını sağlar. UMAP, bu komşulukları kullanarak bir fuzzy simplicial kompleks oluşturur ve bu yapı üzerinden olasılıksal bir benzerlik grafiği tanımlar. Her komşuluk ilişkisi, olasılıksal bir bağlantı ağırlığı ile temsil edilir:

$$w_{ij} = \exp\left(-\frac{d(x_i, x_j) - \rho_i}{\sigma_i}\right).$$

Burada ρ_i minimum mesafedir, σ_i ise lokal ölçek parametresidir ve her noktaya özgüdür. Bu sayede, her noktanın farklı yoğunluktaki bölgelere uyum sağlaması mümkün olur.

Düşük Boyutlu Uzayda Yakınlıkların Modellemesi: Yüksek boyutlu uzayda elde edilen fuzzy topolojik yapı, düşük boyutlu uzayda da yeniden oluşturulmaya çalışılır. Düşük boyutlu temsiller arasında bir olasılık dağılımı tanımlanır:

$$w'_{ij} = \frac{1}{1 + a\left||y_i - y_j|\right|^{2b}},$$

burada a ve b uzaklık fonksiyonunun şekil parametreleridir ve bu parametreler, yüksek ve düşük boyutlu verilerin dağılımları arasındaki uyumu artırmak amacıyla optimizasyon öncesinde sabitlenir.

Olasılık Dağılımları Arasındaki Ayrımın Minimizasyonu: Yüksek boyutlu ve düşük boyutlu fuzzy simplicial kompleksler arasındaki ayrımı minimize etmek için çapraz entropi tabanlı bir maliyet fonksiyonu kullanılır:

$$\sum_{(i,j)} w_{ij} \log\left(\frac{w_{ij}}{w'_{ij}}\right) + (1 - w_{ij}) \log\left(\frac{1 - w_{ij}}{1 - w'_{ij}}\right).$$

Bu hedef fonksiyon, gradyan tabanlı optimizasyon teknikleri aracılığıyla minimize edilir. Bu süreç, düşük boyutlu temsillerin iteratif olarak güncellenmesini sağlar.

3. Bulgular

3.1 Kalite Kriterleri ve Ölçümü

BA teknikleri, yüksek boyutlu veri kümelerinin daha düşük boyutlu temsillerini elde etmeyi amaçlayarak veri analizi, görselleştirme ve ön işleme süreçlerinde yaygın biçimde kullanılmaktadır. Bu tekniklerin performans kalitesinin değerlendirilmesi, indirgenen veri uzayında orijinal yapısal bilgilerin ne ölçüde korunduğunun anlaşılması açısından kritik öneme sahiptir. Literatürde bu değerlendirme genellikle mesafe sıralamaları ve k-komşuluk ilişkilerine dayalı ölçütler aracılığıyla gerçekleştirilmektedir (Venna ve Kaski 2006). Özellikle komşuluk korunumu, güvenilirlik, süreklilik ve sıralama korelasyonu gibi metrikler, yerel ve global yapıların ne derece korunduğunu sayısal olarak ifade etmeye olanak tanımaktadır (Kaski vd. 2003, Venna vd. 2010). Buna ek olarak, topolojik korunumu esas alan yöntemler, veri manifoldunun topolojik özelliklerini (örneğin bağlantılılık, delikler, bileşenler) dikkate alarak indirgenmiş temsillerin kalitesini analiz etmektedir (Moon vd. 2019). Bu bağlamda, farklı boyut azaltma algoritmalarının karşılaştırmalı analizleri, hem nicel ölçütlerle hem de görsel sonuçlar üzerinden gerçekleştirilmektedir. Sonuç olarak, boyut azaltma yöntemlerinin performans kalite değerlendirmesi, veri temsillerinin güvenilirliğini artırmak ve doğru yöntem seçimlerini yapmak açısından vazgeçilmez bir adımdır.

Bu bölümün amacı düşük boyutlu gömme kalitesi üzerinde aşağıda tanımlanacak olan ortak sıralama matrisine dayalı bazı ölçümler belirlemektir. Her yöntem girdi verilerini farklı şekilde ele alır ve farklı amaç fonksiyonlarını optimize eder (Lee ve Verleysen 2009, Lueks vd. 2011). Son yıllarda birçok yeni yöntem önerilmiş olmakla birlikte, bu yöntemlerin değerlendirilmesi ve karşılaştırılması konusundaki soru hâlâ açık bir şekilde gündemdedir. Bu çalışmada, mesafe sıralaması ve k komşuluklara dayalı bazı mevcut kalite ölçütleri incelenmektedir. Aşağıda boyut azaltma yöntemlerinin performans kalitesini değerlendirmek için bazı kriterlerden bahsedilmektedir.

Yüksek boyutlu uzayda m_i ve m_j arasındaki uzaklık $d_{ij} = d(m_i, m_j)$ ve düşük boyutlu uzayda (parametre uzayında) x_i ve x_j arasındaki uzaklık $\tilde{d}_{ij} = d(x_i, x_j)$

olsun. Burada $\tilde{d}_{ii} = d_{ii} = 0$ olduğu açıktır ve m_j'nin m_i'ye göre olan rankı r_{ij} aşağıdaki şekilde yazılır.

$$r_{ij} = |\{k: d_{ik} \leq d_{ij}, 1 \leq k < j \leq N\}|$$

burada $|.|$ küme kardinalitesini yani kümenin eleman sayısını göstermektedir. Benzer olarak x_j'nin x_i'ye göre olan rankı

$$\tilde{r}_{\mathrm{ij}} = \left|\{k: \tilde{d}_{\mathrm{ik}} \leq \tilde{d}_{\mathrm{ij}}, 1 \leq k < j \leq N\ \}\right|$$

şeklindedir.

Ortak sıralama matrisi Q, j mesafe sıralamasının i sıralamasına dönüştüğü noktaların sayısı olarak tanımlanabilir:

$$Q = [q_{kl}] \text{ öyle ki } q_{kl} = \left|\{(i,j) : r_{ij} = k, \tilde{r}_{ij} = l\}\right|.$$

Boyut azaltma yönteminin performansını değerlendirmek için ortak sıralama matrisinden türetilen ölçütler aşağıda belirtilmiştir.

Q_{NX} kriteri: Hem yüksek hem de düşük boyutlu uzayda k −yıncı en yakın komşulara ait nokta sayısıdır.

$$Q_{NX}(k) = \frac{1}{Nk} \sum_{i=1}^{k} \sum_{j=1}^{k} q_{ij}\ .$$

$\boldsymbol{Q_{lokal}}$**:** Yerel mesafelerin korunmasını ölçer.

$$Q_{lokal} = \frac{1}{k_{maks}} \sum_{k=1}^{k_{maks}} Q_{NX}(k),$$

öyle ki

$$k_{maks} = \arg max\ LCMC(k)$$
$$= \arg max\ (Q_{NX}(k) - \frac{k}{N-1}).$$

Q_{global}**:** Global mesafelerin korunmasını ölçer.

$$Q_{global} = \frac{1}{N - k_{maks}} \sum_{k=k_{maks}}^{N-1} Q_{NX}(k).$$

Kofenetik korelasyon: Orijinal modellenmemiş veri noktaları arasındaki çift yönlü mesafelerin korunma kalitesini ölçer. Yüksek ve düşük boyutlu bir uzayda mesafe matrislerinin (kofenetik matrisler olarak bilinen dendrogramlarda) üst veya alt üçgenleri arasındaki korelasyon aşağıdaki gibi verilir:

$$\frac{\sum_{i<j}(d_{ij} - \bar{d})(t_{ij} - \bar{t})}{\sqrt{\sum_{i<j}(d_{ij} - \bar{d})^2 \sum_{i<j}(t_{ij} - \bar{t})^2}}.$$

Burada t_{ij}, model noktaları arasındaki dendrogramatik mesafeyi $\bar{t}$, t_{ij} değerlerinin ortalamasını ve $\bar{d}$, d_{ij} değerlerinin ortalamasını göstermektedir.

Toplam korelasyon: Değişkenlerin eksenlerle olan toplam korelasyonunu hesaplar.

3.2. Sayısal Sonuçlar

Bu bölümde, Swiss Roll ve delikli Swiss Roll manifoldları kullanılarak manifold öğrenme teknikleri için Q_{lokal}, Q_{global}, kofenetik korelasyon ve toplam korelasyon için sayısal bir karşılaştırma sunulmaktadır (*Çizelge 1* ve *Çizelge 2*). Hangi manifold öğrenme yönteminin en etkili olduğunu belirlemek kolay değildir. En uygun yaklaşım özel hedefe bağlıdır. Örneğin, yerel özelliklerin korunmasına odaklanılıyorsa, en iyi Q_{lokal} sonuçları veren teknik seçilmelidir.

Aşağıdaki bütün hesaplamalarda İzomap, LLE ve HLLE için en yakın k komşu sayısı $k = 10$ ve t-SNE ve UMAP için perpleksite de 10 olarak alınmıştır. Sayısal hesaplamalar R programlama dili ve 'dimRed' paketi kullanılarak gerçekleştirilmiştir. Bu paket, R programlama dili içerisinde boyut indirgeme yöntemlerini ortak bir çerçevede birleştiren ve bu yöntemlere nesne yönelimli bir yapı üzerinden erişim imkânı sunan kapsamlı bir araçtır (Kraemer vd. 2018). dimRed, yalnızca algoritmaların uygulanmasını kolaylaştırmakla kalmayıp, aynı zamanda elde edilen düşük boyutlu temsillerin görselleştirilmesini sağlayan çizim fonksiyonları ve yöntemlerin kalitesini değerlendirmeye yönelik nicel ölçütler de sunar. Bu yönüyle, hem görsel analiz hem de sayısal değerlendirme açısından kapsamlı bir araç takımı sağlar.

Çizelge 1: Swiss Roll

$n = 5000$	HLLE	LLE	İzomap	t-SNE	UMAP
Q_{lokal}	0.6495	0.6355	0.9362	0.7704	0.6841
Q_{global}	0.7178	0.6912	0.7284	0.6316	0.7062
Kofenetik korelasyon	0.5772	0.4346	0.2784	0.2819	0.3952
Toplam korelasyon	0.3708	0.3033	0.3706	0.1575	0.2613

Çizelge 2: Delikli Swiss Roll

$n = 5000$	HLLE	LLE	İzomap	t-SNE	UMAP
Q_{lokal}	0.5832	0.6408	0.8692	0.7963	0.6770
Q_{global}	0.6896	0.6617	0.7206	0.7137	0.6959
Kofenetik korelasyon	0.4966	0.4004	0.2929	0.5388	0.3931
Toplam korelasyon	0.3827	0.3438	0.3393	0.3384	0.1989

Çizelge 1'de Swiss Roll manifolduna uygulanan manifold öğrenme algoritmalarının performansları, çeşitli yapısal kalite ölçütleri aracılığıyla nicel olarak değerlendirilmiştir. Bu kapsamda HLLE, LLE, İzomap, t-SNE ve UMAP algoritmaları; lokal kalite Q_{lokal}, global kalite Q_{global}, kofenetik korelasyon ve toplam korelasyon gibi ölçütler üzerinden karşılaştırılmıştır. Elde edilen sonuçlar, İzomap algoritmasının hem lokal hem de global yapı koruma açısından en yüksek değerleri ürettiğini ve bu nedenle manifoldun geometrik yapısını düşük boyutlu uzaya aktarırken en başarılı yaklaşım olduğunu ortaya koymuştur. HLLE algoritması ise özellikle kofenetik korelasyon ve toplam korelasyon açısından yüksek performans sergileyerek, hiyerarşik yapıların ve genel mesafe ilişkilerinin korunmasında güçlü bir alternatif olarak öne çıkmıştır. t-SNE, lokal komşuluk ilişkilerini korumada başarılı olsa da, global yapının bütünsel temsili açısından daha düşük sonuçlar üretmiştir. Bu durum, t-SNE'nin yerel yapıya odaklı çalışma prensibiyle uyumlu olup, söz konusu yöntemin yalnızca lokal yapıların analizi için tercih edilmesi gerektiğini göstermektedir. UMAP algoritması ise hem lokal hem de global ölçütlerde dengeli bir performans sergileyerek, farklı yapıların temsilinde uyarlanabilir bir yöntem olarak değerlendirilebilir. LLE algoritması ise tüm kriterlerde görece daha düşük performans sergilemiş ve sınırlı bir yapısal koruma kapasitesi sunmuştur.

Çizelge 2'de delikli Swiss Roll manifolduna uygulanan manifold öğrenme algoritmaları, çeşitli yapısal kalite ölçütleri aracılığıyla nicel olarak değerlendirilmiştir. Elde edilen sonuçlara göre, İzomap algoritması hem lokal hem de global kalite açısından en yüksek başarıyı göstermiştir. t-SNE, özellikle lokal yapının korunmasında yüksek bir performans sergilemiş ve aynı zamanda kofenetik korelasyon bakımından da en yüksek değere ulaşmıştır. HLLE, toplam korelasyon ölçütü bakımından öne çıkarken, genel yapının korunmasına yönelik dengeli bir temsil sağlamıştır. LLE ve UMAP ise tüm ölçütlerde görece düşük performans göstermiştir. Genel olarak değerlendirildiğinde, İzomap ve t-SNE algoritmaları delikli manifold yapılarında hem lokal hem de global yapıların temsili açısından güçlü performans sergilemiş, HLLE ise bütüncül yapının korunmasına yönelik dikkate değer sonuçlar üretmiştir.

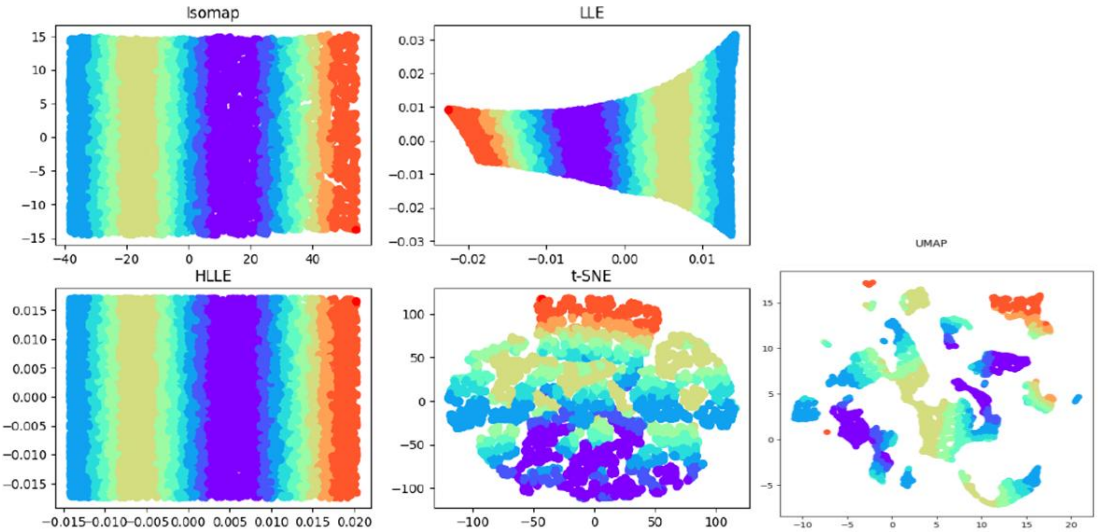
Çizelge 1 ve Çizelge 2 birlikte değerlendirildiğinde, manifold öğrenme yöntemleri arasında kayda değer performans farklılıklarının bulunduğu gözlemlenmektedir. Bu durum, yöntemlerin etkinliğinin veri setinin yapısal özelliklerine bağlı olarak değişiklik gösterebildiğini ortaya koymaktadır. Elde edilen bulgular, boyut indirgeme sürecinde tercih edilecek algoritmanın; veri setinin geometrik özellikleri, analiz hedefi ve korunması istenen yapısal ilişkiler dikkate alınarak belirlenmesi gerektiğine işaret etmektedir. Bu çerçevede, yöntem seçiminin bağlamsal nitelik taşıdığı ve genel geçer bir yaklaşımın her veri yapısında aynı düzeyde başarılı olamayacağı anlaşılmaktadır.

3.3 Manifold Öğrenme Tekniklerini Kullanarak Elde Edilen Düşük Boyutlu Manifoldlar

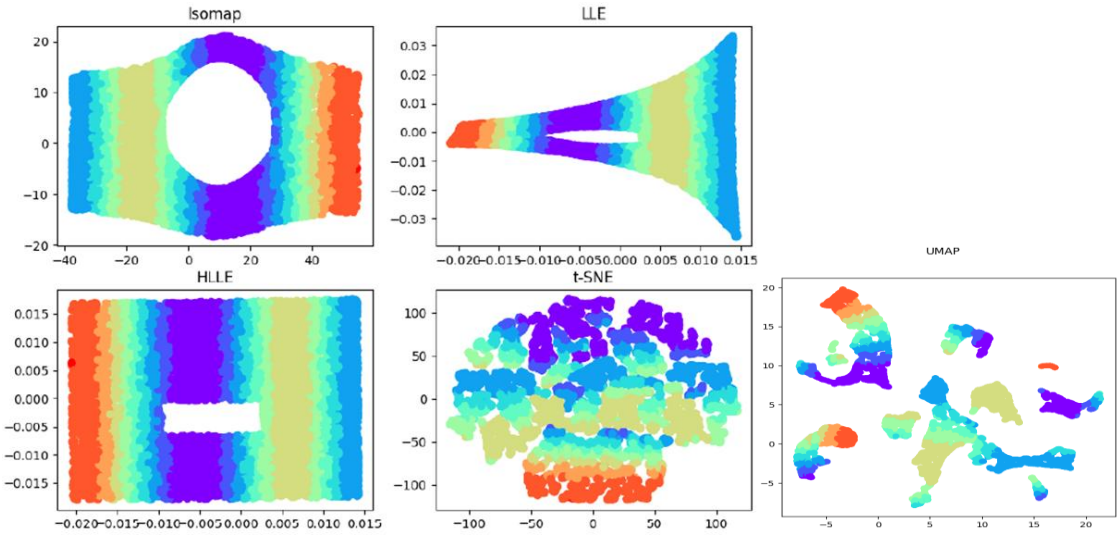
Bu bölümde, ilgili manifoldlar üzerinde İzomap, LLE, HLLE, t-SNE ve UMAP algoritmaları kullanılarak elde edilen düşük boyutlu gömülme sonuçlarına yer verilmektedir. Bu çalışmada, manifold öğrenme algoritmalarının uygulanması ve sonuçların görselleştirilmesi süreci Python programlama dili aracılığıyla gerçekleştirilmiştir. Uygulama aşamasında, scikit-learn (Pedregosa vd. 2011) adlı açık kaynaklı makine öğrenimi kütüphanesinden yararlanılmıştır. Söz konusu kütüphane, çok sayıda manifold öğrenme algoritmasının kolay ve esnek bir biçimde uygulanmasına olanak tanımaktadır. Her bir algoritma sonucunda elde edilen düşük boyutlu temsiller, iki boyutlu düzleme indirgenmiş ve sonuçlar 'matplotlib' kullanılarak görselleştirilmiştir. Bu görselleştirmeler, farklı manifold yapılarına karşılık gelen veri kümelerinin düşük boyutlu uzayda nasıl temsil edildiğini analiz etmeye ve algoritmalar arasındaki yapısal koruma düzeylerini kıyaslamaya imkân tanımaktadır.

Şekil 2'de Swiss Roll veri yapısı üzerinde uygulanan beş farklı manifold öğrenme algoritması tarafından elde edilen düşük boyutlu gömülme sonuçları yer almaktadır. Bu temsiller aracılığıyla, algoritmaların veri yapısına ilişkin farklı yapısal özellikleri nasıl yansıttıkları gözlemlenmektedir. İzomap yöntemi ile elde edilen gömülme, orijinal manifoldun global yapısını koruyarak düzlemsel bir açılım sağlamış; bu durum, yöntemin uzaklık bilgilerini bütünsel olarak yansıtmadaki başarısını göstermektedir. LLE ile elde edilen temsil, Swiss Roll kıvrımlarının doğrusal olmayan bir şekilde açıldığını ortaya koymakta ve bu yöntemin lokal komşuluk ilişkilerini korumada etkili, ancak global yapı açısından sınırlı olduğu izlenimini vermektedir. HLLE tarafından üretilen gömülme sonucu, lokal doğrusal yapıları koruma eğiliminde olup, düzlemsel bir açılım sağlamıştır. t-SNE, benzer örnekleri kümeler halinde gruplayarak lokal komşulukları yüksek doğrulukla yansıtan bir gömülme üretmiş, ancak bu süreçte veri setinin global yapısının korunması zayıf kalmıştır. Son olarak, UMAP algoritmasıyla elde edilen görsel temsil, hem lokal hem de belirli ölçüde global yapısal özellikleri yansıtan çok kümeli ve parçalı bir yapı ortaya koymuştur. Özellikle veri noktaları arasındaki lokal benzerliklerin korunmasında başarılı bir dağılım elde edilmiştir.

Şekil 3'te delikli Swiss Roll veri yapısı üzerinde uygulanan beş farklı manifold öğrenme algoritmasının iki boyutlu gömülme sonuçları sunulmaktadır. İzomap tarafından elde edilen gömülme, hem global yapıyı hem de manifold üzerinde yer alan deliğin topolojik yapısını başarıyla korumuştur. Özellikle, delik yapısının düzgün biçimde korunmuş olması, algoritmanın jeodezik mesafeleri doğru biçimde modelleyebildiğini göstermektedir. LLE yöntemiyle elde edilen gömülmede, lokal komşuluklar korunurken, deliğin etkisi kısmen bastırılmış ve yapı ciddi biçimde bükülerek bozulmuştur.



Şekil 2: Swiss Roll için manifold öğrenme tekniklerinin verdiği sonuçlar



Şekil 3: Delikli Swiss Roll için manifold öğrenme tekniklerinin verdiği sonuçlar

Bu durum, LLE'nin global topolojiye karşı duyarsız kaldığını göstermektedir. HLLE, delikli yapının varlığını açık şekilde yansıtan bir gömülme sunmuştur. Lokal ilişkilerin korunmasında başarılı olan bu yöntem, deliğin konumunu ve yapısını düşük boyutlu veri temsiline etkili bir biçimde aktarmıştır. t-SNE, veri noktalarını kümeler halinde gruplayarak lokal benzerlikleri ön plana çıkarmış; ancak global yapı ve delik yapısı belirgin biçimde kaybolmuştur. Bu durum, yöntemin global geometriyi yansıtma kapasitesinin sınırlı olduğunu göstermektedir. UMAP, lokal komşuluk yapılarını başarılı şekilde korumuş, ancak global topolojiyi özellikle de delik yapısını açık biçimde yansıtamamıştır. Elde edilen gömülmede veriler ayrık kümeler halinde sunulmuş; bu durum, UMAP'in lokal yoğunluklara öncelik verdiğini göstermektedir. Sonuçlar, UMAP'in lokal yapı analizi için uygun olduğunu, ancak karmaşık topolojik özelliklerin korunmasında sınırlı kalabildiğini ortaya koymaktadır.

4. Sonuçlar ve Tartışma

Bu çalışmada, manifold tabanlı boyut azaltma yöntemlerine ilişkin temel fikirler, algoritmik yaklaşımlar ve sayısal performans değerlendirmeleri kapsamlı biçimde incelenmiştir. Temsili yöntemler olarak İzomap,

LLE, HLLE, t-SNE ve UMAP ele alınmıştır. Bölüm 3.2'de bu yöntemlerin sayısal performansları sunulmuş, Bölüm 3.3'te ise parametre uzayı hem konveks hem de konveks olmayan manifoldlar üzerindeki gömülme sonuçları karşılaştırmalı olarak değerlendirilmiştir.

Şekil 2'de yer alan sonuçlar, Swiss Roll veri kümesi için en uygun yerleştirmelerin İzomap ve HLLE yöntemleri tarafından gerçekleştirildiğini ortaya koymaktadır. Swiss Roll manifoldunun kıvrımlarının başarılı bir şekilde açılarak, verinin iki boyutlu uzaya düzgün bir şekilde izdüşümünün sağlandığı gözlemlenmiştir. Buna karşılık, konveks olmayan delikli Swiss Roll manifolduna ilişkin Şekil 3'teki sonuçlar, İzomap algoritmasının global yapıyı koruma konusundaki başarısının daha karmaşık topolojide sürdürülemediğini ortaya koymaktadır. Aynı şekilde, LLE, t-SNE ve UMAP yöntemlerinin her iki manifold yapısında da tatmin edici gömülmeler üretemediği dikkat çekmektedir.

Bölüm 3.2'de sunulan sayısal analizlerde, İzomap, LLE ve t-SNE algoritmalarının Öklid mesafesine dayalı ölçütlerde HLLE'ye kıyasla daha yüksek skorlar elde ettiği belirlenmiştir. Ancak konveks olmayan yapıların dikkate alındığı durumlarda, HLLE'nin Swiss Roll ve delikli Swiss Roll manifoldlarda olduğu gibi kıvrımları düzgün bir şekilde açarak topolojik bütünlüğü daha başarılı şekilde yansıttığı ve bu bağlamda en uygun gömülmeyi sunduğu Şekil 3'te görülmektedir.

Bu bulgular doğrultusunda, manifold öğrenme tabanlı boyut azaltma tekniklerinin seçiminde, veri setinin dayandığı manifoldun yapısal özelliklerinin ve analizde korunması hedeflenen ölçülerin dikkate alınmasının zorunlu olduğu sonucuna varılmaktadır. Her bir yöntemin farklı veri yapılarında gösterdiği performans, yöntem seçiminin analizin amacına ve verinin topolojik karmaşıklığına göre dikkatle yapılması gerektiğini göstermektedir.

Teşekkür

Yazar, çalışmaya katkı sağlayan yapıcı değerlendirmeleri ve dikkatli okumaları dolayısıyla değerli hakemlere teşekkür eder.

Etik Standartlar Bildirgesi
Yazarlar tüm etik standartlara uyduklarını beyan ederler.

Yazarlık Katkı Beyanı
Yazar 1: Kaynaklar, Araştırma, Deney, Görselleştirme, Yazma

Çıkar Çatışması Beyanı
Yazarların bu makalenin içeriğiyle ilgili olarak beyan edecekleri hiçbir çıkar çatışması yoktur.

5. Kaynaklar

Amir, E. A. D., Davis, K. L., Tadmor, M. D., Simonds, E. F., Levine, J. H., Bendall, S. C., Shenfeld, D. K., Krishnaswamy, S., Nolan, G. P. and Pe'er, D., 2013. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nature biotechnology*, 31(6), 545-552.

Becht, E., McInnes, L., Healy, J., Dutertre, C. A., Kwok, I. W., Ng, L. G., Ginhoux, F. and Newell, E. W., 2019. Dimensionality reduction for visualizing single-cell data using UMAP. *Nature biotechnology*, 37(1), 38-44.

Belkin, M. and Niyogi, P., 2003. Laplacian eigenmaps for dimensionality reduction and data representation, *Neural Computation*, 15(6):1373-1396.

Bode, S., He, A. H., Soon, C. S., Trampel, R., Turner, R. and Haynes, J. D. (2011). Tracking the unconscious generation of free decisions using uitra-high field fMRI. *PloS one, 6*(6), e21612.

Cayton, L., 2005. Algorithms for manifold learning. Technical Report CS2008-0923, *University of California, San Diego.*

Charte, D., Charte, F., and Herrera, F., 2021. Reducing data complexity using autoencoders with class-informed loss functions*. IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12), 9549-9560.

De Ridder, D. and Duin, R. P., 2002. Locally linear embedding for classification. *Pattern Recognition Group, Dept. of Imaging Science and Technology, Delft University of Technology*, Delft, The Netherlands, Tech. Rep. PH-2002-01, 1-12.

De Silva, V. and Tenenbaum, J. B., 2004. Sparse multidimensional scaling using landmark points (Vol. 120). *Technical report, Stanford University.*

Diaz-Papkovich, A., Anderson-Trocmé, L., Ben-Eghan, C., and Gravel, S., 2019. UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. *PLoS genetics*, 15(11), e1008432.

Dijkstra, E. W., 2022. A note on two problems in connexion with graphs. *In Edsger Wybe Dijkstra: his life, work, and legacy* (pp. 287-290).

Donoho, D. L. and Grimes, C., 2003. Hessian eigenmaps: Locally linear embedding techniques for high dimensional data, *The Proceedings of the National Academy of Sciences (PNAS),* 100:5591-5596.

Floyd, R. W., 1962. Algorithm 97: shortest path. *Communications of the ACM*, 5(6), 345-345.

Jolliffe IT, Cadima J., 2016. Principal component analysis: a review and recent developments *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences* 374(2065):20150202. https://doi.org/10.1098/rsta.2015.0202

Kaski, S., Nikkilä, J., Oja, M., Venna, J., Törönen, P. and Castrén, E., 2003. Trustworthiness and metrics in visualizing similarity of gene expression. *BMC bioinformatics*, 4, 1-13.

Kobak, D. and Berens, P., 2019. The art of using t-SNE for single-cell transcriptomics. *Nature Communications* 10, 5416. https://doi.org/10.1038/s41467-019-13056-x.

Kobak, D. and Linderman, G. C., 2021. Initialization is critical for preserving global data structure in both t-SNE and UMAP. *Nature biotechnology*, 39(2), 156-157.

Kraemer, D., Reichstein, M., and Mahecha, M. D., 2018. dimRed and coRanking—Unifying dimensionality reduction in R. *The R Journal*, 10(1), 342–358. https://doi.org/10.32614/RJ-2018-038.

Kullback, S. and Leibler, R. A., 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1), 79-86.

Lee, J., 2010. Introduction to topological manifolds (Vol. 202). *Springer Science & Business Media*.

Lee, J. A. and Verleysen, M., 2007. Nonlinear dimensionality reduction, *Springer*.

Lee, J. A. and Verleysen, M., 2009. Quality assessment of dimensionality reduction: Rank-based criteria. *Neurocomputing.,* 72(7-9):1431–1443.

Lueks, W., Mokbel, B., Biehl, M. and Hammer, B., 2011. How to Evaluate Dimensionality Reduction? Improving the Co-ranking Matrix. arXiv:1110.3917.

McInnes, L., Healy, J. and Melville, J., 2018. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.

Mikolov, T., Chen, K., Corrado, G. and Dean, J., 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Moon, K. R., Van Dijk, D., Wang, Z., Gigante, S., Burkhardt, D. B., Chen, W. S., Yim, K., van den Elzen, A., Hirn, M. J., Coifman, R. R., Ivanova, N. B., Wolf, G., and Krishnaswamy, S., 2019. Visualizing structure and transitions in high-dimensional biological data. *Nature biotechnology*, 37(12), 1482-1492.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, É., 2011. Scikit-learn: Machine learning in Python. *The Journal of machine Learning research,* 12, 2825-2830.

Roweis, S. T. and Saul L. K., 2000. Nonlinear dimensionality reduction by locally linear embedding, *Science*, 290:2323-2326.

Tenenbaum, J. B., De Silva, V. and Langford, J.C., 2000. A Global Geometric Framework for Nonlinear Dimensionality Reduction, *Science*, 290:2319-2323.

Van Der Maaten, L. and Hinton, G., 2008. Visualizing Data using t-SNE, *Journal of Machine Learning Research*, 9 2579-2605.

Venna, J. and Kaski, S., 2006. Local multidimensional scaling. *Neural Networks*, 19(6-7), 889-899.

Venna, J., Peltonen, J., Nybo, K., Aidos, H. and Kaski, S., 2010. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *Journal of Machine Learning Research,* 11(2).

Zhang, Z., 2003. Learning metrics via discriminant kernels and multidimensional scaling: Toward expected euclidean representation. In Icml, 2,872-879.