YTÜ FEN EDEBİYAT FAKÜLTESİ

MATEMATİK BÖLÜMÜ

TOPOLOJİK VERİ ANALİZİ

Çağrı ZAMANOĞLU 19025029

Danışman: Doç.Dr. Erdal GÜL

ÖZET

21.yüzyılda yaşayan insanlar olarak yaşadığımız çağa Bilişim Çağı adını veriyoruz. Bilgi teknolojilerini kullanarak veriler yaratıyor ve bu verileri işliyor, saklıyor, değiş-tokuş ediyoruz. Günümüzde kamuya ait veya özel şirketler ve yapıların en değer verdiği şeylerden olan verilerden anlamlar ve sonuçlar çıkarmaya çalışıyoruz. Verilerin geometrik ve topolojik yapılarını anlamak bizleri bu anlam ve sonuçları çıkarmaya bir adım daha yaklaştırıyor. Bu iş için adımlarımızı atabileceğimiz çeşitli yöntemler var. Bu bitirme çalışmasında bu yöntemlerden topolojik veri analizini yöntemlerini teknik detaylara kapılmadan açıklamaya çalışacağız ve günümüzde kullanılan çeşitli uygulamalarından bahsedeceğiz.

TOPOLOJİK VERİ ANALİZİ

Verilerin topolojik ve geometrik yapısını şekillendirmeye çalışmak çözülmesi için yeni araçlara ihtiyaç duyulan zor problemler ortaya çıkarmıştır. Bu zorluğun birçok nedeni olsa da kabaca üç neden gösterilebilir.

- Veriler soyut iken topolojik ve geometrik nitelikler genellikle sürekli şekiller ile özleştirilir. Bu sebeple geometrik modeller oluşturarak verilerin şekilleri üzerinde bir yaklaşımda bulunabilir ve bu şekillerin altında yatan bilgileri ortaya çıkarabiliriz.
- Veriler tamamen tutarlı olmayabilir. Gerçek dünyadan alınan verilerde görülen veri kirliliği veya gerçekçi olmayan gözlemler şekillendirmeyi zorlaştırır.
- Veriler düşük boyutlu şekiller etrafında yoğunlaşsa bile gömülü oldukları uzayın yüksek boyutluluğu algoritmalarda ve uygulamalarda sorunlar ortaya çıkarabilir. Verinin boyutluluğu arttıkça klasik araçların işlevi azalır.

Yukarıdaki zorlukları gidermek için ortaya çıkan alanlardan biri de topolojik veri analizidir. Topolojik veri analizi uygulamalı topoloji ve hesaplamalı geometri çalışmalarından 21.yüzyılın başlarında ortaya çıkmış bir araştırma alanıdır. Topoloji ve geometrinin yardımıyla verilerin yapısı hakkında nitel ve nicel çıkarımlar yapmamıza olanak sağlar. Özellikle veri kirliliği içeren veri tabanlarında ham veriler ile çalışmak yerine şekilleriyle çalışarak daha kolay bir şekilde çıkarımlar yapılabilir.

Topolojinin ilk teoremi Leonhard Euler'in Königsberg'in Yedi Köprüsü problemi olarak kabul edilir. Königsberg 7 köprü ile birbirine bağlanan 4 kara parçası içeren bir şehirdi. Bu 7 köprünün hepsini sadece bir kez kullanmak koşuluyla tek yolculukta geçmek mümkün müydü? Bu problemin çözümünü ararken Euler günümüzün çizge teorisi ve topoloji alanlarının temelini atmıştır. Noktaların çizgiler ile birleştirilmesi ile oluşan çizgeler ile çözümü aramıştır.

Königsberg'in 7 Köprüsü Problemi [1]

Yukarıdaki çizgede her nokta bir kara parçasını, çizgiler ise köprüleri simgelemektedir. Bu çizgeden görebileceğimiz gibi tek yolculukta, her köprüyü bir kez kullanarak bütün köprülerden geçmek mümkün değildir.

Topolojik veri analizinde kabaca Euler'in Königsberg'in 7 Köprüsü probleminde uyguladığı yöntemi uyguluyoruz. Dünyamızdan bazı verileri alıyoruz (Königsberg'in haritası gibi), problemimizi çözecek bir şekil (Königsberg'in çizgesi gibi) ortaya çıkarmaya çalışıyoruz.

TVA tek bir yöntem değildir. Verilerin şekillerini ortaya çıkarmak için kullanılan yöntemlerin bir koleksiyonudur.

TOPOLOJİK VERİ ANALİZİNİN TEMEL YÖNTEMİ

TVA ile çözmeye çalıştığımız problemi şu şekilde kabaca açıklayabiliriz. Herhangi bir veri kümesini N-boyutlu bir nokta bulutu olarak düşünelim. N değişkenlerini veri kümesinin eksenleri kabul edebiliriz. Bu nokta bulutundan bir şekil oluşturabilir miyiz? Oluşturduğumuz şeklin topolojik/geometrik özelliklerini hesaplamak mümkün müdür? Bu problemi farklı TVA yöntemleri ile çözebiliriz. Genel bir şekilde şu aşamaları takip ederiz:

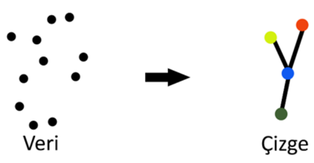
1. Üzerinde çalışılan veri tabanının aralarında belli bir uzaklık bulunan bir sonlu noktalar kümesi olduğu varsayılır. Bu uzaklık verilerin bulunduğu uzayın metriğiyle (örneğin R^d uzayı için Öklid metriği) veya oluşturulabilecek bir ikili uzaklık matrisinden tanımlanmış metrik üzerinde ifade edilir. Metriğin tanımı uygulama tarafından verilmez ise kullanıcı tarafından tanımlanır.
2. Verilerin üzerinde sürekli bir şekil oluşturularak verilerin topolojik ve geometrik özelliği ortaya çıkarılır.
3. Topolojik veya geometrik bilgiler oluşturulan şekilden ortaya çıkarılır. Verinin topolojik/geometrik şeklinin tanımlanması dışında, bu aşamanın zorluğu elde edilen şeklin veri kirliliğini ve karışıklığını önleyerek ayrışmaz ve istikrarlı bir yapı elde edebilmektir.
4. Elde edilen topolojik/geometrik bilgilerin analizi veri hakkında yeni özellikler ve yeni veri açıklayıcılar öğrenmemizi sağlar.

MAPPER ALGORİTMASI

Mapper, verileri interaktif çizgelere dönüştürmeye yarayan bir algoritmadır. Yüksek boyutlu verileri ve araştırma verilerinin analizi için kullanışlıdır. Boyut düşürme, demetleme ve çizge ağları kurarak verinin yapısını anlamayı hedefler.

Genellikle verilerin şeklini belirli bir lens üzerinden görselleştirmek, demetleri ve ilgi çekici topolojik yapıları tespit etmek, yapıları yorumlamak açısından verilerin özellikleri arasında ayrım yapabilmek için kullanılır.

Mapper algoritması Euler'in Königsberg'in 7 Köprüsü probleminde yaptığı gibi bir çizge oluşturur. Ancak köprü ve kara parçaları yerine N-boyutlu bir uzaydaki sonlu noktalar kümesi kullanılır.



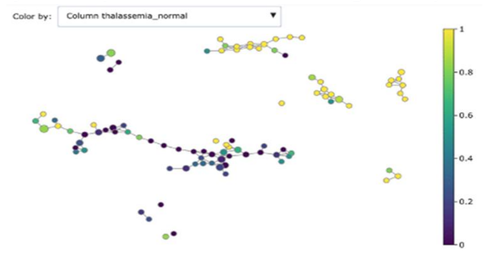
Bir metrik veya bir farklılık ölçüsü bulunan sonlu noktalardan oluşan bir veri kümesi verilsin.

1. Daha düşük boyutlu bir uzaya izdüşüm yapılması için $f: \mathbb{X} \to R$ şeklinde bir filtre fonksiyonu ya da lens kullanılır. Filtre fonksiyonu olarak genellikle Temel Bileşen Analizi kullanılır.
2. $f(\mathbb{X})$'in bir $\mathcal{U}$ örtüsü inşa edilir. Bu örtü genellikle örtüşen sabit uzunluktaki aralıkların bir kümesi formunda olur.
3. Her $U_i \in \mathcal{U}$ $(i \in I)$ için $f^{-1}(U)$ $C_{U,1}...,C_{U,n}$ şeklinde demet kümelerine ayrıştırılır.
4. Köşeleri demet kümeleri olan ve kenarları ortak nokta içeren demet kümeleri arasında çizilen bir çizge oluşturulur.

MAPPER ALGORİTMASI İLE VERİ ANALİZİNİN BİR ÖRNEK UYGULAMASI

Kalp hastalığı olasılığını yaş, kolesterol, tansiyon gibi değerleri kullanarak tahmin etmek için kullanılan Kaliforniya Irvine Üniversitesi'nin kalp hastalıkları veri tabanı üzerinde Mapper'ın uygulaması:

Hastaların kaydedilmiş maksimum nabızlarına göre oluşturulan çizge [2]



Hastaların Akdeniz Anemisi değerlerine göre oluşturulan çizge [2]

Oluşan topolojik yapıda daha açık renkli noktalar (yeşil-sarı) kalp hastalığının olasılığı yüksek, koyu renkli noktalar ise düşük olduğunu gösteriyor. Bu çizgelerden maksimum nabız değerinin düşük olduğu kişilerde kalp hastalığı olasılığının yüksek olduğu çıkarımını yapabiliriz.

DEVAMLI HOMOLOJİ

Delik, veri kirliliğine karşı dirençli bir topolojik özelliktir. Bu yüzden verileri görselleştirmek için delikler kullanılabilir. Bunu yapmak için ilk olarak n adet değişken içeren bir veri kümesi için n-boyutlu bir uzay çizilir. Daha sonra bir yarıçap belirlenir ve her veri noktası üzerinde n-boyutlu bir küre çizilir. Eğer;

- 2 küre kesişiyorsa bu kürelerin üzerine çizildiği noktalar arasında bir doğru parçası çizilir.
- 3 küre kesişiyorsa bu kürelerin üzerine çizildiği noktalar arasında bir üçgen çizilir.
- 4 küre kesişiyorsa bu kürelerin üzerine çizildiği noktalar arasında bir dört yüzlü çizilir.

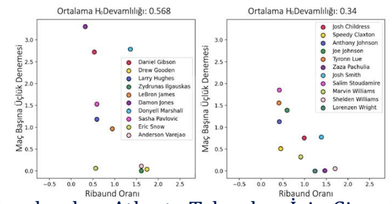
Böylece veriler görselleştirilmiş olur.

Devamlı homoloji veya ısrarlı homoloji bir uzayın topolojik özelliklerini farklı uzay ölçeklerinde hesaplamaya yarayan bir yöntemdir. Geniş yelpazeli uzay ölçeklerinde tekrar eden ısrarlı topolojik özellikler tespit edilir. Bu topolojik özelliklerin büyük olasılıkla üstünde çalışılan uzayın gerçek özelliklerini gösterdiği kabul edilir.

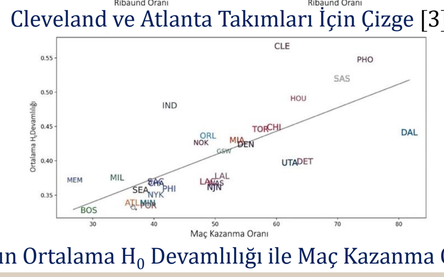
Bir uzayın devamlı homolojisini bulmak için öncelikle uzayın bir simpleks kompleksi haline getirilmesi gerekir. Üstünde çalışılan uzaydaki bir metrik simpleks kompleksinde bir filtre görevi görür ve iç içe geçmiş diziler şeklinde alt kümeler ortaya çıkarır.Böylece devamlı homoloji, delikler ve simpleks kompleksleri ile veri analizi yaparken özel bir yarıçap seçmek yerine, mümkün olabilecek her yarıçapı (örneğin r=0, ... ,∞) seçmemizi sağlar. Her yarıçap için n-boyutlu deliklerinin sayısına göre ayırt edilebilen bir simpleks kompleksi oluşturulur.

DEVAMLI HOMOLOJİ İLE VERİ ANALİZİNİN BİR ÖRNEK UYGULAMASI

NBA takımlarındaki oyuncuların 2007 sezonu istatistiklerinin tutulduğu bir veri tabanı üzerinde, oyuncuların maç başına üçlük denemesi sayısı ve ribaund oranlarının lig ortalaması ile farkı, veri noktaları olarak alındığında; H_0 yani yol bağlantılı bileşenleri gösteren homoloji grubunun yarıçapa göre devamlılığını farklılık ölçütü olarak kabul edelim.



Cleveland ve Atlanta Takımları İçin Çizge [3]



Tüm Takımların Ortalama H_0 Devamlılığı ile Maç Kazanma Oranı İlişkisi [3]

Görülebileceği üzere Atlanta (ATL) H_0 devamlılığı düşük olup az maç kazanmıştır. Ancak Cleveland (CLE) H_0 devamlılığı yüksek olup çok maç kazanmıştır ve NBA Playoffları finallerine kalmıştır. Ayrıca H_0 devamlılığı yüksek başka bir takım olan San Antonio (SAS) NBA Playoffları finallerinde şampiyon olmuştur. Yani belirlediğimiz istatistiklerdeki H_0 devamlılığının başarıyla doğru oranlı olduğunu söyleyebiliriz.

SONUÇLAR

Topolojik veri analizi, bize tıptan basketbola kadar her alanda kullanabileceğimiz, verilerin somutlaştırılmasında ve şekillendirilmesinde işe yarayan yöntemler sunar. Diğer veri analizi yöntemlerinden en büyük farkı verilere sayısal değil görsel yönden yaklaşımı kolaylaştırmasıdır.

KAYNAKÇA

[1] www.wikipedia.org
[2] www.quantmetry.com
[3] www.youtube.com/watch?v=-cfp-tH-vIM