KİMYASAL MOLEKÜLLERİN EŞLENMESİ İÇİN ÇİZGE TEMELLİ ÖRÜNTÜ TANIMA KULLANIMI

YUNUS GÖKÇER

YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ

TOBB EKONOMİ VE TEKNOLOJİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

AĞUSTOS 2015

Fen Bilimleri Enstitü onayı

_____

Prof. Dr. Osman EROĞUL
Müdür

Bu tezin Yüksek Lisans derecesinin tüm gereksinimlerini sağladığını onaylarım.

_____

Doç. Dr. Erdoğan Doğdu
Anabilim Dalı Başkanı

Yunus GÖKÇER tarafından hazırlanan KİMYASAL MOLEKÜLLERİN EŞLENMESİ İÇİN ÇİZGE TEMELLİ ÖRÜNTÜ TANIMA KULLANIMI adlı bu tezin Yüksek Lisans tezi olarak uygun olduğunu onaylarım.

Doç. Dr. M. Fatih DEMİRCİ	Yrd. Doç. Dr. Mehmet Tan
Tez Danışmanı	İkinci Tez Danışmanı

Tez Jüri Üyeleri

Başkan : Doç Dr. Osman ABUL _____

Üye : Doç. Dr. M. Fatih DEMİRCİ _____

Üye : Doç. Dr. H. Sakir BİLGE _____

TEZ BİLDİRİMİ

Tez içindeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edilerek sunulduğunu, ayrıca tez yazım kurallarına uygun olarak hazırlanan bu çalışmada orijinal olmayan her türlü kaynağa eksiksiz atıf yapıldığını bildiririm.

Yunus GÖKÇER

Üniversitesi : **TOBB Ekonomi ve Teknoloji Üniversitesi**
Enstitüsü : **Fen Bilimleri**
Anabilim Dalı : **Bilgisayar Mühendisliği**
Tez Danışmanı : **Doç. Dr. M. Fatih Demirci**
Tez Türü ve Tarihi : **Yüksek Lisans – Ağustos 2015**

YUNUS GÖKÇER

KİMYASAL MOLEKÜLLERİN EŞLENMESİ İÇİN ÇİZGE TEMELLİ ÖRÜNTÜ TANIMA KULLANIMI

ÖZET

Veri gösterimlerinin sınıflandırılmasında kullanılan örüntü tanıma teknikleri biyoenformatik ve kemoenformatik alanlarının önemli bileşenleri olarak görülürler. Kimyasal moleküllerin aktivitelerinin sonuçlarını tahmin edebilmek, laboratuvar ortamında deneyler yaparak elde edilen sonuçlara harcanan zaman ve maliyeti önemli oranda azaltmaya yardımcı olabilir. Bu çalışmada kimyasal moleküller arasındaki benzerlik oranlarını hesaplamaya odaklı bir çizge temelli örüntü tanıma metodunun kullanımı işlenmektedir. Bu metot, kimyasal moleküllerin kanserojenlik oranlarının tahmininde kullanılmaktadır. Kullanılan yöntemde moleküller kenar ağırlıklı çizgeler olarak, her atom bir düğüme karşılık gelecek şekilde ve atomların aralarında oluşturdukları bağlar da kenarlara karşılık gelecek şekilde tasvir edilmektedir. Uygulamada çizge gömme işlemi düğümlerin geometrik uzayda noktalar olarak temsil edilmesiyle gerçekleştirilir. Uzayda temsil edilen noktalar arasındaki benzerlik ölçüsü (uzaklığı) Earth Mover's Distance (EMD) metodu kullanılarak hesaplanır, öyle ki, bu metot dağıtım odaklı taşınım algoritması üzerine temellendirilmiştir. Bu çalışmada kullanılan metot Predictive Toxicology Challenge (PTC) veri seti üzerinde varolan metotlarla karşılaştırıldığında umut verici sonuçlar vermektedir.

Anahtar Kelimeler: Biyoenformatik, Örüntü tanıma, Çizge Eşleme, Sınıflandırma, Kimyasal Molekül Eşleme

University : **TOBB Economics and Technology University**
Institute : **Institute of Natural and Applied Sciences**
Science Programme : **Computer Engineering**
Supervisor : **Associate Professor Dr. M. Fatih Demirci**
Degree Awarded and Date : **M.Sc. – August 2015**

YUNUS GÖKÇER

A GRAPH-BASED PATTERN RECOGNITION FOR CHEMICAL MOLECULE MATCHING

ABSTRACT

Pattern recognition techniques that are used for classification of data representations are important components of bioinformatics and chemical informatics. Prediction of the activity of chemical molecules is a significant process that can help saving time and cost devoted to conduct the actual experiments in the laboratory. We present a new method that uses graph-based pattern recognition to compute the similarity between chemical molecules. Our method is used for prediction of the activity of chemical molecules, that is, the prediction of carcinogenicity of molecules. In our method, molecules are depicted as edge-weighted graphs, where each atom corresponds to a vertex and the bonds between the atoms are depicted as edges. The framework performs graph embedding by representing vertices as points in a geometric space. The similarity measure (distance) between the embedded points is computed using the Earth Mover's Distance (EMD) method, which is based on a distribution-based transportation algorithm. Our method shows promising results on the Predictive Toxicology Challenge (PTC) dataset compared to the existing kernels.

Keywords: Bioinformatics, Pattern Recognition, Graph Matching, Classification, Chemical Molecule Matching

TEŞEKKÜR

Çalışmalarım boyunca yardım ve katkılarıyla beni yönlendiren danışman hocam Doç. Dr. M. Fatih Demirci'ye; 2. danışman hocam Yrd. Doç. Dr. Mehmet Tan'a ve tüm desteklerinden dolayı aileme teşekkürü bir borç bilirim.

İÇİNDEKİLER

ŞEKİL LİSTESİ

Şekil **Sayfa**

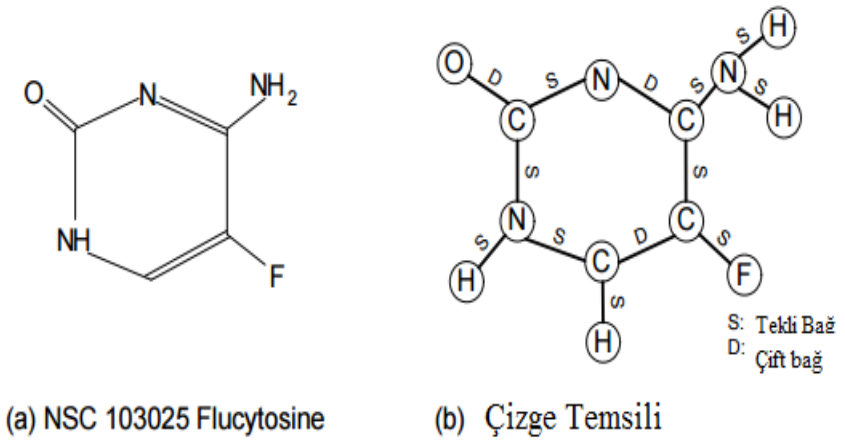
ÇİZELGE LİSTESİ

Çizelge **Sayfa**

1. GİRİŞ

Veri gösterimlerinin sınıflandırılmasında ve benzerlik oranlarının ölçümünde kullanılan örüntü tanıma teknikleri biyoenformatik ve kemoenformatik alanlarının önemli çalışma dallarından biri olarak kabul edilirler. Dahası, kimyasal moleküllerin aktivitelerinin sonuçlarını tahmin edebilmek, laboratuvar ortamında deneyler yaparak elde edilen sonuçlara harcanan zaman ve maliyeti önemli oranda azaltmaya yardımcı olabilir.

Gerçek dünyada var olan moleküler yapılar dijital dünyada çizgeler olarak kolayca temsil edilebilirler. Bunun nedeni, molekül yapılarının, çizge tasvirleri kullanılarak dijital ortamda niteliklerini kaybetmeden tasvir edilebilmeleridir. Bu durumun bir örneği Şekil 1.1'de rahatça görülebilir. Moleküler yapılar çizge olarak tasvir edildiklerinde, çeşitli çizge tabanlı algoritmalar kullanılarak moleküller arasındaki benzerlik oranlarını hesaplamak ve sınıflandırma yapmak mümkündür.



Şekil 1.1. Bir kimyasal molekülün görsel ve çizge temsili [12].

Sınıflandırma problemlerinde, eğitim setinde bulunan ve sınıf değerleri bilinen veriler yardımıyla bir model oluşturulur, daha sonra, yeni karşılaşılan veriler bu model ile karşılaştırılarak sınıflandırılmaya çalışılır. Bu tez çalışmasında son yıllarda büyük bir ün kazanan makine öğrenimi kütüphanesi olan LIBSVM kullanılmaktadır.

LIBSVM'in amacı, destek vektör makinelerinin çeşitli uygulamalarda etkili ve kolayca kullanılabilmesidir. LIBSVM kullanımı basitçe iki aşamadan oluşur: ilk olarak, eğitim seti kullanılarak bir model oluşturulması ve ikinci olarak, oluşturulan model kullanılarak test veri seti üzerinde sınıflandırma tahmininin yapılması [1].

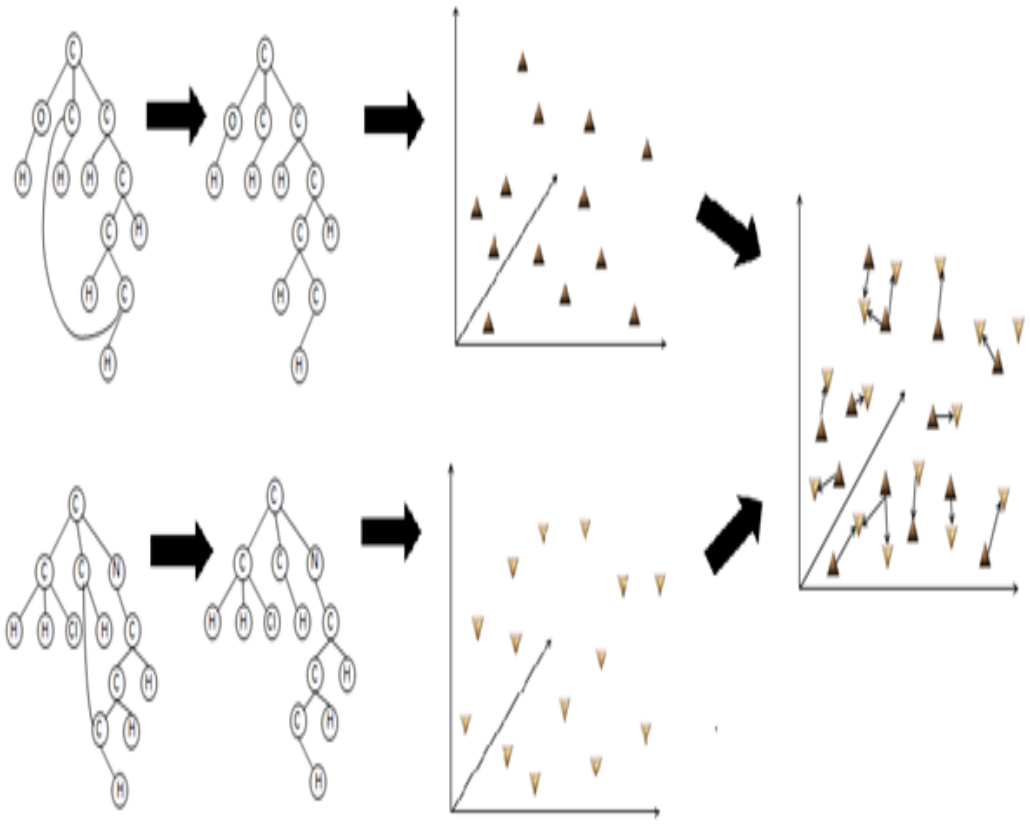
Değişken boyutlu yapısal veriler kullanılarak uygulanan kernel (çekirdek fonksiyon) algoritmaları son yıllarda makine öğreniminin önemli bir parçası haline gelmiştir. Kernel algoritmalarını basit olarak açıklamak gerekirse, iki girdi objesi u ve v ele alınsın, örnek olarak iki molekül gibi, kernel metodu k(u, v), u ve v girdileri arasındaki benzerlik oranını ölçer. Bu kernel ayrıca nitelik gömme uzayında $k(u, v) = \langle \varphi(u), \varphi(v) \rangle$ formunun iç çarpımı olarak da görülebilir [2].

Bu tezde, kimyasal yapıların aktivitelerinin sınıflandırılmasını amaçlayan ve bunu ağırlıklı çizgeler olarak tanımlanan kimyasal moleküller arasındaki benzerlikleri ölçerek yapan bir metot ortaya konmaktadır. Kimyasal moleküllerin aktivitelerinin tahmini işlemi iki ana aşamadan oluşmaktadır:

- İlk olarak, molekülleri temsil eden çizgelerin aralarındaki benzerlikleri temel alan bir metrik oluşturulması için çizge eşleme algoritmasının uygulanması.
- İkinci olarak, çizge eşleme işleminin sonuçlarını bir destek vektör makinesi kullanarak kimyasal moleküllerin aktivite sınıflarının tahmininin yapılması.

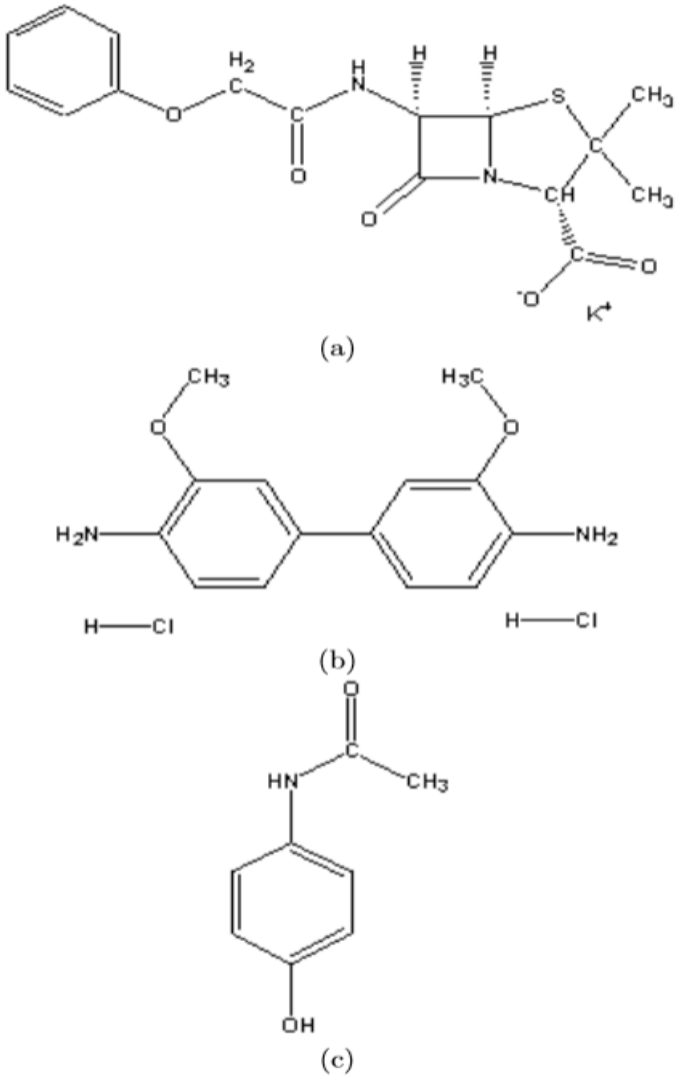
Yapılan çalışmayı kısaca özetlemek gerekirse, öncelikli olarak, kimyasal moleküllerin koordinat bilgileri dikkate alınarak Floyd Warshall algoritması kullanılır ve uzaklık matrisi oluşturulur. Daha sonra, ağaç temsili kullanılarak çizgede bulunan düğümler düşük boyutlu bir geometrik uzaydaki noktalar olarak gömülür. Earth Mover's Distance (EMD) algoritması kullanılarak çok boyutlu iki dağılım arasındaki benzerlik oranı hesaplanır. Son olarak, EMD algoritmasının oluşturduğu sonuç dikkate alınarak destek vektör makinesi kullanılır ve sınıflandırma sonuçları oluşturulur. Yapılan çalışmanın genel görünümü Şekil 1.2'de görülmektedir.

Çizge eşleme algoritması [3] ve [4]'te ortaya konan çalışmalar temel alınarak uygulanmıştır ve bu çalışmalara göre çizgedeki düğümler bir çizge gömme tekniği kullanılarak aynı geometrik uzayda tanımlanır. Çizge üzerinde atfedilen bu düğümler, geometrik uzayda ağırlıklı noktalar olarak eşlenir ve noktalar arasındaki en kısa yol uzaklıkları aralarındaki Öklid uzaklıkları dikkate alınarak hesaplanır. Gömme işleminin gerçekleşebilmesi için çizgelerin ağaç da olarak tanımlanabilir olması gereklidir. Eğer çizgeler ağaç olarak tanımlanamıyor ise çizgelerin metrik ağaç yaklaşımlarının oluşturulması için bir ön işleme basamağı gereklidir. Dolayısıyla, çizge eşleme problemi, nokta eşleme olarak yeniden düzenlenmiş olacaktır. Ağırlıklı noktaların eşlenmesi probleminin çözümü, Earth Mover's Distance (EMD) [5] algoritması kullanılarak, bir dağılımın diğerine dönüşümü için harcanan minimum maliyet göz önünde bulundurularak çözülür.



Şekil 1.2. Kullanılan algoritmanın örnek görüntüsü. Moleküller çizgeler olarak belirtildikten sonra, çizgelerin minimum örten ağaçları hesaplanır. Her ağaç geometrik uzaya gömülür ve aralarındaki uzaklık (eşleşme oranı, benzerlik) Earth Mover's Distance algoritması kullanılarak hesaplanır.

Bu çalışmada Predictive Toxicology Challenge (PTC) veri seti [6] kullanılmıştır ve bu veri setinde birkaç yüz kimyasal bileşenin dişi fareler, dişi sıçanlar, erkek fareler ve erkek sıçanlar üzerindeki toksikoloji sonuçları raporlanmıştır. Veri setinde kimyasal elementler çeşitleri, elementlerin koordinat bilgileri ve moleküler yapıların şekilleri belirtilmiştir.



Şekil 1.3. PTC veri setinde bulunan örnek moleküller. (Tüm moleküllere http://www.predictive-toxicology.org/ptc adresinden ulaşmak mümkündür.)

Tezin geri kalanında 2. bölümde literatür çalışmaları, 3. bölümde örüntü tanıma ve örüntü tanımada çizge temelli metotların kullanımı, 4. Bölümde makine öğrenmesi, 5. bölümde Earth Mover's Distance algoritması ve kullanım alanları, 6. bölümde tezde sunulan metot ve üzerine yapılan çalışmalar, 7. bölümde deneysel sonuçlar ve son olarak da 8. bölümde sonuçlandırma bulunmaktadır.

2. LİTERATÜR ÇALIŞMALARI

Geçmişte yapılan çalışmalar tarandığında, çizgelerin çeşitli niteliklerinden faydalanan birçok çizge temelli çekirdek fonksiyon önermesine rastlamak mümkündür.

[10, 11, 12]'deki çalışmalarda düzensiz gezinme metotları kullanılmıştır. [10]'da ortaya konan çalışmada tanımlanan çizge tabanlı çekirdek fonksiyonda en kısa yollar temel alınmıştır ve polinomiyal zamanda çalışır. Bu metotta iki kenar bir en kısa yolda iki kez bulunamaz, dolayısıyla suni olarak ortaya çıkabilecek yanlış yüksek benzerlik sonuçlarından kaçınılmış olunur.

[11]'de Neighborhood Subgraph Pairwise Distance Kernel (NSPD) isimli bir çekirdek fonksiyon tanımlanmıştır. Bu yöntemde çizgeler giderek büyüyen uzaklıklarda küçük yarıçaplı komşuluk alt çizgelerine dönüştürülür. Daha sonra bu alt çizgeler üzerinde tam eşleme yöntemi kullanılarak çizgeler arasındaki ayrılık hesaplanır. Bu çalışma çeşitli biyoenformatik ve kemoenformatik görevlerinde kullanılmak için uygun olduğu belirtilmiştir.

[12]'de kimyasal bileşimlerin sınıflandırılması için yaygın altyapı keşfine dayanan bir yöntem ortaya konmuştur. Bu yöntem büyük veri setleri üzerinde de iyi sonuçlar vermektedir. Bu çalışmada ortaya konan modelde öncelikle kimyasal bileşimler içinde yeterli sıklıkta rastlanan altyapılar bulunur. Tüm yaygın altyapılar keşfedildikten sonra, bu altyapılar değerlendirilerek bir sınıflandırma modeli oluşturulur.

[14]'te ortaya konan çalışmada iki çizge arasındaki ayrılık oranı tüm düzensiz gezinmelerin değerlendirilmesi sonucunda hesaplanır. Açıklamak gerekirse, bu yöntemde etiketli köşe ve kenar bilgisi bulanan çizgeler üzende düzensiz gezinme uygulanarak eş zamanlı lineer denklemlerin çözümü ile çizgeler arasındaki ayrılık oranı hesaplanır.

Ek olarak, [10]'da ortaya konan çalışmada, verilen çizgeler, düğümler arasındaki en kısa yol bilgisinin bilindiği en kısa yol çizgelerine dönüştürülerek sınıflandırma işlemi gerçekleştirilir. Tüm alt çizgelerin sayılıp dökülmesi (enumeration) işlemi NP-Complete olarak bilinmektedir[13], dolayısıyla, [12]'de sunulan çalışmada yalnızca sıklıkla rastlanan alt çizgeler tanımlayıcı olarak kullanılmıştır. [11]'de belirli bir yarıçap içerisindeki her düğümün komşulukları karşılaştırılarak bir alt çizge eşleme çekirdek fonksiyonu tanımlanmıştır

[17, 19]'daki çalışmalarda alt ağaçlar kullanılmıştır. [17]'de sınırlı uzunluğa sahip bir alt ağaç çekirdek fonksiyonu tanımlanmıştır. Bu çekirdek fonksiyon çizgeler içerisinde bulunan ve ağaç temsili olarak tanımlanan yapılardaki ortak örüntülerin ortaya çıkarılmasına dayanır. Bu çalışma belirli veri setleri üzerinde başarılı sonuçlar vermesine rağmen genel olarak ölçeklenebilirlik açısından sorunludur. Literatürdeki çalışmalarda ölçeklenebilirlik önemli bir ilgi alanı olarak ortaya çıkmaktadır, bunun nedeni de genellikle yapılan çalışmalarda, büyük çizgeler üzerinde hızlı analiz yapabilmenin önemli bir öncelik olarak kabul edilmesidir [20].

[12, 18]'de ise alt çizgeler kullanılmıştır. [18]'de çizgeler içerisinde bulunan en büyük ortak alt çizgeler bulunarak çizgeler arasındaki karşılaştırma işlemi yapılır. Burada yerel olarak en uygun örüntülerin çıkarılması yerine çizgeler arasındaki en büyük ortak alt çizgeler içerisindeki nitelikler çıkarılır. Ayrıca, [18]'de ortaya konan çalışmaya göre de enumeration işlemi yalnızca rastlanan en büyük alt çizgelere uygulanmıştır. [15]'te küçük boyutlu nitelikli çizgeler için alt çizgeler temel alınarak oluşturulan bir çekirdek fonksiyon anlatılmaktadır. Bu yöntemde çizgeler arasındaki benzerlik oranı hesaplanırken en büyük ortak alt çizgeler üzerinde benzerlik araştırması yapılması yerine, alt çizgeler arasındaki eşleşme miktarları hesaplanır. Dolayısıyla, bu yöntem polinomiyal zamanda çalışır.

[19]'da yayılma tabanlı çekirdek fonksiyonlar ailesinde yer alan Weisfeiler-Lehman çekirdek fonksiyon grubu sunulmuştur. Burada, yayılma tabanlı çekirdek fonksiyonlar ayrılık oranını hesaplarlarken yapısal bilgiyi düğüm ve kenarlar boyunca yayarlar. Bu algoritmadaki temel fikir, düğümlerin bulundurduğu etiket bilgilerinin sıralı olarak komşu düğüm bilgileriyle çoğaltılması ve çoğaltılan bu

etiket bilgilerinin sıkıştırarak yeni isimler verilmesine dayanmaktadır. Bu işlem düğüm etiketleri farklılık gösterene kadar ya da önceden belirlenen sabit bir miktarda tekrarlanır. Yakın zamanda ortaya konan bir çalışma olan [21]'de yayılma tabanlı çekirdek fonksiyonlar ele alınmış ve hesaplama gereksinimlerini azaltmak için bölgesel olarak hassas kırpma (hashing) işlemi uygulanmıştır. Yine yakın zamanda sunulan bir çalışma olan [16]'da akımlar üzerinde çizge sınıflandırması işlemi üzerinde çalışılmış ve veri setinin devamlı olarak büyümesi nedeniyle tek bir nitelik uzayında çizgelerin eşleştirilmesinin zorluğu incelenmiştir.

3. ÖRÜNTÜ TANIMADA ÇİZGE TEMELLİ TEKNİKLERİN KULLANIMI

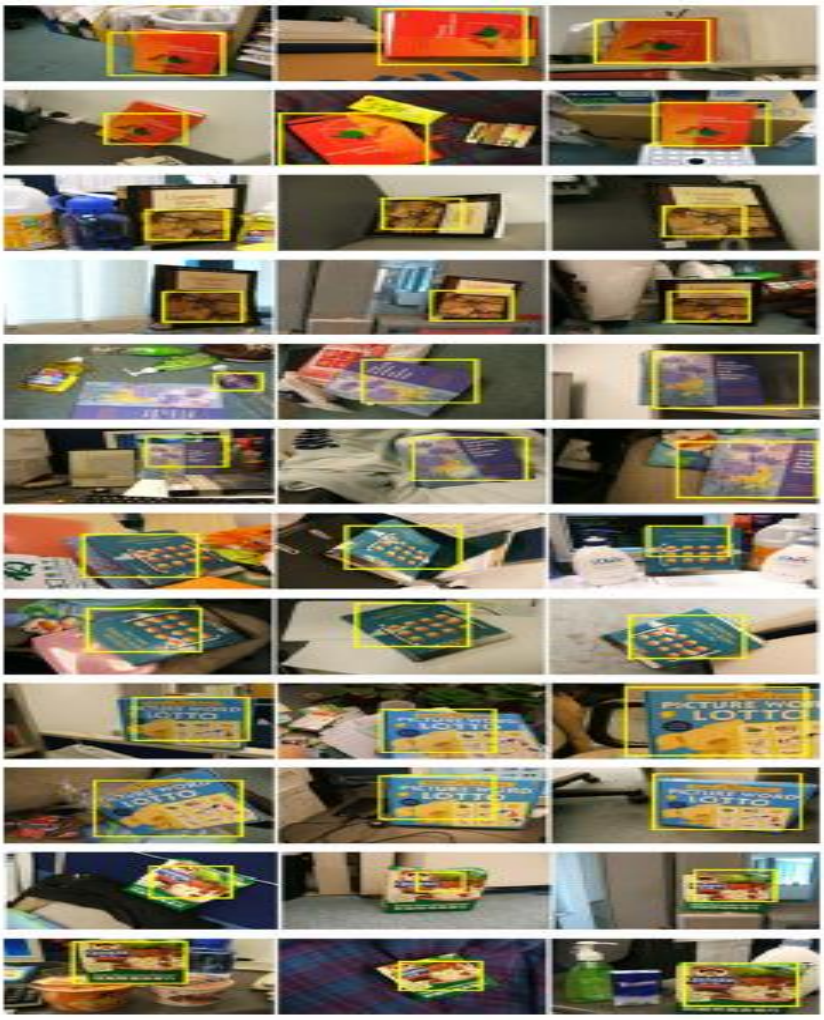
Örüntü tanıma işlemi makine öğrenmesi alanının içinde değerlendirilen, veriler içerisindeki düzenlilik ve kalıpların bulunması işlemi olarak tanımlanır. Bir nesne için verilen ölçümlerin değerlendirilmesi sonucu nesneyi doğru sınıfa yerleştirme işlemine örüntü tanıma denir. Bu işlem genellikle bir bilgisayar yardımıyla otomatik olarak gerçekleştirilir. Tanıma işleminin uygulanacağı nesneler, nesnelerle ilgili ölçümler ve bu nesnelerin ait olduğu sınıflar dünyada var olan birçok şey arasından seçilebilir. Bu nedenle birbirinden farklı birçok örüntü tanıma görevi bulunmaktadır. Bazı örüntü tanıma görevleri günlük olaylar arasından seçilebilir, örnek olarak konuşma tanıma işlemi gösterilebilir, diğer örüntü tanıma görevleri de daha genel görevler olabilirler.

Bazı örüntü tanıma işlemleri insanlar için kolay ve önemsiz gibi görünse de, bu durum ilgili örüntü tanıma işleminin basit olduğunu göstermez. Örüntü tanıma işleminin uygulanmasının nedeni, verilen parametreler göz önünde bulundurularak otomatikleşmiş bir tanıma işlemi veya otomatikleşmiş bir karar verme mekanizması ihtiyacı nedeniyledir. Yaklaşık yarım yüzyıldır yapılan çalışmalara rağmen, örüntü tanımanın kullanılabileceği yeni uygulama alanlarının ortaya çıkması ve henüz çözülmemiş birçok problem olması nedeniyle örüntü tanıma aktif bir araştırma alanı olmaya devam etmektedir [34]. Şekil 3.1'de bir örüntü tanıma örneği gösterilmektedir.

3.1 Örüntü Tanımada Genel Olarak Kullanılan Terminoloji

- Örüntü tanımada bir nesne ile ilgili karakteristik ve ayırıcı özellikler taşıyan değişkene nitelik denir.
- Nitelikler genellikle örüntü tanıma objeleriyle ile ilgili ölçüm ve değerlendirmeler barındırırlar.
- d elemanlı nitelikler setinin anlamlı bir şekilde yer aldığı d boyutlu kolon vektörüne nitelik vektörü adı verilir. Nitelik vektörü tanımlanmak istenen nesnenin imzasını temsil eder.

- Nitelik vektörünü içinde barındıran *d* boyutlu uzaya nitelik uzayı denir.
- Bir nesnenin sahip olduğu nitelik seti ve sınıf bilgisi bütününe örüntü denir.
- Verilen bir nesnenin içinde bulunduğu kategoriye sınıf adı verilir.
- Eğitim, örüntü tanıma sisteminin nitelik vektörleri ile bu vektörlere ilişkin sınıf/etiket bilgileri arasındaki bağlantıyı öğrendiği aşamadır. Bu aşama sonrasında *d* boyutlu nitelik uzayında oluşan karar sınırı farklı sınıflara ait örüntülerin ayrıştırılmasında kullanılır.



Şekil 3.1. Çeşitli nesneler arasında örüntü tanıma işleminin gerçekleştirilmesi örneği. Aynı satırda bulunan sarı çerçeve içerisinde yer alan örnekler birbirleriyle başarılı bir şekilde eşlenmiştir.

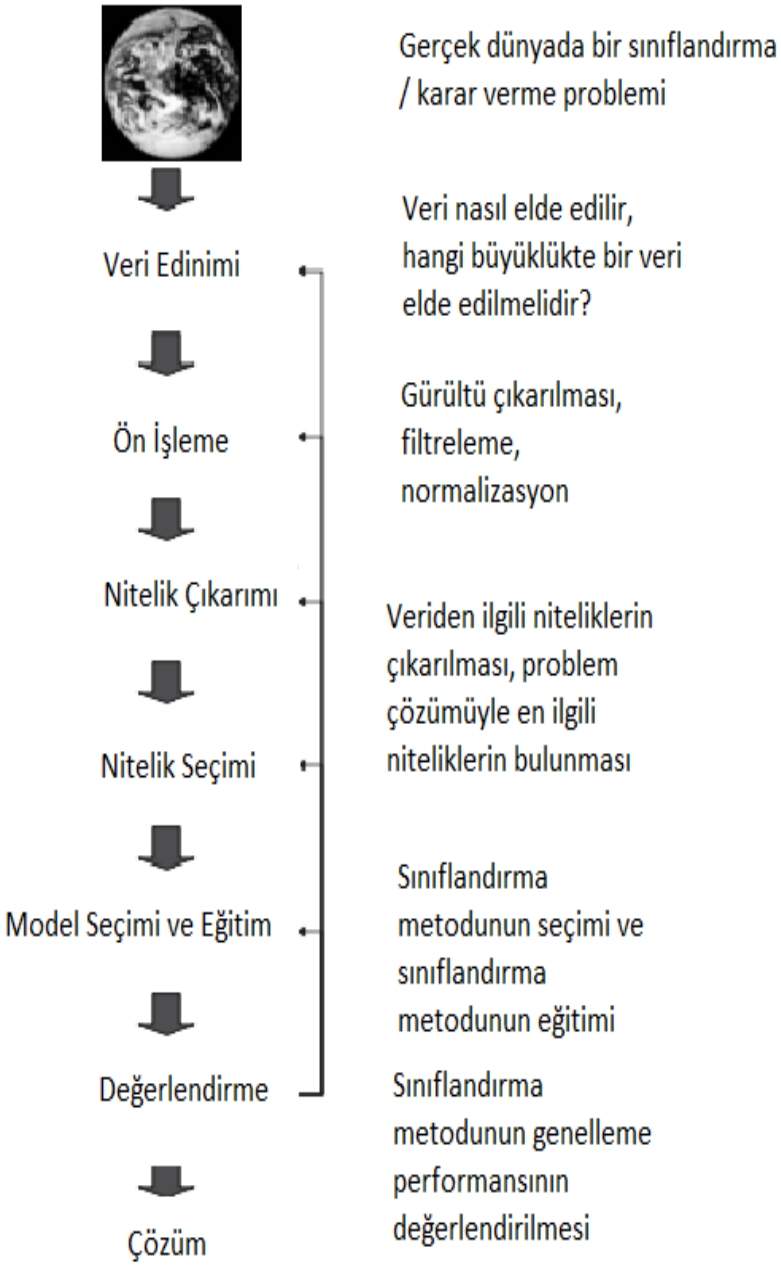
3.2 Örüntü Tanıma Bileşenleri

Bazı örüntü tanıma işlemleri insanlar için kolay ve önemsiz gibi görünse de, bu durum ilgili örüntü tanıma işleminin basit olduğunu göstermez. Örüntü tanıma işleminin uygulanmasının nedeni, verilen parametreler göz önünde bulundurularak otomatikleşmiş bir tanıma işlemi veya otomatikleşmiş bir karar verme mekanizması ihtiyacı nedeniyledir. Yaklaşık yarım yüzyıldır yapılan çalışmalara rağmen, örüntü tanımanın kullanılabileceği yeni uygulama alanlarının ortaya çıkması ve henüz çözülmemiş birçok problem olması nedeniyle örüntü tanıma aktif bir araştırma alanı olmaya devam etmektedir [34].

Genellikle, örüntü tanıma konusu incelenirken yalnızca sınıflandırıcı modeli ve eğitim algoritması göz önünde bulundurulur. Ancak, bütün bir örüntü tanıma sisteminde birçok farklı bölümden söz edilebilir. Bu bölümler veri edinimi, ön işleme, nitelik çıkarımı, nitelik seçimi, model seçimi ve eğitim, değerlendirme kategorileridir. Bu bölümler Şekil 3.2'de gösterilmektedir ve bu bölüm içerisinde açıklanmaktadırlar.

3.2.1 Veri Edinimi

Uygun ve başarılı bir sınıflandırma algoritmasının haricinde, başarılı bir örüntü tanıma sisteminde olması gereken özelliklerden biri de yeterli ve temsil değeri yüksek eğitim ve test veri setleri kullanımıdır. Temsil değerinin yüksek olması, nitelik vektörleri ile doğru sınıf bilgilerinin eşleştirilmesi için gerekli olan başarılı karar sınırının oluşturulmasına olumlu olarak yardım etmektedir. Karar sınırının oluşturulması aşamasında hangi büyüklükte bir veri setine ihtiyaç olduğuna dair kesin bir sayı bulunmamaktadır.



Şekil 3.2. Örüntü tanıma bileşenleri.

3.2.2 Ön İşleme

Ön işleme aşamasındaki temel amaç edinilen veri seti içerisindeki gürültünün olabildiğince temizlenmesi işlemidir. Bu aşama, bazı durumlarda göz ardı edilmesine rağmen önemli bir aşama olarak kabul edilir. Gürültünün nedenleri ve kaynakları ile ilgili yeterli bilgiye ulaşılabiliyor ise veri seti üzerinde çeşitli filtreleme teknikleri kullanılabilir. Ayrıca, veri seti üzerinde normalizasyon işlemi uygulanarak veri setinin sahip olduğu genişlik küçülttürülebilir ve bu sayede sınıflandırıcının performansı arttırılabilir. Ön işleme sırasında aykırı noktaların çıkarılması işlemi yapılarak da sınıflandırıcının daha başarılı sonuçlar vermesi sağlanabilir.

3.2.3 Nitelik Çıkarımı

Nitelik çıkarımı işlemi, sınıflandırma için gerekli olan, yeterli miktarda öğretici ve ayırıcı niteliklerin bulunması işlemidir. Çok boyutluluğun azaltılması örüntü tanıma sisteminin başarısını arttıran bir etmendir. Bu nedenle tercihen az ama yeterli miktarda etkili niteliklerin çıkarımı önemlidir. Ayrıca, küçük ve yeterli bir nitelik seti sınıflandırma algoritmasının zaman ve bellek gereksinimlerini düşürmekte, karmaşıklığın azalmasına yardımcı olmakta ve ‘overfitting’ ihtimalini azaltmaktadır.

3.2.4 Nitelik Seçimi

Nitelik seçimi işlemi, nitelik çıkarımı algoritmasının uygulanması sonucu ortaya çıkan nitelikler arasından bir alt set seçimi işlemidir. Çıkarılan nitelikler kümesinden sınıflandırma aşamasında kullanılacak olanlar seçilirken önemli olan en iyi ayrımı yapacak niteliklerin seçilmesidir. Bu sayede, seçilen nitelikler ile eğitilen sınıflandırıcının iyi bir genelleme performansına sahip olması sağlanabilir.

3.2.5 Model Seçimi ve Eğitim

Daha önce anlatılan aşamaların tamamlanmasından sonra sınıflandırma modelinin (sınıflandırıcı) seçimi yapılabilir. Bu aşama örüntü tanıma sistemleri için en önemli aşama olarak görülebilir. Sınıflandırma metodu veri setinin yapısı, ortaya çıkarılmak

istenen sonuçlar, sınıflandırma metotlarının veri setine uygunluğu göz edilerek seçilebilirler. Sınıflandırma algoritması seçildikten sonra bu algoritma bir eğitim seti üzerinde çalıştırılarak bir model oluşturulabilir. Bu model, bir sonraki aşamada test seti üzerinde denenerek istenen örüntü tanıma işlemi gerçekleştirilebilir.

3.2.6 Değerlendirme

Bir model oluşturulduktan, eğitim sona erdikten sonra, modelin genelleme performansının ölçülmesi amacıyla model bir test seti üzerinde çalıştırılır. Burada kullanılan test setindeki veriler modelin daha önce karşılaşmadığı verilerdir. Ortaya çıkan sonuç algoritmanın performansını gözler önüne serer. Genelleme performansının ölçümünde kullanılan en yaygın yöntem bir veri setini eğitim ve test setleri olarak iki parçaya bölmek, eğitim setini modeli oluşturmak için ve test setini de performans ölçümünü yapmak için kullanmaktır.

3.3 Örüntü Tanımada Kullanılan Çizge Temelli Teknikler

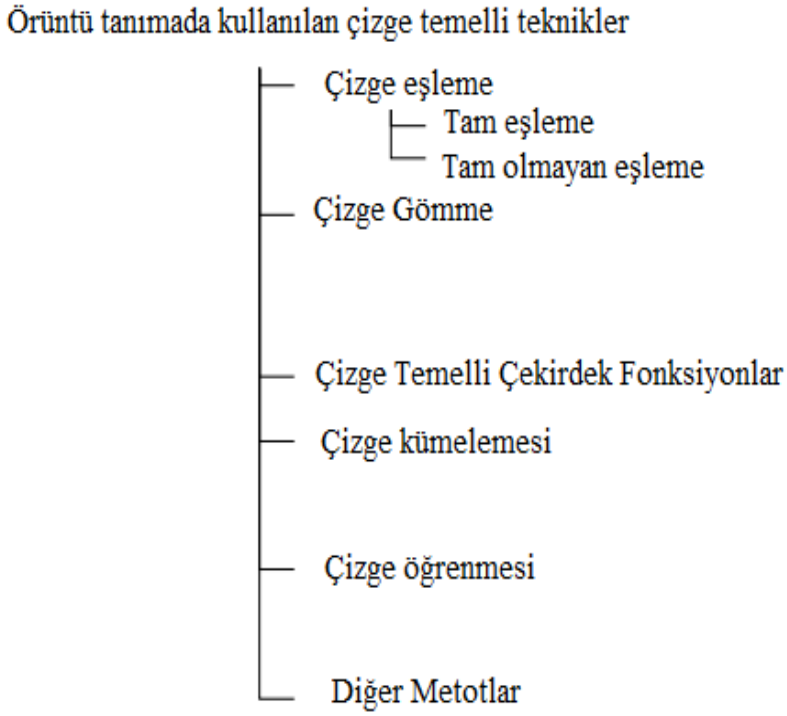
Biyoenformatik, kemoenformatik, ilaç keşfi, web veri madenciliği, sosyal ağlar gibi alanlarda var olan yapılar doğal yollarla çizgeler olarak tanımlanabilmektedirler. Dolayısıyla, çizge temelli yapılar incelenerek ve bu yapılarda yer alan örüntüler üzerinde çalışılarak birçok alanda var olan problemlerin çözümlerine dair önemli bir katkı yapılabilir. Çizgelerin düğümleri objeleri, kenarları da objeler arasındaki ilişkileri temsil edebilirler [39].

Şekil 3.3. Çizge temsilleri üzerinde basit bir örüntü tanıma örneği. G ve G' çizgelerinin benzerlikleri kırmızı renk ile gösterilmiştir.

Çizge temelli tekniklerin başarılı bir şekilde uygulandığı altı önemli başlık şu şekilde sıralanabilir [27]:

- Doküman işleme
- Biometrik tanımlama
- Görüntü veri tabanları
- Video analizi.
- Biyolojik ve biyomedikal uygulamalar
- 2D ve 3D görüntü işleme ve analizi.

Yetmişlerin sonundan itibaren, çizge temelli teknikler örüntü temsili ve sınıflandırma uygulamalarında yer almaya başladı [41]. Yakın zamanda, örüntü tanımada çizgelerin kullanımında bir artış yaşandı. Bunun nedeni, çizge temelli algoritmaların hesaplama maliyetinin halen birçok durumda yüksek olmasına rağmen hesaplama gücünün diğer yeni yöntemlerle karşılaştırılabilir olmasıdır [26].



Şekil 3.4. Örüntü tanımada kullanılan çizge temelli teknikler şeması.

İki çizgenin karşılaştırması işleminde kesin eşleşme için gerekli olan bağlayıcı kısıtlar bazı durumlarda çok katı olabilmektedir. Çoğu uygulamada, gözlenen çizgeler bazı nedenlerden dolayı deformasyona sahip olabilmektedirler. Dolayısıyla, eşleme işleminin toleransa sahip olması önemlidir, bu da bazı kısıtların gevşetilmesi yoluyla küçük farklılıklara uyum sağlanarak gerçekleştirilir. Eşleme işlemi sırasında herhangi bir deformasyon yoksa veya beklenmiyorsa bile toleransa sahip olmak önemlidir. Bu nedenle, bazı kesin olmayan eşleme metotları eşleme maliyetini benzeşmezlik/ayrılık oranı olarak tanımlarlar.

3.3.1 Çizgeler Üzerinde Örüntü Eşleme Problemi

İki objenin birbiriyle karşılaştırılması ya da bir modelin bir objeyle karşılaştırılması işleminde, objeleri oluşturan yapısal veriler çizgeler olarak tanımlandıklarında, karşılaştırma işleminde bir tür çizge eşleme yönteminin uygulanması gerekmektedir. Eşleme işlemi temel olarak, iki çizgenin sahip olduğu düğüm ve kenarlar arasında, bazı kısıtlara uyan benzer yapıların birbirleriyle eşlenmesi işlemidir. Dahası, çizge eşleme metotları yapıları itibarıyla ikiye ayrılmaktadırlar [26]:

Tam Eşleme:

Tam eşleme metotları iki obje arasında eşleme işlemini gerçekleştirirken objeler veya alt yapıları arasında katı bir uyuşma aramaktadırlar. Tam eşlemede iki çizgenin düğümleri arasındaki eşleştirme işleminde kenarlar korunmalıdır, açıklamak gerekirse, eğer ilk çizgedeki iki düğüm bir kenar ile bağlı ise eşleştirme işlemi yapılan ikinci çizgede de bu düğümler bir kenar ile bağlı olmak durumundadırlar.

Tam Olmayan Eşleme:

Tam eşlemede ortaya çıkan katı kısıtlamalar bazı durumlarda çizgelerin karşılaştırılması için fazla sert kalabilmektedir. Tam olmayan eşleme metotları iki obje arasında eşleme işlemini gerçekleştirirken çizgeler arasında kabul edilebilir düzeyde bazı yapısal farklılıklar olsa dahi eşleme işlemini gerçekleştirebilirler.

Çizgeler üzerinde örüntü tanıma işlemi problemi şu şekilde tanımlanabilir [44]:

Bir çizge G = (V, E), V köşelerinden ve E kenarlarından oluşsun. Her e ∈ E (v_i, v_j) , $v_i, v_j \in$ V arasında yer alır.

Örüntü sorgusu P = (E_p, E_p) olarak tanımlanır. Bu örüntü sorgusu G ve P arasındaki örüntü tanıma işleminin gerçekleşmesi için G çizgesinin sahip olması gereken yapısal ve semantik gereklilikleri temsil etmektedir.

Buradaki amaç G çizgesi içerisinde bulunan, P örüntüsüyle eşleşen, M altçizge setini bulabilmektir.

Örüntü tanımada çizge yapıları kullanılmak istenirse, oluşturulan çizgelerin işlenmesinin nasıl yapılacağı sorusu ortaya çıkmaktadır. Bunun nedeni, çizgelerin tanımlandığı uzayda örüntü tanıma metotlarının uygulanabileceği matematiksel özelliklerin bulunmamasıdır [42]. Bu durumda çizge gömme işlemi devreye girmektedir.

3.3.2 Çizge Gömme

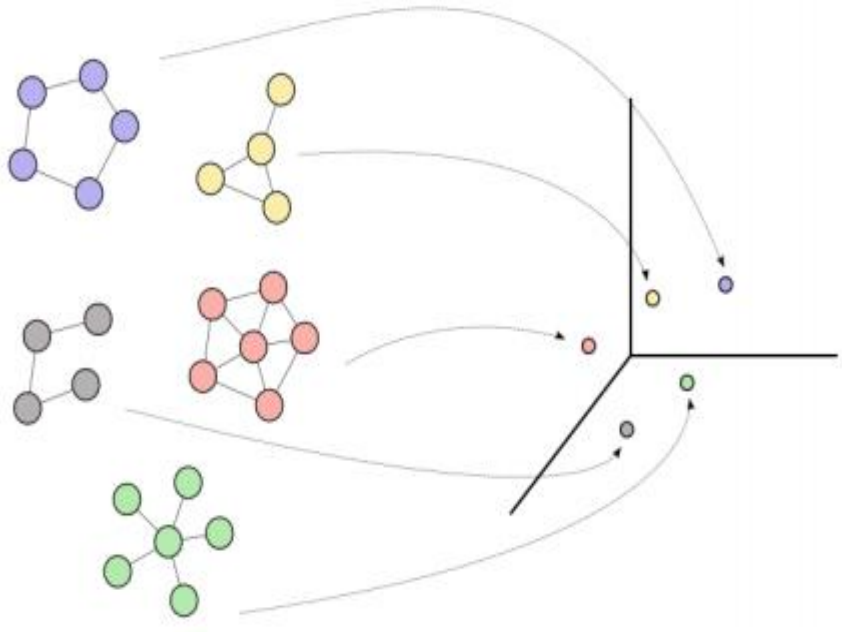
Temel olarak çizge gömme, çizgelerin yapısal özelliklerini koruyacak şekilde vektörel uzayda yeniden tanımlanması işlemidir.

Çizge gömme teknikleri çizgeleri yüksek boyutlu uzaylara gömerek çeşitli örüntü tanıma metotlarının uygulanması için gerekli olan matematiksel işlemlerin yapılabilmesini sağlamaktadırlar. Bu nedenle, çizge gömme işlemi, çizge eşleme ve sınıflandırma problemleri için önemli bir kavram olarak değerlendirilmektedir.

Çizge gömme denince iki farklı kavram ortaya çıkmaktadır:

- Çizge içerisindeki yapısal olarak benzer yapıya sahip olan düğümlerin vektör uzayında birbirlerine yakın noktalara gömülmesi
- Çizgelerin tümünün vektör uzayında noktalar olarak gömülmesi ve benzer çizgelerin vektör uzayında yakın noktalara gömülmesi

Çizge eşleme metotları örtülü çizge gömme ve belirgin çizge gömme metotları olarak ikiye ayrılırlar. Örtülü çizge gömme metotları çizge temelli çekirdek fonksiyonları üzerine kuruludur. Çizgeler vektör uzayına gömülmek yerine, çekirdek fonksiyon işlemleri hali hazırda var olan çizge uzayında yapılır. Örtülü çizge gömme işleminde nokta çarpımının tüm özellikleri karşılanır. Ayrıca, örtülü çizge gömme işleminde vektör uzayına gömme yapılmadığı için, vektör uzayında gerçekleştirilebilecek olan bazı işlemler uygulanamazlar.



Şekil 3.5. Çizge gömme işlemi örneği. Bu örnekte çizgeler renk bilgileriyle vektör domaininde temsil ediliyorlar.

Belirgin çizge gömme metotları verilen çizgeleri nitelik uzayında açıkça yeniden temsil eder (gömer) ve bu sayede vektör uzaylarında kullanılabilecek olan yöntemlerin gömülen çizgeler üzerinde kullanılmaları için zemin hazırlarlar. Belirgin çizge gömme metotlarının önemli bir özelliği, farklı büyüklüğe ve düzene sahip çizgelerin önceden kararlaştırılmış büyüklükteki bir nitelik vektörüne gömülmesidir.

4. MAKİNE ÖĞRENMESİ

Makine öğrenmesinin temeli çeşitli algoritmalar ve veriler kullanarak makinelerin davranış ve tahminler oluşturması ve karşılaşılan verilere göre gelişim gösterebilmesi işlemidir. Buradaki amaç bir görevi tamamlamak, doğru tahminler yapabilmek veya akıllı davranış sergilemek olabilir. Yapılan öğrenme işlemi her zaman gözlem, veri, örnek veya talimatlar temel alınarak gerçekleştirilir. Dolayısıyla, genel olarak makine öğrenimi geçmiş deneyimler kullanılarak gelecekte daha iyi davranabilme ile ilgilidir. Makine öğrenmesinde önemli olan nokta otomatik metotlar yaratabilmektir; bir başka deyişle, öğrenme algoritmaları insan yardım ve müdahalesine ihtiyaç duymadan otomatik olarak öğrenme işlemini gerçekleştirebilmelidirler.

Makine öğrenmesinin temel amacı genel amaçlı kullanıma uygun kullanışlı algoritmalar yaratmaktır. Bu algoritmalar verimli olmalı, dolayısıyla zaman ve alan verimini göz önünde bulundurmalıdırlar. Öğrenme konusunda önemli olan bir nokta da öğrenme işleminin yapılabilmesi için hangi büyüklükte bir veriye ihtiyaç olduğudur [28].

Makine öğrenmesinin genel problemi potansiyel hipotezler uzayında verilen bilgiye en iyi uyacak olan hipotezin seçilmesidir. Kullanılan veri etiketlenmiş, yani sınıf bilgisine sahip veya etiketlenmemiş olabilir; eğer kullanılan veri etiketli ise problem danışmanlı öğrenme problemidir. Eğer etiketler kategorik ise problem sınıflandırma problemi, eğer etiketler gerçek değerli ise problem regresyon problemidir. Kullanılan veri etiketlenmemiş ise problem danışmansız öğrenme problemidir ve amaç sınıf bilgisi bulunmayan veri içerisindeki farklı grupları keşfetmektir.

Kimyasal bileşimler genellikle çizgeler şeklinde temsil edilebilirler ve bu nedenle de kemoenformatik alanında değişken boyutlu çizgeler üzerinde çeşitli makine öğrenmesi metotları kullanılabilmektedir. Çizgelerin kullanılmasıyla ortaya çıkan yapısal verilerde makine öğrenmesi metotlarının kullanılması veri içerisindeki anlamı, düzenlilikleri ve örüntüleri ortaya çıkarmada önemlidir. Dolayısıyla, son

yıllarda değişken uzunluklu yapısal veriler üzerinde çekirdek fonksiyonların kullanımı makine öğreniminde önemli bir faktör olarak ortaya çıkmıştır [38].

Makine öğrenmesiyle ilgili olarak birçok örnek verilebilir. Bunlardan bazıları şu şekilde sıralanabilir:

- Optik Karakter Tanıma
- Yüz tespiti
- Spam filtreleme
- Metinlerde konu belirleme
- Konuşma algılama
- Medikal tanı koyma
- Müşteri sınıflandırma
- Dolandırıcılık tespiti
- Hava durumu tahmini

Makine öğrenmesinde çok sayıda ve farklı özelliklere sahip öğrenme algoritmaları mevcuttur, bu nedenle problem için uygun bir öğrenme algoritması seçmek önemlidir.

4.1 Makine Öğrenme Algoritmalarının Temel Bileşenleri

Çok sayıda öğrenme algoritması olmasına rağmen bu algoritmaları üç bileşende tanımlamak mümkündür, temsil, değerlendirme ve optimizasyon. Bu bileşenler kısaca şöyle tanımlanabilirler [29]:

4.1.1 Temsil:

Öğrenici bilgisayar sisteminin üstesinden gelebileceği biçimsel (formal) bir dilde temsil edilir. Bir diğer açıdan, öğrenici için bir temsil seçilmesi işlemi, öğrenicinin öğrenebilmesi mümkün olan sınıflandırıcı setinin seçilmesi işlemiyle aynıdır. Bu sınıflandırıcı setine öğrenicinin hipotez uzayı denmektedir. Hipotez uzayında bulunmayan bir sınıflandırıcı algoritma tarafından öğrenilemez.

4.1.2 Değerlendirme:

İyi sınıflandırıcıları kötü olanlardan ayırmak için bir değerlendirme fonksiyonu gereklidir. Algoritma tarafından kullanılan değerlendirme fonksiyonu ile sınıflandırıcı tarafından kullanılan ve optimize edilmesi istenen değerlendirme fonksiyonu farklılık gösterebilirler.

4.1.3 Optimizasyon:

Sınıflandırıcılar arasında en iyi sonucu ortaya koyanın belirlenmesi için bir metot gereklidir ve bu metot öğrenicinin verimli olması için önemli bir rol oynar. Genellikle öğreniciler standart optimizasyon niteliklerini kullanarak başlarlar ve daha sonra en iyi sonucu verecek şekilde uyarlanmış optimizasyon niteliklerine geçiş yaparlar.

4.2 Makine Öğrenmesinde Kullanılan Algoritma Yapıları

Makine öğrenmesinde kullanılan algoritmalar yapılarına göre şu gruplara ayrılabilirler [31]:

4.2.1 Danışmanlı (Denetimli) Öğrenme:

Öğrenme algoritmasının yarattığı fonksiyon verilen girdileri istenen sonuçlar ile eşler. Danışmanlı öğrenme işleminin klasik bir örneği sınıflandırma problemidir. Öğrenme algoritması (öğrenici) girdi verileri ve sınıf verileri arasındaki bağları ortaya çıkarır.

4.2.2 Danışmansız (Denetimsiz) Öğrenme:

Denetimli öğrenme işleminin aksine etiketli örnekler olmadan bir model oluşturulması işlemidir. Amaç, ilgi çekici yapıları ortaya çıkarmaktır, bilgi keşfi olarak da adlandırılabilmektedir.

4.2.3 Yarı Denetimli Öğrenme:

Etiketli ve etiketsiz örnekler birlikte kullanılarak bir sınıflandırma fonksiyonu yaratılması işlemidir.

4.2.4 Güçlendirici Öğrenme:

Belirtilen dünya içerisinde ortaya konan gözlemlere dayanarak yapılan öğrenme işlemidir. Ortaya konan her davranışın çevrede bir etkisi vardır ve çevre buna bağlı olarak öğrenme algoritmasını yönlendirici geri bildirimler verir.

4.2.5 Uyum Sağlama ile Öğrenme:

Danışmanlı öğrenmeye benzer bir yapıda olmasına karşın açıkça bir sınıflandırma fonksiyonu oluşturmak yerine sonuçları öğrenme girdileri, öğrenme çıktıları ve karşılaşılan yeni verilere göre belirler.

4.2.6 Öğrenmek için Öğrenme:

Öğrenme algoritması geçmiş deneyimlere dayanarak kendi tümevarımsal çıkarımlarını oluşturur.

Bu tez çalışmasında danışmanlı (denetimli) öğrenme teknikleri kullanıldığı için diğer öğrenme gruplarına değinilmemiş olup danışmanlı öğrenme konusu bir sonraki bölümde detaylı bir şekilde işlenmektedir.

4.3 Danışmanlı (Denetimli) Öğrenme Açıklaması

Danışmanlı öğrenme, sınıf bilgisine sahip bir eğitim seti kullanılarak genel bir hipotez oluşturulması ve bu hipoteze bağlı olarak gelecek veriler için sınıf tahmini yapma işlemidir. Danışmanlı öğrenme metotlarının hedefi girdi verileri ve sınıf verileri arasındaki bağları ortaya çıkarmaktır. Ortaya çıkan bağlar model adı verilen

yapıda temsil edilir ve modellerin amacı olayları tanımlamak ve açıklamak olup hedef niteliklerin sınıflarının tahminini yapmak için kullanılırlar.

Danışmanlı öğrenme örnekler aracılığıyla öğrenme işlemidir ve bu işlem sırasında eğitim ve test aşamaları için iki veri seti kullanılır. Eğitim aşamasındaki ana fikir algoritmanın eğitim setinde bulunan etiketli örnekleri kullanarak öğrenme işlemini gerçekleştirmesi ve test seti içerisinde bulunan veriler üzerinde en iyi muhtemel tahminleri yapabilmesidir. Örnek olarak, bir eğitim seti elma ve armut görsellerine ve bu görsellerin hangi sınıfa ait olduğunu belirten (elma, armut) etiketlere sahip olsun. Test seti de yine aynı gruba ait, etiketlenmemiş meyvelerden oluşur ve amaç test setinde bulunan etiketsiz meyveleri doğru bir şekilde (elma veya armut olarak) sınıflandıracak bir kural oluşturmaktır.

Danışmanlı öğrenmede eğitim seti n elemanlı (x_1, y_1), (x_2, y_2), ... (x_n, y_n) ikililerinden oluşmaktadır ve her x_i bir veri için sahip olunan ölçümü veya ölçümleri temsil eder, her y_i de bu verinin etiketi olarak tanımlanır.

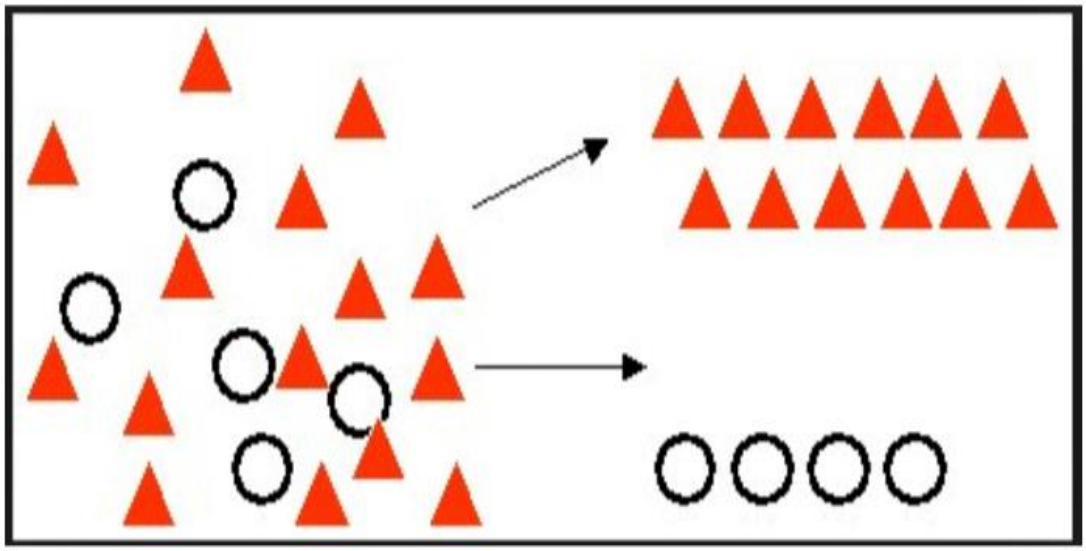
Danışmanlı öğrenmede test seti de m elemanlı $(x_{n+1}, x_{n+2}, \ldots, x_{n+m})$ ölçümlerden oluşur ve etiket bilgisi yoktur. Daha önce belirtildiği gibi amaç test setinde bulunan veriler için olabildiğince doğru etiket tahminleri yapabilmektir.

4.4 Sınıflandırmanın Temelleri

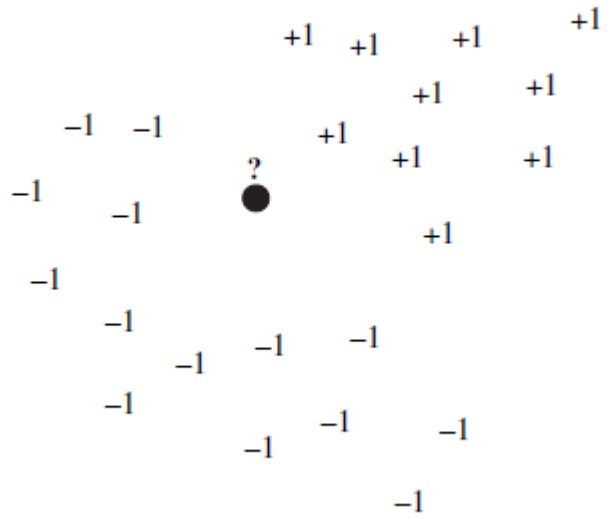
Verilen bir verinin ait olduğu sınıfı doğru tahmin etmeye çalışma işlemine sınıflandırma denir. Verinin sınıfının belirlenebilmesi için, bir algoritmanın, çeşitli nitelik ve sınıflandırma sonuçlarına sahip eğitim setini işlemesi, daha sonra da sınıflandırma tahmininde kullanılabilecek olan nitelikler arasındaki ilişkileri keşfetmesi gerekmektedir. Daha sonra, algoritma önceden işlemediği bir test veri seti üzerinde çalışarak bu veri setinin içerisindeki verileri sınıflandırmaya çalışmalıdır. Bu işlemlerin sonucunda, algoritma test setinde bulunan verilerin sınıf tahminlerini üretir. Tahminlerin doğruluk oranı algoritmanın başarı oranıyla doğru orantılıdır.

Sınıflandırma işleminde amaç x girdilerini y sonuçlarıyla eşlemektir, burada, $y \in \{1, ..., C\}$, ve C sınıf sayısı olarak alınır. Eğer C = 2 ise, sınıflandırma işlemi ikili sınıflandırma, öyle ki $y \in \{0, 1\}$; ya da, C > 2 ise, sınıflandırma işlemi çok sınıflı sınıflandırma işlemi olarak adlandırılır. Eğer sınıf etiketleri birbirini dışlayan şekilde değil ise, yani bir eleman birden çok etikete sahip olabiliyor ise, sınıflandırma işlemi çok etiketli sınıflandırma işlemi olarak adlandırılır.

Sınıflandırma işlemini formüle etmenin yolu sınıflandırma problemini fonksiyon yaklaştırım işlemi olarak görmektir. Açıklamak gerekirse, bir f fonksiyonu için y = f(x) olduğu ve öğrenme işlemindeki amacın da verilen etiketli bir eğitim seti ve $\hat{y} = \hat{f}(x)$ kullanılarak f fonksiyonuna yaklaşmak olduğu kabul edilebilir. Buradaki ana amaç farklı girdi verileri için doğru sınıflandırma tahminleri yapabilmektir [30].



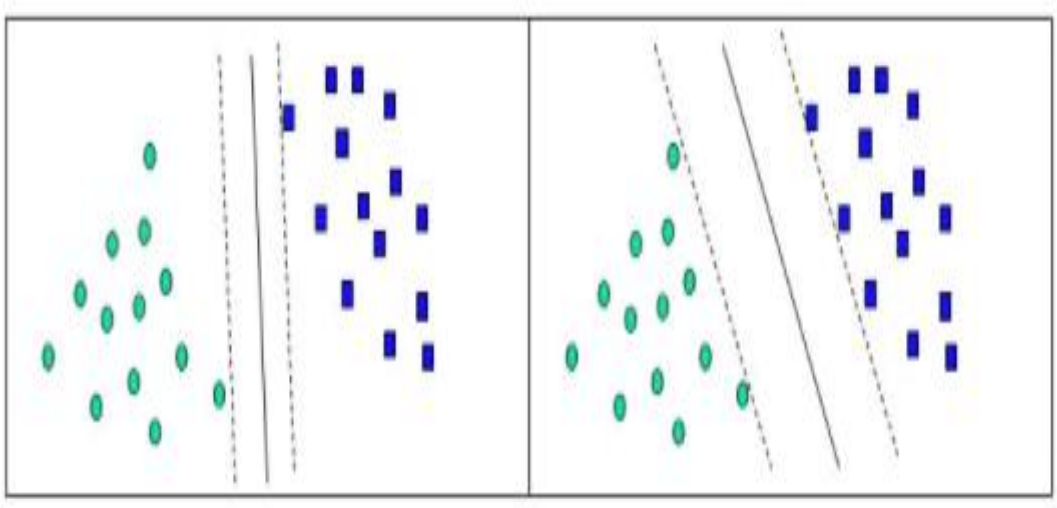
Şekil 4.1. Sınıflandırma işlemi örneği; bu örnekte algoritma verileri başarılı bir şekilde 2 ayrı sınıfa ayırmaktadır.



Şekil 4.2 İkili sınıflandırma örneği. Algoritma (?) ile tanımlanan nesnenin sahip olduğu niteliklere bakarak nesnenin hangi sınıfa ait olduğunu doğru bir şekilde tespit etmelidir.

4.5 Destek Vektör Makineleri

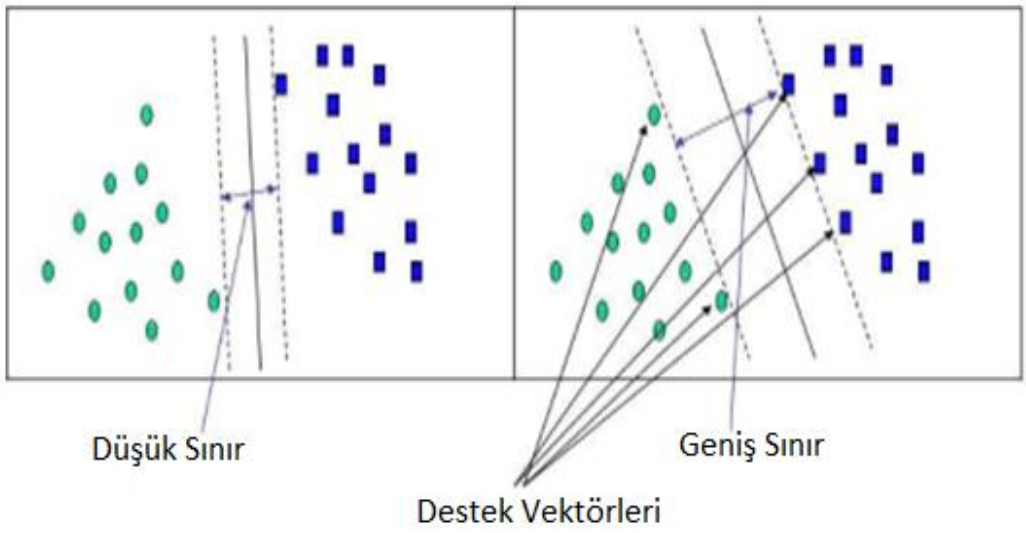
Destek vektör makineleri istatistiksel öğrenme teorisine dayalı danışmanlı öğrenme yöntemi olup Vapnik [22] tarafından tanımlanmış bir sınıflandırma tekniğidir. Destek Vektör Makineleri örüntü tanıma, veri madenciliği gibi alanlarda önemli bir kullanım alanına sahiptir.



Şekil 4.3. Destek Vektör Makineleri kullanılarak yapılan sınıflandırma örneği.

Makine öğrenmesi ve veri madenciliği alanlarında sınıflandırma probleminin çözümüne ne dair ortaya konan çalışmalar önemli bir yer tutmaktadır. Tıp alanında hastalık teşhisi, biyoloji alanında canlıların sınıflandırılması, bankacılıkta riskli müşterilerin tahmini, kimya alanında ilaçların etkilerinin belirlenmesi, sosyal medya alanında spamların saptanması gibi problemlerde çeşitli sınıflandırma yöntemlerinden faydalanılmaktadır.

İkili sınıflandırmada destek vektör makineleri kullanımında verilen eğitim seti $\{x_1, \dots, x_n\}$, $x_i \in R^d$ nitelikler vektörü, ve $\{ y_1, \dots, y_n \}$, $y_i \in \{-1, 1\}$ etiket değerleridir. Burada destek vektör makinelerinin amacı eğitim setinde bulunan farklı etiket değerine sahip olan verilerin en büyük sınırdaki hiper düzlem ile ayrılasını sağlamaktır. Hiper düzleme en yakın noktadaki eğitim verileri destek vektörleri olarak isimlendirilirler. Şekil 4.4.'te destek vektör makinelerinin kullanımı bir örnekle görsel olarak tanımlanmaktadır.



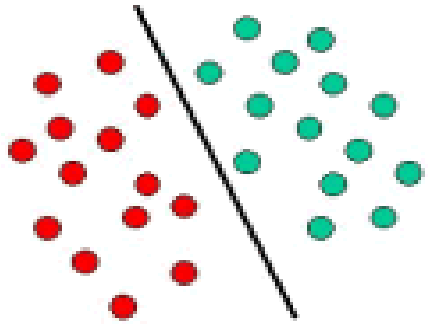
Şekil 4.4. Destek Vektörleri ve Sınırlar.

Destek vektör makineleri lineer (doğrusal) olarak ayrılabilen ve lineer (doğrusal) olarak ayrılamayan verilerin sınıflandırılmasında kullanılabilir. Doğrusal olmayan bir eşleme ile n boyutlu veri kümesi m > n olacak şekilde m boyutlu yeni bir veri

kümesine dönüştürülür ve yüksek boyutta doğrusal sınıflandırma işlemi yapılabilir [45]. Uygun bir dönüşüm ile her zaman veri bir hiper düzlem ile iki sınıfa ayrılabilir.

4.5.1 Lineer Olarak Ayrılabilen Verilerde DVM:

Eğitim için kullanılan n elemanlı eğitim setinde bulunan veriler hiper düzlem ile doğrudan ayrılabilmektedirler. Bu durum Şekil 4.5'te açıkça görülmektedir. Kırmızı ve yeşil etikete sahip veriler hiper düzlem yardımıyla doğrusal olarak ayrılmışlardır. Destek vektör makinelerinin amacı hiper düzlemin iki sınıftaki uç örneklere aynı uzaklıkta olmasını sağlamaktır.



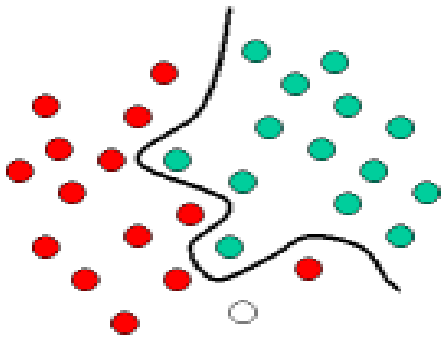
Şekil 4.5. Lineer olarak ayrılabilen verilerde DVM ile sınıflandırma örneği [36].

4.5.2 Lineer Olarak Ayrılamayan Verilerde DVM:

Lineer olarak ayrılabilen verilerde eğitim setindeki elemanlar iki farklı sınıfa doğrusal bir düzlem ile ayrılabilmektedirler; ancak, bu durum her zaman geçerli olmayabilir.

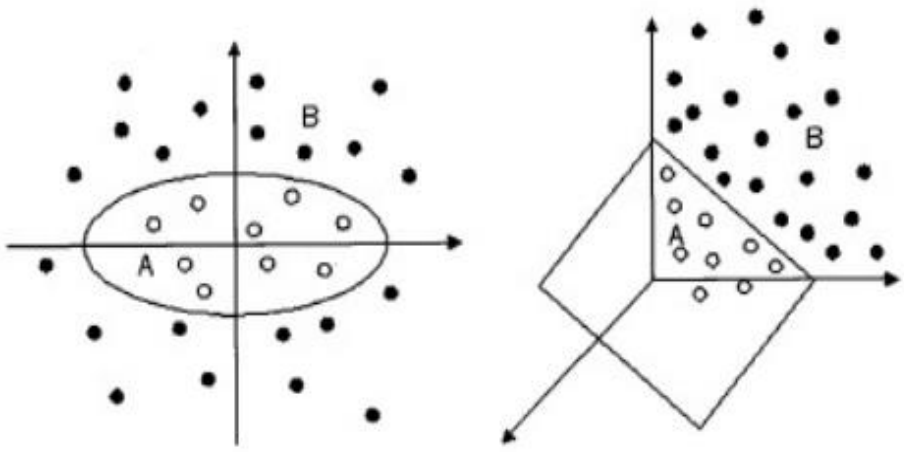
Lineer olarak ayrılamayan verilerde doğrusal bir düzlem ile veriler iki farklı sınıfa başarılı bir şekilde ayrılamazlar. Bu durumun bir örneği Şekil 4.6'da gösterilmektedir, öyle ki, yeşil ve kırmızı elemanları birbirinden ayırmak için bir eğri kullanılması gerektiği görülmektedir. Lineer olarak ayrılamayan verilerde en uygun

ayrımın sağlanması için doğrusal hiper düzlemlerden daha farklı bir yapıya ihtiyaç vardır.



Şekil 4.6. Doğrusal bir düzlem ile lineer olarak ayrılamayan verilerin bir örneği [36].

Verilerin lineer olarak ayrılamadığı durumda lineer sınıflandırıcılar yerine lineer olmayan sınıflandırıcılar kullanılabilir. Dolayısıyla $x \in R^d$ nitelikler vektörü daha fazla boyutlu bir nitelik uzayında yeniden tanımlanarak bu çok boyutlu nitelik uzayında arzulanan başarılı doğrusal sınıflandırıcıları elde etmek mümkün olabilir. Şekil 4.7'de lineer olarak ayrılamayan bir eğitim setinin R^2'den R^3'e taşınması ve daha sonra başarılı bir şekilde sınıflandırılması örneği gösterilmektedir.



Şekil 4.7. Lineer olarak ayrılamayan eğitim setinin R^2'den R^3'e taşınması örneği.

Farklı veri setleri üzerinde sınıflandırma probleminin çözümüne yönelik destek vektör makinelerinin kullanılmasında doğru çekirdek fonksiyonun seçimi ve parametrelerin en uygun şekilde kullanılması önemlidir. Literatürdeki çalışmalarda genellikle daha iyi sonuçlar verdiği ve uygulanması kolay olduğu için radyal tabanlı çekirdek fonksiyonun kullanıldığı görülmüştür [32].

Birçok çalışmada Destek Vektör Makineleri'nin uygulanması işleminde çekirdek fonksiyonu seçilirken radyal tabanlı fonksiyon kullanılmaktadır [1]. Fakat, her problemin çözümü için radyal tabanlı çekirdek fonksiyonunun kullanımı en iyi sonucu vermeyebilmektedir, dahası çekirdek fonksiyonların başarı oranı problem tanımına ve kullanılan veri setine göre değişkenlik gösterebilmektedir.

Bir Destek Vektör Makinesi temel olarak tanımlanmak istenirse [33]:

1. Nitelik uzayında maksimum ayırma oranına sahip hiper düzlemdir.
2. Girdi uzayında bir çekirdek fonksiyon tarafından oluşturulur.
3. Verileri iki sınıf arasında sınıflandırmak amacıyla karar fonksiyonu olarak kullanılır.

4.5.3 Çekirdek Fonksiyonlar

İdeal bir çekirdek fonksiyon aynı sınıftaki nesneler için yüksek bir benzerlik sonucu ortaya koyarken farklı sınıf etiketine sahip nesneler için daha düşük benzerlik sonucu ortaya koymalıdır. Burada çekirdek fonksiyonunun görevi birbirlerine benzer nesneleri birbirlerine yakın, farklı nesneleri de birbirlerinden uzak olarak nitelik uzayına eşlemektir.

Yoğun olarak kullanılan çekirdek fonksiyonlar aşağıda belirtilmiştir [37]:

- Lineer Çekirdek Fonksiyon
- Radyal Tabanlı Çekirdek Fonksiyon
- Polinomiyal Çekirdek Fonksiyon
- Sigmoid Çekirdek Fonksiyon

Lineer Çekirdek Fonksiyon:

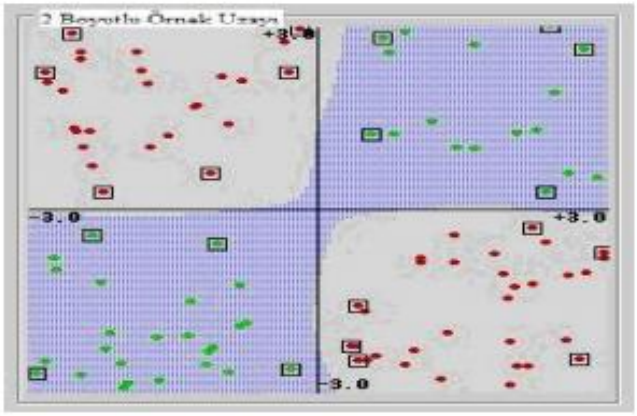
Bu çekirdek fonksiyon nitelik sayısı fazla ise ve daha yüksek boyutlu bir nitelik uzayında gerek duyulmuyor ise başarılı sonuçlar ortaya koyabilir. Gürültülü veriler ile iyi sonuçların ortaya çıkma şansı azdır. Gösterimi şu şekildedir:

$$K(x_i, x_j) = x_i^T \cdot x_j \tag{4.1}$$

Radyal Tabanlı Çekirdek Fonksiyon:

Radyal tabanlı çekirdek fonksiyon eğitim verilerini daha yüksek boyutlu bir nitelik uzayında tanımladıktan sonra lineer olarak ayrılamayan verilerin sınıflandırılmasında kullanılabilir. Ancak, lineer çekirdek fonksiyonlarının aksine nirelik sayısının çok fazla olduğu durumlarda kullanılmak için uygun olmayabilir. Gösterimi şu şekildedir:

$$K(x_i, x_j) = \exp(-\gamma |x_i - x_j|^2) \tag{4.2}$$

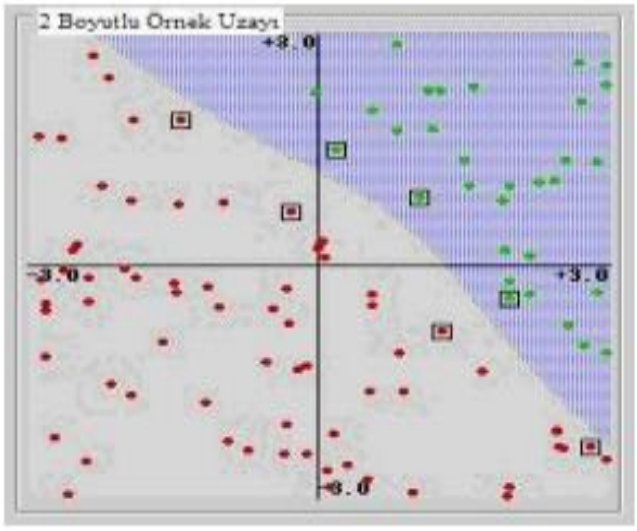


Şekil 4.8. Örnek Uzayda Radyal Tabanlı Fonksiyon Gösterimi [40]

Polinomiyal Çekirdek Fonksiyon:

Polinomiyal çekirdek fonksiyonda kullanılan formül aşağıda gösterilmektedir:

$$\mathrm{K}(x_i, x_j) = (\gamma . x_i . x_j + C)^d \tag{4.3}$$

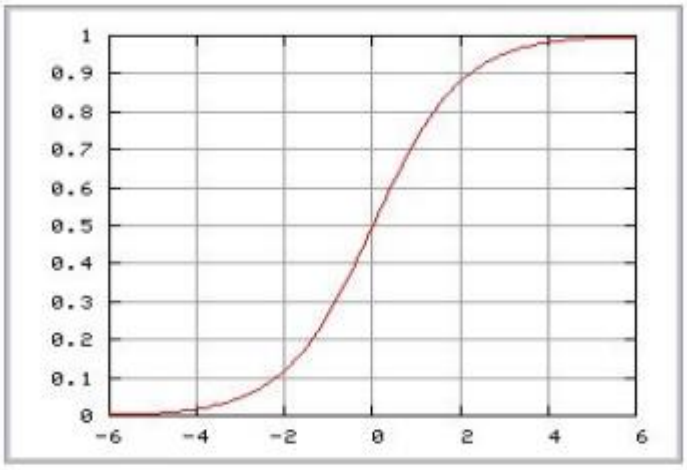


Şekil 4.9. Örnek Uzayda Polinomiyal Fonksiyon Gösterimi [40].

Sigmoid Çekirdek Fonksiyon:

Sigmoid çekirdek fonksiyon ile sınıflandırma yapılırken aşağıdaki formül kullanılır:

$$K(x_i, x_j) = \tanh(\gamma . x_i . x_j + C) \quad (4.4)$$



Şekil 4.10. Sigmoid Fonksiyon Gösterimi [40].

4.6 LIBSVM Kütüphanesinin Kullanımı

LIBSVM destek vektör makineleri ve regresyon için hazırlanan bir kütüphanedir. LIBSVM'nin amacı kullanıcıların destek vektör makinelerini uygulamalarında kolayca kullanabilmelerini sağlamaktır. LIBSVM kütüphanesinde kullanılan formülasyonlar şu şekildedir:

C-destek vektör sınıflandırması (C-SVC), ν-destek vektör sınıflandırması (ν-SVC), dağılım tahmini (tek-sınıflı SVM), ε-destek vektör regresyonu (ε-SVR) ve ν-destek vektör regresyonu (ν-SVR).

Sınıflandırma Destek Vektör Makineleri:

C-destek vektör sınıflandırması (C-SVC):

Bu destek vektör makinesi türünde aşağıdaki optimizasyon problemi çözülmektedir.

$$\frac{1}{2} w^T w + \mathrm{C} \sum_{i=1}^{l} \xi_i \quad (4.5)$$

Kısıtlar da aşağıdaki gibidir:

$$y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i \quad (4.5a)$$

$$\xi_i \geq 0, i = 1, \ldots, l \quad (4.5b)$$

Burada C kapasite sabiti, w katsayılar vektörü, b bir sabit ve ξ_i de ayrılamayan verileri temsil eden bir parametredir.

ν-destek vektör sınıflandırması (ν-SVC):

ν-destek vektör sınıflandırmasında farklı olarak $\nu \in (0, 1]$ parametresi tanımlanmıştır. Bu destek vektör makinesi türünde aşağıdaki optimizasyon problemi çözülmektedir.

$$\frac{1}{2} w^T w - vp + \frac{1}{N} \sum_{i=1}^{l} \xi_i \quad (4.6)$$

Kısıtlar da aşağıdaki gibidir:

$$y_i(w^T \phi(x_i) + b) \geq p - \xi_i \quad (4.6a)$$

$$\xi_i \geq 0, i = 1, \ldots, N \quad (4.6b)$$

$$p \geq 0 \quad (4.6c)$$

Regresyon Destek Vektör Makineleri:

$$y = f(x) + \text{gürültü} \quad (4.7)$$

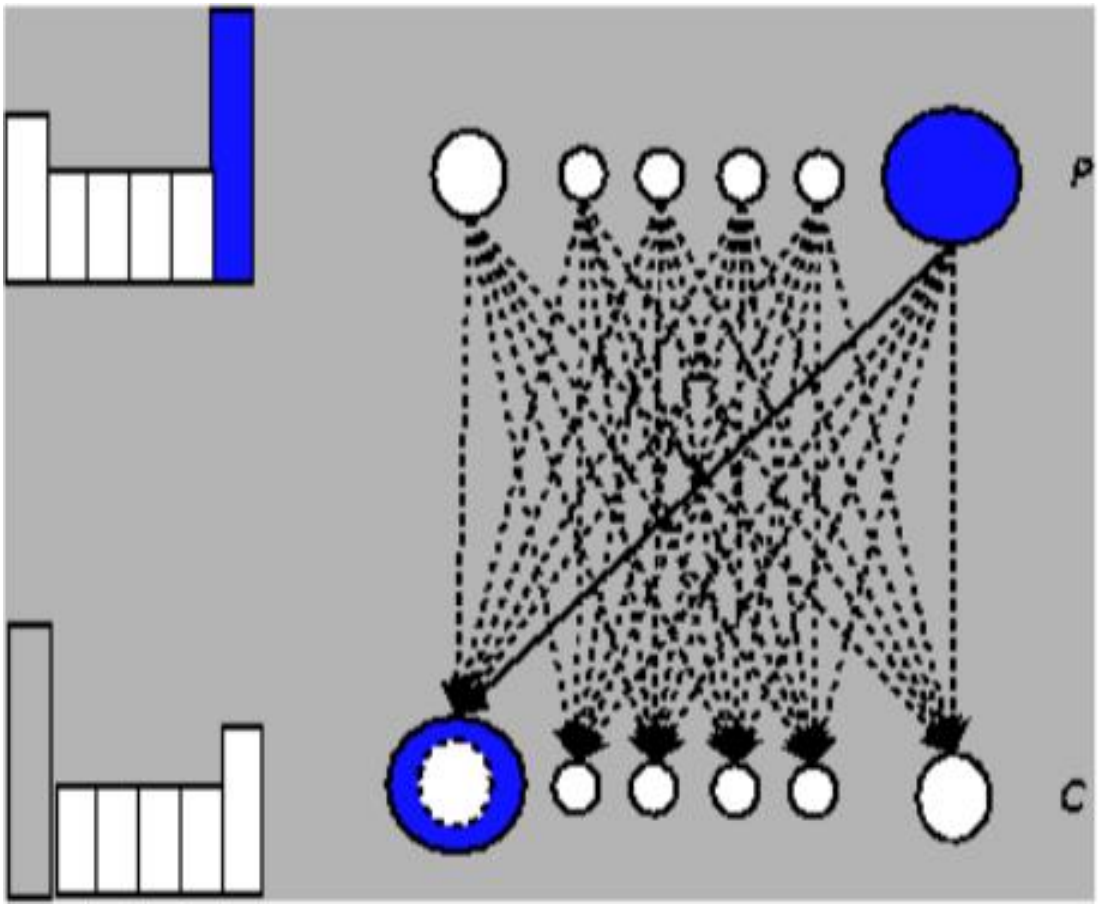
Buradaki amaç f fonksiyonu için destek vektör makinelerinin karşılaşmadığı yeni durumların tahmininin doğru şekilde yapılmasıdır. DVM modeli örnek bir set

üzerinde çalıştırılarak hata fonksiyonunun optimizasyon işlemi yapılabilir ve f fonksiyonunun başarılı bir regresyon analizi yapması sağlanabilir.

5. EARTH MOVER'S DISTANCE (EMD) ALGORİTMASI

The Earth Mover's Distance (EMD) yöntemi, öznitelikler arasındaki yer uzaklığı bilinen bir öznitelik uzayındaki, iki çok boyutlu dağılım arasındaki ayrılık oranını hesaplama işlemidir.

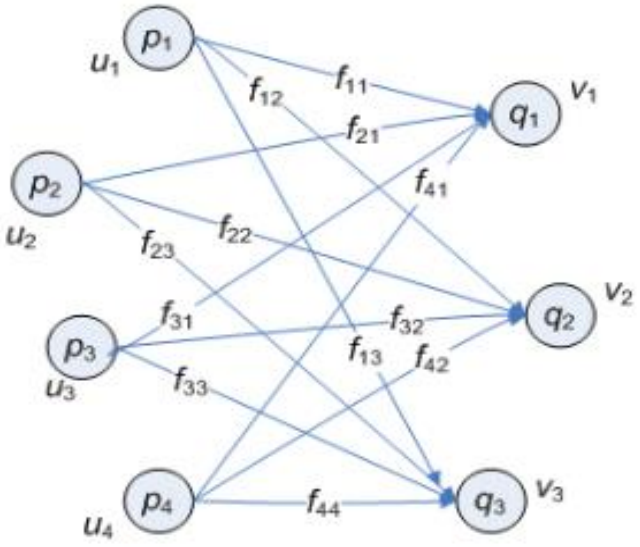
EMD algoritmasını açıklamak için kullanılacak en iyi benzetme, verilen iki dağılım için, dağılımlardan birinin dünyanın kütlesinin uzayda düzgün olarak dağılmış olması ve diğer dağılımın da aynı uzayda dağılmış delikler olarak değerlendirilmesi olarak görülebilir. Daha sonra, EMD algoritması kütleleri deliklere doldurmak için gerekli en az iş miktarını ölçer. İş miktarı, bir birim dünya kütlesinin yer uzaklığı kullanılarak deliklerden birinin içine yerleştirilmesi olarak hesaplanır.



Şekil 5.1. İki benzer fakat tam olarak aynı olmayan dağılım arasındaki EMD sonucu.

Dağılımlar imza adı verilen yapılar kullanılarak temsil edilirler ve bir dağılımın içinde bulunan çeşitli bölümler imzanın da niteliklerini oluştururlar. İmzalar çeşitli boyutlarda olabilirler; basit dağılımlar kısa imzalar ile karmaşık dağılımlar da uzun imzalar ile temsil edilirler.

EMD sonucunun hesaplanmasında dağıtım probleminin çözümü temel alınır [23]. Kısa bir örnekle açıklamak gerekirse, birkaç dağıtıcının elinde belli miktarda dağıtılmayı bekleyen malzeme vardır ve bu malzemeleri yine birkaç tüketiciye belli bir kota dahilinde ulaştırmaları gerekmektedir. Her dağıtıcı ve tüketici için bir birim malzemenin dağıtım masrafı belirlidir. Dolayısıyla, dağıtım problemi dağıtıcılardan tüketicilere giden ve tüketicilerin ihtiyacını karşılayan en az maliyetli malzeme akışının hesaplanması olarak tanımlanır. Dağıtım problemindeki dağıtıcılar ve tüketiciler imzalar olarak tanımlanır ve dağıtım masrafı da yer uzaklığı olarak alınırsa imza eşleme işlemi kolaylıkla dağıtım problemi olarak ortaya konabilir. Ortaya çıkan sonuç da bir imzanın diğerine dönüştürülmesi için harcanan en küçük iş miktarı olarak hesaplanır. Şekil 5.2'de iki imza arasındaki EMD uygulamasının bir örneği görülmektedir.



Şekil 5.2. İki imza arasındaki Earth Mover's Distance uygulaması örneği.

5.1 EMD Algoritmasının Hesaplanması

Earth Mover's Distance algoritması Rubner [8] tarafından, iki dağılım arasındaki ayrılık oranının imzalar kullanılarak hesaplanması amacıyla ortaya konmuştur. Burada A elemanlı bir imza S = $\{ s_j = (w_j, m_j) \}_{j=1}^{N}$ seti olarak ve m_j j'ninci elemanın pozisyonu w_j de j'ninci elemanın ağırlığı olarak tanımlanır.

P ve Q olarak adlandırılan iki imza P = $\{(p_i, u_i)\}_{j=1}^{m}$ ve Q = $\{(q_j, v_j)\}_{j=1}^{n}$ şeklinde m ve n imzaların büyüklüğünü belirtecek şekilde tanımlandığında bu iki imza arasındaki EMD sonucu dağıtım probleminin çözümü olarak modellenmiş olur. Bu örnekte Q'nun v_j'de bulunan elemanı tüketici ve P'nin de u_i'de bulunan elemanı dağıtıcı olarak düşünülür; dolayısıyla, q_j talebi, p_i de arzı temsil etmektedir. Sonuç olarak, EMD arz ve talep arasındaki dağıtımın en az iş harcanarak çözülmesi olarak formüle edilir:

$$\text{EMD (P, Q)} = \min_{F = \{f_{ij}\}} \frac{\sum_{i,j} f_{ij} d_{ij}}{\sum_{i,j} f_{ij}} \tag{5.1}$$

Ayrıca şu kısıtlara sahiptir:

$$\sum_j f_{ij} \leq p_i \tag{5.1a}$$

$$\sum_j f_{ij} \leq q_i \tag{5.1b}$$

$$\sum_{i,j} f_{ij} = \min\{ \sum_i p_i , \sum_j q_j \} \tag{5.1c}$$

$$\sum_j f_{ij} \leq p_i \tag{5.1d}$$

$$f_{ij} \geq 0 \tag{5.1e}$$

F = $\{f_{ij}\}$ akış setini temsil eder. Her f_{ij} akışı i'inci arzdan j'ninci talebe giden miktarı temsil eder. Ek olarak, v_{ij} ve u_i arasındaki yer uzaklığı da d_{ij} olarak tanımlanmıştır [24].

5.2 EMD Algoritmasının Kullanım Alanları

EMD algoritmasının birçok kullanım alanı bulunmaktadır ve bunlardan önemli olan bazıları görüntü erişimi, kilit nokta eşleştirme, kenar ve köşe bulma gibi işlemlerdir. EMD algoritmasını görüntü erişimi ve nesne tanıma alanlarında uygulanması işlemleri aşağıdaki alt başlıklarda açıklanmıştır.

5.2.1 Görüntü Erişiminde EMD Kullanımı

Dünya üzerinde her gün çok büyük miktarda dijital görüntü ve video verisi olmak üzere görsel bilgi ortaya çıkmaktadır. Görüntüler içerisindeki ortak örüntülerin çıkarılması işlemi birçok uygulama alanı için önemli bir adım halini almıştır. Öncelikli olarak, bu ortak örüntüler sınıflandırma ve tarama amacıyla, ikincil olarak da video dizinleme, kümeleme ve özetleme işlemleri için temel oluşturabilirler. Ancak, ortak örüntü tespiti problemi arama ölçeğinin ve problem alanının büyük olmasından dolayı zor bir problemdir. Ortak örüntü tespiti yapılabilmesi için çekim açısı, üst üste oturma, rotasyon, ölçeklendirme, segmentasyon konuları göz ardı edilmeden ele alınmalıdır. Var olan çözümlerin bir bölümü bu konuların bazılarını göz ardı ederek hedef görüntü kategorisini daraltmaktadır [25]. Görüntü erişimi problemi ortak örüntü konumun bulunması ve uzaysal arama ölçeğinin küçültülmesiyle daha kolay bir hal alabilir.

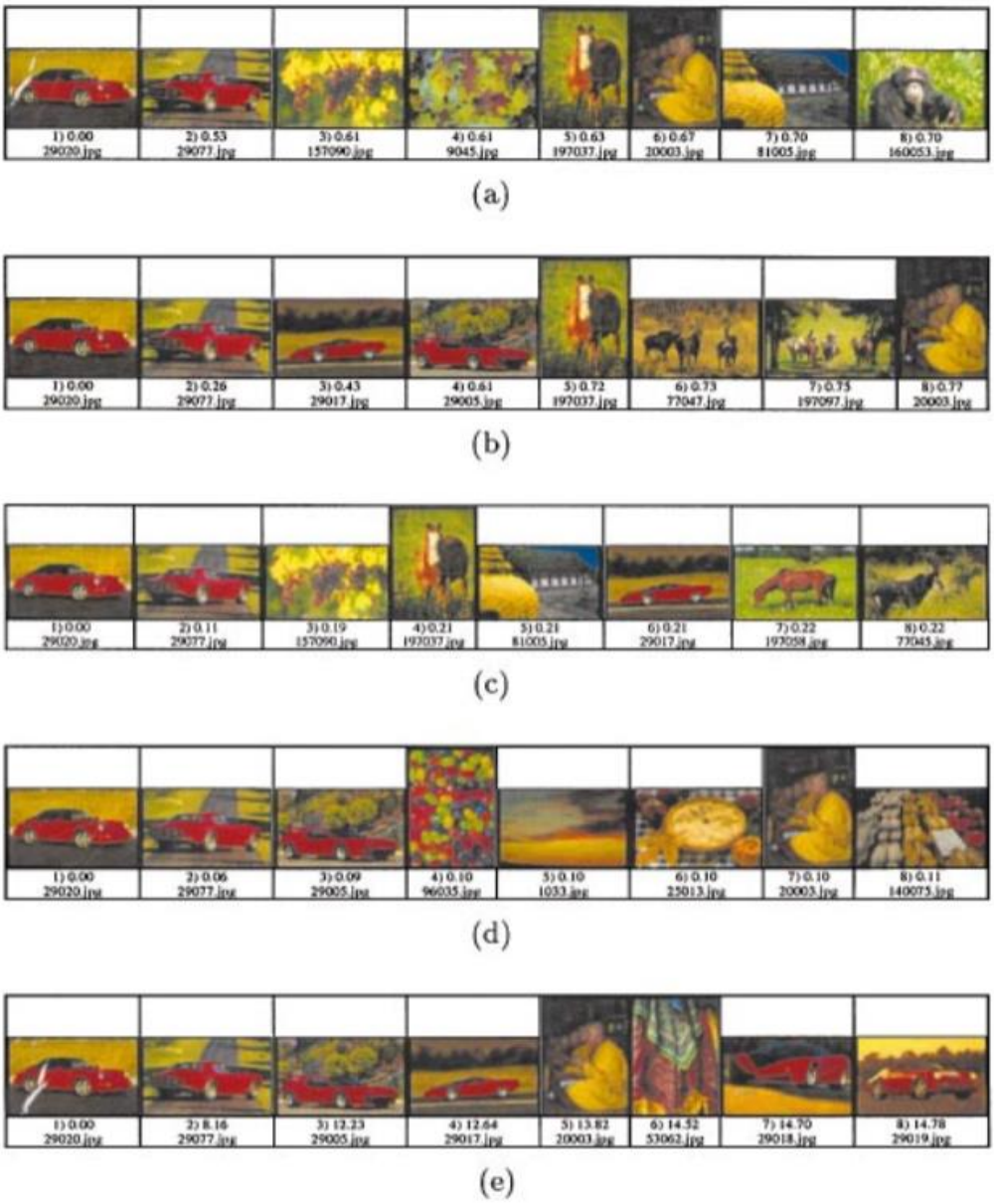
Şekil 5.3. İki resim arasındaki ortak örüntünün farklı arka planlarda tespiti.

Şekil 5.3'te iki resim arasındaki ortak örüntünün farklı arka planlarda tespitinin yapılması işlemi ortaya konmaktadır. Şekildeki örüntü tespiti işleminde görüldüğü gibi iki resim arasındaki ortak örüntüler etrafında üst üste oturma, ölçekleme, farklı çekim açısı, ışıklandırma farkı gibi problemler var ise örüntü bulma işlemi zordur.

Görüntü erişiminde Earth Mover's Distance algoritmasını kullanmak histogram eşleme tekniklerinin kullanımına göre daha istikrarlı sonuçlar vermektedir. Bunun nedeni, EMD değişken uzunluklu temsiller üzerinde kullanıldığı için histogramlar üzerinde geçerli olan nicemleme (quantization) ve diğer binning problemlerinden kaçınılmasıdır.

Rubner ve arkadaşları tarafından ortaya konan çalışmada [8] EMD algoritmasının renk ve doku alanı üzerindeki görüntü erişimi performansı değerlendirilmiştir. Bu çalışmaya göre değerlendirilen iki dağılım için nicelikli bir benzeşmezlik ölçüsü ortaya koymak ve ayrılık oranını olabildiğince doğru tahmin edebilmek önemlidir. Bu durum görüntü erişiminde önemlidir, ancak özellikle doku ayrımı ve renk algısı için özellikle gereklidir.

[8]'de uygulanan yöntem ile elde edilen sonuçlar Şekil 5.4'te gösterilmektedir. Burada kırmızı bir araba sorgusu üzerinde algoritma çalıştırılmış ve sorguya en yakın görüntülere erişilmiştir. Şekilde elde edilen görüntüler sorguya benzerlik oranlarına göre soldan sağa doğru sıralanmışlardır.

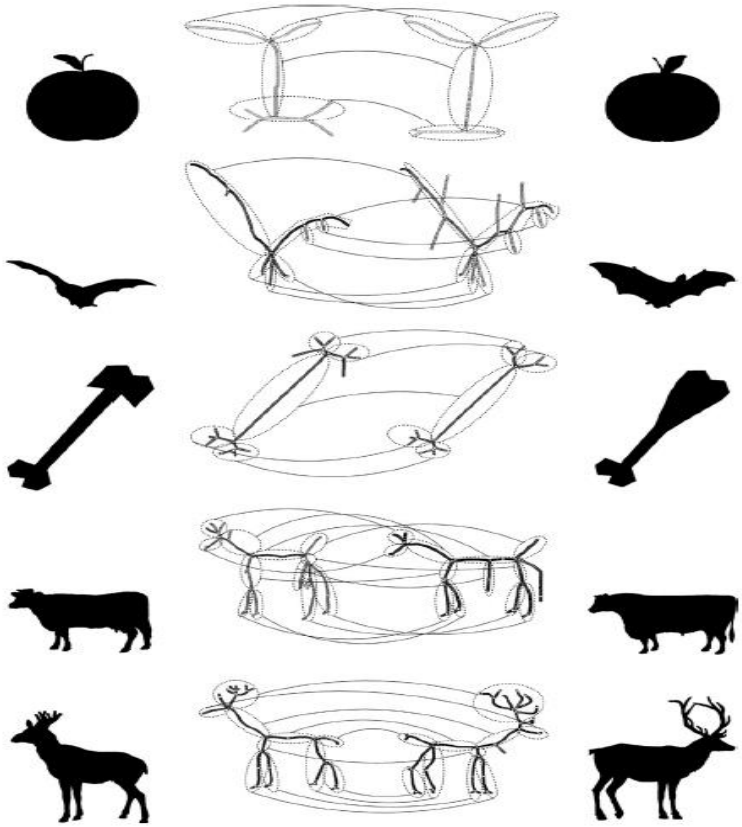


Şekil 5.4. Rubner ve arkadaşları tarafından EMD algoritmasının uygulandığı karşılaştırmalı görüntü erişimi örneği. (e)'de EMD algoritması sonuçları ve (a), (b), (c), (d) de ise karşılaştırılan diğer algoritmaların verdiği sonuçlar ortaya konmaktadır. En soldaki görüntüler sorgu görüntüsü olup, diğer görüntüler sorgu görüntüsüne benzerlik oranlarına göre soldan sağa sıralanmaktadırlar. En iyi sonucu (e)'nin verdiği görülmüştür.

5.2.2 Neste Tanımada EMD Kullanımı

Nesne tanıma ve sınıflandırma işlemleri için nesne niteliklerini çoklu olarak eşleyebilme durumu önemli bir süreçtir. Birebir eşleme yapan algoritmalar iki benzer

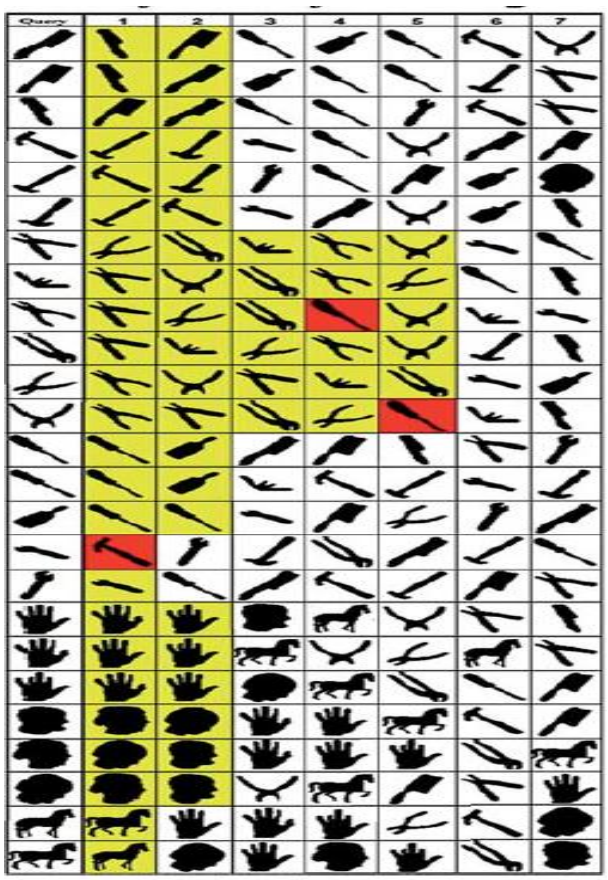
nesne nitelikleri arasında oluşabilecek segmentasyon ve bitiştirme problemleriyle, ölçeklendirme farklılıklarıyla baş etmekte zorlanırlar. [5]'te yapılan çalışmada nesneler arasındaki benzeşmezlik (ayrılık) oranı hesaplanırken, öncelikli olarak nesnelerin nitelikleri çizgeler ile tanımlanır ve bu çizgeler metrik ağaçlar olarak dönüştürülürler. Daha sonra, yapılan çalışmaya göre bu ağaç temsilleri geometrik uzayda l_1 normu altında izometrik olarak gömülürler. Bu yöntemde ortaya konan çalışmada gürültü problemi ortadan kaldırılmıştır. En son olarak, gömülen noktalar EMD algoritması kullanılarak çoklu olarak eşleştirilirler. Şekil 5.5'te [5]'te ortaya konan çalışmanın veri seti üzerindeki görsel tasviri ortaya konmaktadır.



Şekil 5.5. [5]'te ortaya konan çalışmaya göre bazı şekiller için ortaya çıkan eşleştirme sonucu. Elipslerin içerisinde bulunan ve birbirleriyle eşleşen iskelet grupları çoklu eşleşmeleri göstermektedirler.

Nesnelerin niteliklerini çoklu olarak eşleyebilme durumu nesneleri sınıflandırmada önemli bir koşul olarak ortaya çıkmaktadır. Birebir nitelik uyuşmasını dikkate alan

katı eşleşme uygulamalarında nesne nitelikleri tam anlamıyla birbirleriyle uyuşmuyor ise, yani benzerlikleri daha soyut ise, başarısız sonuçlar ortaya çıkmaktadır. [4]'te ortaya konan çalışmada çizge temsilleri kullanılarak çoklu eşleşme probleminin çözümü ele alınmıştır. Burada öncelikli olarak karşılaştırılacak çizgelerin metrik ağaçlar kullanılarak bir temsilleri oluşturulur. Daha sonra, çizgelerin düğüm bilgileri kullanılarak bu temsiller geometrik uzayda yeniden tanımlanırlar (gömülürler). Böylece çoklu çizge eşleme problemi çoklu geometrik nokta eşleme problemi haline dönüştürülmüş olur ve problem EMD algoritmasının etkili bir şekilde kullanımına uygundur. [4]'te yapılan çalışmanın bir veri seti üzerindeki sonuçları Şekil 5.6'da ortaya konmaktadır.



Şekil 5.6. [4]'te ortaya konan çalışmanın veri seti üzerindeki eşleştirme sonuçları. Sarı ile işaretlenenler doğru eşleştirmeleri, kırmızı ile işaretlenenler yanlış eşleştirmeleri ortaya koymaktadır. Diğer kategorilerdeki eşleşmeler ise beyaz ile gösterilmiştir.

5.2.3 EMD Algoritmasının Kullanım Avantajları

EMD algoritmasının avantajları aşağıda sıralanmıştır:

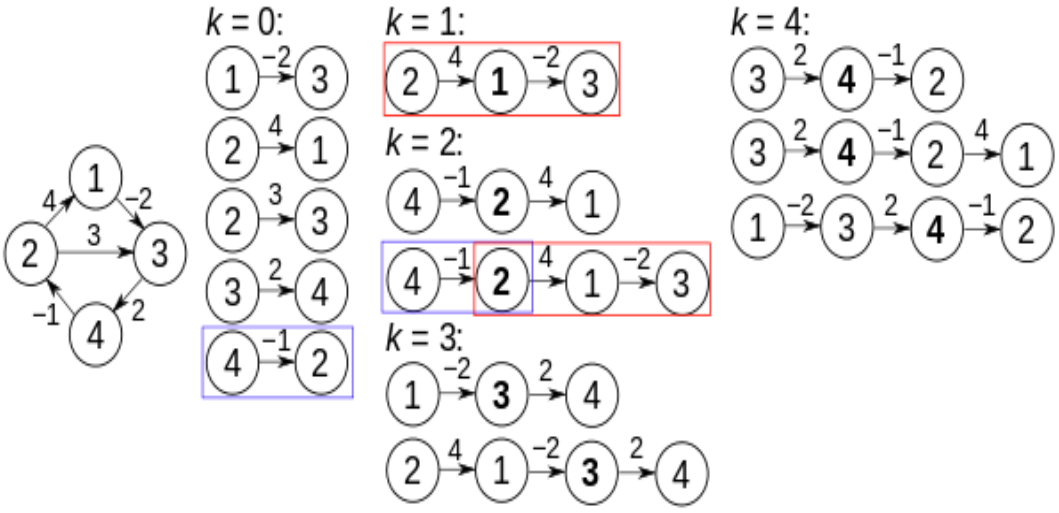
- Tekli öğeler arasındaki uzaklık kavramını doğal olarak setler, dağılımlar arasına yayar.
- Değişken boyutlu imzalar için uygulanabilir. Bu durumda imzalar histogramları kapsarlar. Ayrıca, imzaların kompakt yapıdadır ve elemanların hareket maliyetleri yakınlık kavramını nicemleme problemleri olmadan uygun bir şekilde yansıtırlar.
- Kısmi eşleşmelere doğal olarak izin veren bir yapıdadır; bu nedenle, görüntü erişimi, üst üste oturma, karışıklık problemleri için önemlidir.

6. SUNULAN ÇALIŞMA

Bu bölümde tez çalışmasında ortaya konan metot, denenen ve değerlendirilen teknikler ile ilgili anlatımlar yer almaktadır.

6.1 Sunulan Güncel Metot

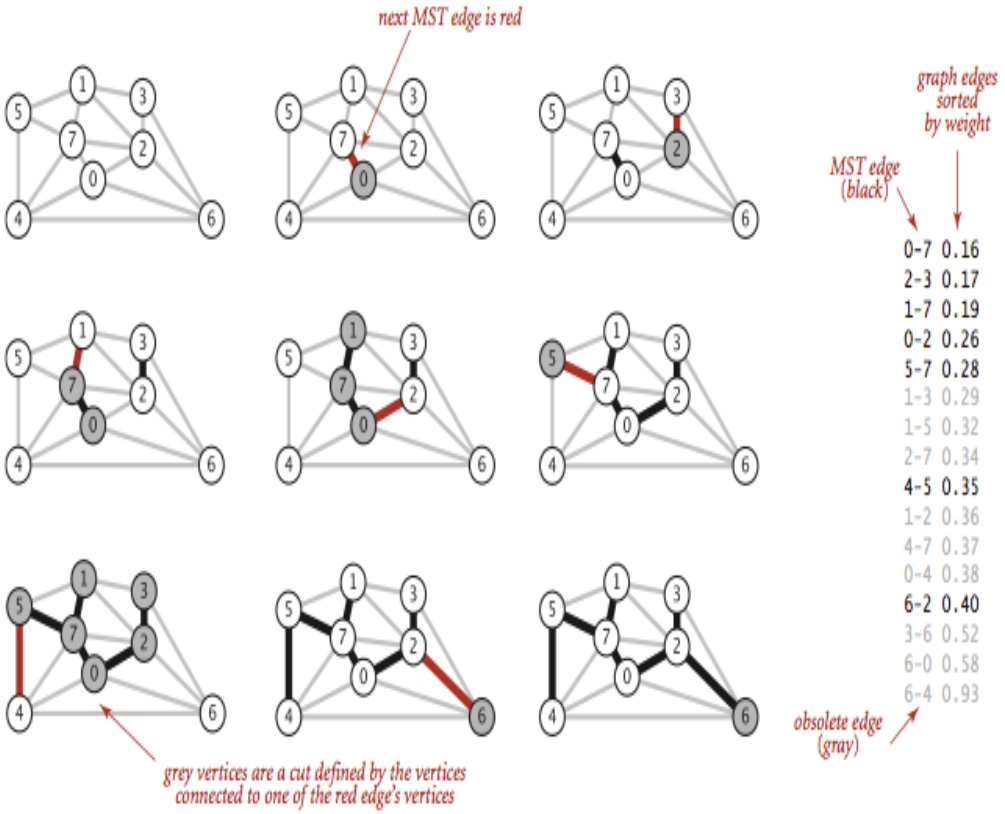
Giriş bölümünde değinildiği üzere bu çalışmada kimyasal moleküllerin eşlenmesi için çizge temelli örüntü tanıma tekniği kullanımı uygulanmıştır. Moleküller kenar ağırlıklı çizgeler olarak, atomlar düğüm, atomlar arasındaki bağlar da kenarlar olarak betimlenir. Bir kenar ağırlığı atomların bulundukları pozisyonlar göz önünde bulundurularak aralarındaki Öklid uzaklığına bağlı olarak belirlenir. Bu çalışmada yapılan ilk işlem kimyasal moleküllerin koordinat bilgileri dikkate alınarak Floyd Warshall algoritmasının kullanılması ve uzaklık matrisinin oluşturulması işlemidir.



Şekil 6.1. Bir çizge üzerinde Floyd Warshall algoritmasının çalışma prensibinin gösterimi.

Bu çalışmada kullanılan uygulama, çizgenin sahip olduğu düğümleri geometri uzayındaki noktalar olarak gömer ve düğümler arasındaki uzaklık atomların arasında bulunan gerçek uzaklık olarak kabul edilir. Gömme işlemi sonrasında, gömülü noktalar arasındaki uzaklığın hesaplanması için kabul görmüş bir dağıtım odaklı transportasyon algoritması olan Earth Mover's Distance (EMD) [5] kullanılır. EMD

algoritması noktalar arasındaki benzerlikleri bulmak için kullanılır. Bu çalışmada kullanılan yaklaşım Şekil 1.2'de gösterilmektedir; buna göre, iki çizge, gömme işlemi yapılarak aynı geometrik uzaya yerleştirilir ve EMD kullanılarak gömülen nokta setleri arasında eşleştirme yapılır. Daha detaylı olmak gerekirse, molekülleri temsil eden iki çizge için, bu çalışmadaki yaklaşım öncelikli olarak çizgeleri geometrik uzaydaki noktalar olarak tanımlamak ve bunu yaparken düğümler arasındaki uzaklığı uzayda tanımlanan noktalar arasındaki uzaklık olarak yeniden tanımlamaktır. Bu aşamadaki amaç (çizge gömme aşaması), çizge eşleme problemini, etkili ve başarılı algoritmaların kullanılabilir olduğu geometrik uzayda nokta eşleme problemi olarak yeniden formüle etmektir. Bu formülasyonun yapılabilmesi için girdi olarak kullanılan çizgelerin aynı zamanda ağaç olarak tanımlanabilir olması gereklidir. Ancak, eğer girdi olarak kullanılan çizgeler ağaç olarak tanımlanamıyorlar ise, başka bir deyişle, döngüye sahiplerse, çizgelerin metrik ağaç yakınsamasını oluşturmak için bir ön işleme adımı gereklidir.



Şekil 6.2. Kullanılan Kruskal MST algoritmasının örnek bir çizge üzerinde çalışma prensibinin gösterimi [35].

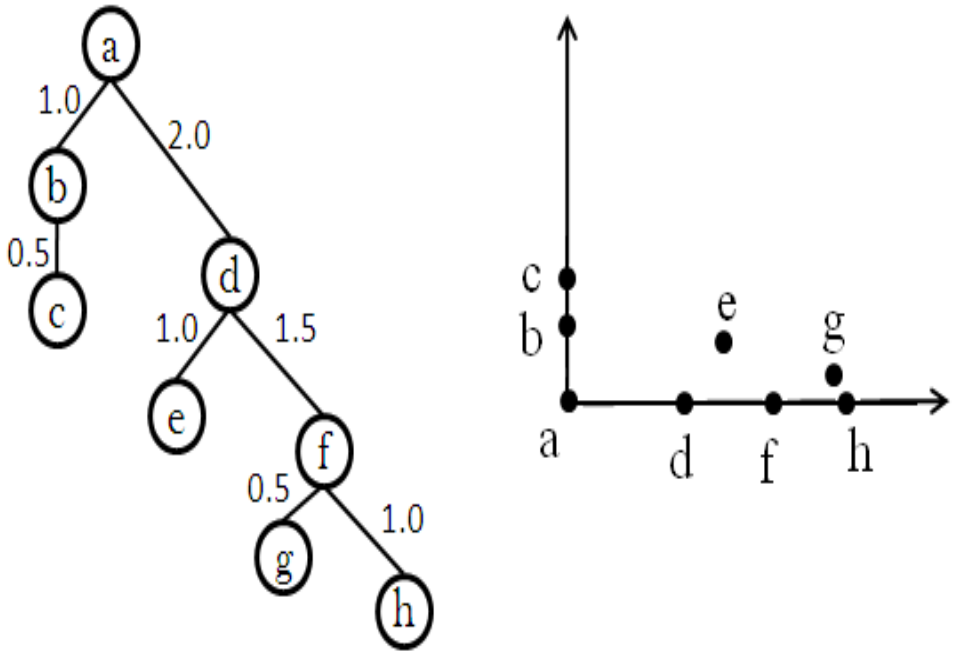
Bu çalışmada, bir girdi çizgesini ağaca dönüştürme işlemi o çizgenin en kısa örten ağacını bulmakla sağlanır, ağacın düğümleri atomları temsil eder ve ağacın kenarları yakın olan atomları birbirine bağlar. Ağacın kökü diğer tüm düğümler ile en kısa yol toplamlarını minimize edecek şekilde seçilir. Şekil 1.2'deki ilk adım bu işlemi göstermektedir.

Girdi çizgelerinin ağaçlar olarak tanımlanmasından sonra, çizge gömme işlemi gerçekleştirilir. Çizge gömme işlemi, en uzun ayrık kenarlı kökten yaprağa yolun bulunmasıyla başlar. Bu yol seviye 1 yolu olarak adlandırılır. Daha sonra, bu yol ağaçtan silinerek bir orman oluşturulmuş olunur. Aynı işlem, örnek olarak, bir sonraki en uzun ayrık kenarlı kökten yaprağa yol bulunur (seviye 2 yolu) ve ağaçtan silinir. Bu işlem tüm yollar silinene kadar sürer. Gömme işlemi sırasında çıkarılan her yol bir koordinat eksenine karşılık gelir ve toplam yol sayısı ağacın gömüldüğü uzayın boyutluluğunu belirler. Düğüm *v*'nin geometrik uzaydaki koordinatlarını belirlemek için, öncelikle *v* ile kök arasında eşsiz *P(v)* yolu bulunur. Eğer *P(v)*'nin l_1 ağırlığına sahip ilk bölümü seviye 1 yolu P^1'i takip etsin, ikinci bölümü l_2 ağırlığına sahip seviye 2 yolu P^2'yi takip etsin ve son bölümü l_d ağırlığına sahip seviye *d* yolu P^d'yi takip etsin. *P(v)*'nin ortaya çıkan ayrışma dizisi $<P^1,...,P^d>$ ve ağırlık dizisi $<l_1,...,l_d>$ olsun. Eğer *P(v)*'nin ayrışma dizisi seviye i yolu P^i'ye sahip ise, onun karşılığı olan koordinat ağırlık dizisinde belirtilen l_i ağırlık değerine sahip olur. Aksi takdirde, ilgili koordinat 0 olacaktır. Bu işlem sırasında ağacın kökünün her zaman orijine eşleştirildiği gözlemlenebilir. Şekil 2'de gösterilen örnek ağaç 4 boyutlu bir geometrik uzaya gömülmüştür ve gömülen noktaların koordinatları: *a* = (0,0,0,0), *b*=(0,1.0,0,0), *c*=(0,1.5,0,0), *d*=(2.0,0,0,0), *e*=(2.0,0,1.0,0), *f*=(3.5,0,0,0), *g*=(3.5,0,0,0.5), ve *h*=(4.5,0,0,0)'dir. Bu gömme işlemi Manhattan uzaklığı altında izometriktir. Anımsamak gerekirse, d boyutlu geometrik bir uzayda, $X = [x_1,x_2,...,x_d]^t$ ve $Y = [y_1,y_2,...,y_d]^t$ noktaları arasındaki Manhattan uzaklığı $d_1(X,Y)$ şöyle hesaplanır:

$$d_1(X,Y) = \sum_{k=1}^{d} |xk - yk| \qquad (6.1)$$

Uygulanan gömme işlemi sonrasında iki farklı ağacın farklı boyutlarda temsil edilme olasılığı bulunmaktadır. Bu nedenle eşleme işlemi yapılmadan önce tarafların aynı

boyuta getirilmeleri gerekmektedir; dolayısıyla, az boyutta temsil edilen dağılıma dolgu işlemi yapılarak boyutlar eşitlenir. Bu işlem hakkında detaylı bilgi [5] referanslı çalışmada yer almaktadır.



Şekil 6.3. Örnek ağaç, a köküne sahiptir ve 4 boyutlu uzaya gömülmüştür. Gömülen düğümlerin koordinatları: a=(0,0,0,0), b=(0,1.0,0,0), c=(0,1.5,0,0), d=(2.0,0,0,0), e=(2.0,0,1.0,0), f=(3.5,0,0,0), g=(3.5,0,0,0.5), ve h=(4.5,0,0,0)'dir. Bu gömme işlemi Manhattan uzaklığı altında bozukluk içermez.

Çizge gömme işlemi tamamlandıktan sonraki adım gömme işlemi sonrasında ortaya çıkan nokta dağılımları arasındaki uzaklıkların Earth Mover's Distance (EMD) algoritması kullanılarak hesaplanmasıdır. EMD algoritması [8], transportasyon problemi üzerine kuruludur [9]. 2 dağılım arasındaki mesafeyi bulmak için, öncelikle dağılımlar içindeki her nokta d-boyutlu geometrik uzayda koordinat ve ağırlık bilgisiyle temsil edilir. Bir noktanın ağırlığı, o noktanın temsil ettiği atomun ağırlığıyla aynıdır. İlk dağılım başlangıç pozisyonu, 2. dağılım da varış pozisyonu olarak düşünülürse EMD tarafından ortaya konan sonuç, ilk dağılım 2. dağılıma ulaşırken (dönüşürken) nokta ağırlıklarının minimum yer değiştirme miktarı olarak hesaplanır. Bu çalışmada, EMD sonucu olarak ortaya çıkan uzaklık, verilen 2 molekül için hesaplanan benzeşmezlik oranı olarak değerlendirilmiştir. Bu

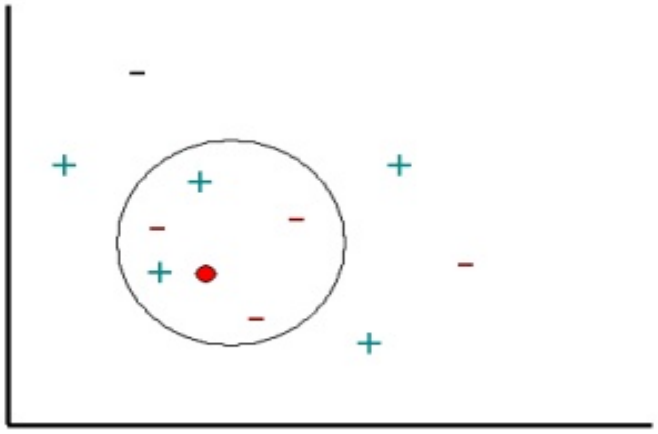
çalışmanın bilinen veri setleri kullanılarak ortaya konan test sonuçları ve saygın bazı yöntemlerle karşılaştırmaları bir sonraki bölümde yer almaktadır.

6.2 Çalışma Sonuçlarının Değerlendirilmesi için Kullanılan Yöntemler

Bu tez çalışmasında ortaya konan metot üzerinde farklı yöntemler denenerek iyileştirme denemeleri ve metot analizi yapılmıştır. Aşağıda bu yöntemler açıklanmaktadır. 6.2 bölümünde açıklanan yöntemler birlikte veya ayrı ayrı kullanılmış ve ortaya çıkan sonuçların önemli bir bölümü 7. bölümde yer alan deneysel sonuçlarda, Tablo 7.1'de raporlanmıştır.

6.2.1 Sınıflandırma Aşamasında k-En Yakın Komşuluk Algoritması Kullanımı

K-en yakın komşuluk (kNN) kuralı örüntü sınıflandırmasında kullanılan en eski ve basit yöntemlerden biridir. Buna rağmen, bazı durumlarda bu algoritma iyi sonuçlar verebilmektedir. kNN'ye göre her etiketlenmemiş örnek, eğitim setinde bulunan k-en yakın komşusunun sahip olduğu etiketlere bakılarak sınıflandırılır. kNN'nin performansı için önemli olan kavram en yakın k komşunun belirlenmesi için kullanılan uzaklık (benzeşmezlik oranı) metriğidir.



Şekil 6.4. k-en yakın komşuluk örneği. Kırmızı noktayla gösterilen örnek 1-en yakın komşuluk kullanılırsa (+), 2-en yakın komşuluk kullanılırsa belirsiz, 5-en yakın komşuluk kullanılırsa (-) olarak sınıflandırılır [43].

Algoritmayı kısaca açıklamak gerekirse:

1. Pozitif bir k sayısı belirlenir. Yeni bir veri değerlendirilmeye başlanır.
2. Veri setinde bulunan ve yeni veriye en yakın uzaklığa sahip k tane eleman seçilir.
3. Seçilen k tane elemanın sahip olduğu sınıf bilgileri incelenir en fazla etikete sahip sınıf grubu belirlenir.
4. Yeni veri en fazla etikete sahip sınıf grubuyla etiketlenir.

kNN algoritmasının avantajları aşağıda sıralanmıştır:

- Anlaşılması ve programlanması kolaydır.
- Kayıp verilerin üstesinden gelebilir.
- Uzaklık metriği hesaplanması dar bir uzayda gerçekleşir.
- Çoğunluk tarafından kabul edilmeyen veriler reddedilebilir.

kNN algoritmasının dezavantajları aşağıda sıralanmıştır:

- Lokal yapılardan etkilenir.
- Gürültüye karşı hassastır.
- Yüksek bellek hacmi gerektirebilir.
- Yaygın olan sınıflar sonucu domine edebilirler.
- En yakın komşular uzakta olabilirler ve sonucu yanlış etkileyebilirler.

6.2.2 Atom Ağırlıkları ve Atom Çeşitliliğinin Çizgeler Üzerindeki Etkisi

Kimyasal moleküller eşlenirken moleküllerin içerisinde bulunan elementlerin türü ve atom ağırlıkları, moleküllerin yapısı hakkında önemli bilgiler vermektedir. Bu nedenle, eşleme işlemi yapılırken atom ağırlık bilgilerinin ve elementlerin isim bilgilerinin EMD algoritmasının uygulama aşamasında değerlendirilmesi eşleme probleminde kullanılabilecek önemli bir etmen olarak ortaya çıkmaktadır.

EMD algoritması uygulanırken, imza içerisindeki elemanların ağırlık bilgisi olarak elementlerin atom ağırlık bilgileri kullanılmıştır. Buradaki amaç, kimyasal moleküllerin çizgeler olarak tanımlanması ve daha sonra da eşlenmesi aşamasında

moleküllerin olabildiğince gerçeğe yakın olarak temsil edilmesidir. Atom ağırlıklarının kullanılması eşleme sırasında benzer moleküllerin birbirleriyle eşlenmesini kolaylaştıracaktır.

Karmaşık yapıdaki benzer moleküllerin birbirleriyle eşlenmesi işleminde az miktarda bulunan (yüzde ondan daha az) bazı elementler çizgenin yapısını değiştirerek benzer moleküllerin birbirleriyle eşlenmesini zorlaştırabilmektedir. Bu nedenle bir molekülün içinde bulunma oranı yüzde ondan daha az olan elementlerin molekülü temsil eden çizgeden ve dolayısıyla eşleme işleminden çıkarılması bazı durumlarda eşleme işleminin başarı oranını arttırabilmektedir. Çizelge 7.1'de ortaya konan sonuçlarda %10 kuralı adıyla belirtilen sonuçlar, molekülde bulunma oranı yüzde ondan daha az olan elementlerin çizge temsilinden çıkarılması ile elde edilmiştir.

6.2.3 Earth Mover's Distance (EMD) Algoritması Üzerinde Yapılan İyileştirme Denemeleri

EMD algoritmasının yapısı incelenecek olursa iki molekül arasındaki ayrılık oranının hesaplanmasında akışların kullanıldığı görülür, ayrıca, EMD algoritmasıyla ilgili detaylı açıklama 5. bölümde yer almaktadır.

İki molekül arasındaki ayrılık oranı hesaplanırken, molekülün içerisindeki benzer yapılar arasında düşük bir akış oranı, farklı yapılar arasında da yüksek bir akış oranı olması istenmektedir.

Bu tezde ortaya konan çalışmada, farklı elementlerin düşük akış miktarıyla birbirlerine eşleşmesi halinde sabit ve önceden belirlenmiş bir ceza parametresi iki imza arasındaki uzaklık sonucuna eklenmektedir. Burada ortaya çıkması beklenen sonuç, iki kimyasal molekül birbirleriyle eşlenirken, birbirlerinden farklı elementlere sahip moleküllerin benzerlik oranının düşük, aynı elementlere sahip, birbirlerine yakın yapısal şekildeki moleküllerin benzerlik oranının yüksek olmasıdır. Bu durumda istenmeyen eşlenmelerin yaşanması ihtimali düşürülmüştür.

Ceza parametresinin sayısal olarak belirlenmesinde kesin bir yöntem kullanılmamaktadır. Bu parametre, algoritmanın uygulama aşamasında akışlar ve benzerlik oranları incelenerek, veri setlerinin yapısına göre istenen şekilde ayarlanabilir.

7. DENEYSEL SONUÇLAR

7.1 Veri Setleri

Bu çalışmada Predictive Toxicology Challenge (PTC) [6] tarafından sunulan ve birkaç yüz kimyasal bileşenin dişi fareler (FM), dişi sıçanlar (FR), erkek fareler (MM) ve erkek sıçanlar (MR) üzerindeki toksikoloji sonuçlarını rapor eden veri seti kullanılmıştır. Veri setinde yaklaşık 400 adet kimyasal bileşen yer almaktadır ve bu bileşenler basit ve küçük olanlardan orta ölçekli ve çoklu döngüye sahip olanlara kadar değişiklik göstermektedir [7]. Veri seti US National Institute for Environmental Health Sciences – US National Toxicology Program (NTP) tarafından sağlanmaktadır.

7.2 Deneysel Düzen

Bu çalışmada C-SVC (C-Support Vector Classification) ile 10 katlı çoklu doğrulama kullanılmaktadır, öyle ki, 9 kat eğitim için 1 kat da test etmek için kullanılır. SVM ile ilgili tüm parametreler sadece eğitim sırasında optimize edilir. Yapılan tüm deney rastgeleliği arttırmak için 10 kez tekrarlanmıştır. Yapılan deneyin sonuçları Tablo 1'de gösterilmektedir.

7.3 Deneysel Sonuçlar

Yapılan çalışmada ortaya çıkan sonuçlar kabul edilen ve bilinen çalışmalar ile karşılaştırılmıştır, dahası, bu çalışmada ortaya çıkan sonuçlar WL ve NSPD kernelleri ile karşılaştırılmıştır. WL ve NSPD kernelleri ilgili açıklama ilgili çalışmalar bölümünde yer almaktadır. Bu çalışmada uygulanan metodun PTC veri seti üzerinde kullanılmasıyla ortaya çıkan sınıflandırma başarı yüzdeleri WL ve NSPD metotlarının PTC veri seti üzerindeki başarı yüzdeleri ile Tablo 1'de karşılaştırılmıştır.

Çizelge 7.1. PTC veri seti üzerinde WL, NSPD ve bu çalışmada ortaya konan metot ile 10 katlı çoklu doğrulama kullanılarak elde edilen ortalama sınıflandırma keskinlik yüzdeleri (%)

Metot / Veri Seti	FM	FR	MR	MM
Sunulan güncel metot için en iyi sonuçlar	63.89	66.95	66.96	60.75
WL	63.01	67.51	67.27	56.62
NSPD	63.57	66.94	69.39	60.13
Sunulan güncel metot (5-en yakın komşuluk)	59.35	65.54	45.91	46.74
Sunulan güncel metot (5-en yakın komşuluk ve %10 kuralı)	58.10	64.65	51.73	58.33
Sunulan güncel metot (5-en yakın komşuluk ve atom ağırlıkları = 1.0)	44.45	63.86	55.82	51.19
Sunulan güncel metot (5-en yakın komşuluk, %10 kuralı ve atom ağırlıkları = 1.0)	46.44	65.20	53.56	53.75

Sonuç olarak, ortaya konan çalışma PTC veri seti üzerinde umut vadeden sonuçlar ortaya koymaktadır. Çizelge 7.1'de görülebileceği üzere bu çalışmada ortaya teklif edilen yöntem dişi fare (FM) ve erkek fare (MM) setleri için en iyi, dişi sıçanlar (FR) için NSPD'den daha iyi ve erkek sıçanlar (MR) veri seti için WL ve NSPD'den az farkla daha kötü sonuçlar ortaya koymaktadır. Sonuçlar göstermektedir ki bu çalışmada teklif edilen yöntem, biyoenformatik alanında sınıflandırma amacıyla kullanılan diğer yöntemlerle kıyaslanabilecek başarıya sahiptir. Dahası, bu çalışmada ortaya konan yöntem biyoenformatik alanında daha önce sınıflandırma amacıyla kullanılmamış bir metot olması nedeniyle yeni bir yaklaşım ortaya koymaktadır.

Bu çalışmada ortaya konan yöntemin biyoenformatik alanı için yeni ve keşfedilmemiş olması nedeniyle, eğer uygulanması mümkün ise çeşitli geliştirme ve düzenlemeler ile yöntemin verimliliği ve etkinliği artırılabilir.

8. SONUÇ VE GELECEK ÇALIŞMALAR

Sınıflandırma, tahmin ve benzerlik ölçümleri biyoenformatik alanı için çok önemli işlemler olarak kabul edilebilir. Bu tezde biyoenformatik alanında yeni bir yöntem ortaya konmuştur, bu yöntem kimyasal moleküller arasındaki benzerlik oranlarını hesaplamaya odaklı bir çizge temelli örüntü tanıma metodudur.

Bu çalışmada uygulanan yöntem kısaca özetlenecek olursa, ilk olarak moleküller çizgeler olarak temsil edilir; sonrasında, çizgelerin ağaç temsilleri oluşturulur. Ağaç temsilleri çizge gömme işleminde kullanılır ve çizgeler alternatif domainde tanımlanmış olurlar. Gömülen noktalar arasındaki uzaklık Earth Mover's Distance (EMD) algoritması ile bulunur. EMD algoritmasının verdiği sonuçlar moleküller arasındaki ayrılık ölçüsü olarak kullanılır. En son olarak da SVM (destek vektör makineleri) kullanılarak sınıflandırma işlemi sonuçlandırılır.

Bu tezde ortaya konan yöntem biyoenformatik alanında yeni bir yaklaşım ve PTC veri seti üzerinde kıyaslanabilir sonuçlar ortaya koymaktadır. EMD algoritması öncelikli olarak görüntü erişimi ve görüntü veri tabanları için ortaya konan bir metot olmasına rağmen, bu tezde çizge gömme metoduyla birlikte farklı bir alanda kullanılmıştır. Sunulan yöntem biyoenformatik alanında sınıflandırma amacıyla daha önce önerilmemiş ve başarılı bir şekilde kullanılmamış olması nedeniyle taze ve farklıdır. Ayrıca, deneysel sonuçlara bakılacak olursa, bu tezde sunulan yöntem, biyoenformatik alanında sınıflandırma amacıyla kullanılan diğer yöntemlerle kıyaslanabilecek bir başarı oranına sahip olması nedeniyle ilgi çekicidir.

Bu tezde ortaya konan metodun geliştirilmesi için gelecekte yapılacak çeşitli çalışmalar mevcuttur. Gelecek çalışmaları maddeler halinde aşağıda sıralanmaktadır:

- Earh Mover’s Distance algoritması üzerinde, akışlar (flow) ile ilgili yapılacak çeşitli bilinçli değişikliklerin başarı oranlarını nasıl etkilediğinin araştırılması
- Çizge gömme işlemi sırasında farklı bir algoritmanın kullanılması ve bu tezde kullanılan mevcut yöntemle karşılaştırılması

- Bu tezde kullanılan metodun konuyla ilgili farklı veri setleri üzerinde denenmesi ve bu veri setleri kullanılarak diğer modern yöntemler ile karşılaştırılması
- EMD algoritması yerine farklı bir metodun kullanılması ve EMD algoritması ile karşılaştırılması
- PTC veri setinde bulunan moleküllerin yapısal olarak analizlerinin ortaya konması
- Eşleme işleminin görsel olarak tasviri ile bu tezde sunulan metodun başarı analizinin çeşitlendirilmesi
- Destek Vektör Makineleri üzerinde uygulanacak en iyileme çalışmaları

KAYNAKLAR

[1] Chih-Chung Chang, Chih-Jen Lin, LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2011, 2:27:1--27:27.

[2] Maji, S., Mehta, S. A Netflow distance between labeled graphs applications in chemoinformatics. Undergraduate Thesis. Department of Computer Science of Indian Institute of Technology, Kanpur 1.1 (2006): 1-6.

[3] Rubner, Y., Tomasi, C., Guibas, J., Adaptive Color-Image Embeddings for Database Navigation. Proceedings of the 1998 Asian Conference on Computer Vision, Hong Kong, China, 1998, 104-111.

[4] Demirci, M.F., Shokoufandeh, A., Keselman,Y., Bretzner, L., Dickinson, S., Object recognition as many-to-many feature matching. International Journal of Computer Vision 69 (2) 2006, 203–222.

[5] Demirci, M.F, Osmanlioglu, Y., Shokoufandeh, A., Dickinson, S., Efficient many-to-many feature matching under the l1 norm. Computer Vision and Image Understanding 115 (7) 2011, 976-983

[6] Helma, C., King, R., Kramer, S., Srinivasan, A., Predictive toxicology challenge, 2001.

[7] Swamidass, S. J., Chen, J., Bruand, J., Phung, P., Ralaivola, L., & Baldi, P., Kernels for small molecules and the prediction of mutagenicity, toxicity and anti-cancer activity. Bioinformatics, 2005, 21.suppl 1: i359-i368.

[8] Rubner, Y., Tomasi, C., Guibas, L. J., The earth mover's distance as a metric for image retrieval. International journal of computer vision, 2000, 40.2: 99-121.

[9] Ahuja, R. K., Magnanti, T. L., Orlin, J. B., Network Flows: Theory, Algorithms, and Applications. pages 4–7. Prentice Hall, Englewood Cliffs, New Jersey, 1993.

[10] Borgwardt, K. M., Kriegel, H. P., Shortest-Path Kernels on Graphs. In Jaiwei Han and Benjamin Wah, editors, Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM 2005), pages 74–81, Washington, DC, USA, November 2005. IEEE Computer Society.

[11] Costa, F., De Grave, K., Fast Neighborhood Subgraph Pairwise Distance Kernel. In Johannes Furnkranz and Thorsten Joachims, editors, Proceedings of the 26th International Conference on Machine Learning (ICML 2010), pages 255–262, Haifa, Israel, June 2010. Omnipress.

[12] Deshpande, M., Kuramochi, M., Wale, N., Karypis, G., Frequent Substructure-Based Approaches for Classifying Chemical Compounds. IEEE Transactions on Knowledge and Data Engineering, 17(8):1036–1050, 2005.

[13] Gärtner, T., Flach, P., Wrobel, S., On Graph Kernels: Hardness Results and Efficient Alternatives. Proceedings of the 16th Annual Conference on Computational Learning Theory and 7th Kernel Workshop (COLT/Kernel 2003) August 24-27, 2003.

[14] Kashima, H., Tsuda, K., Inokuchi, A., Marginalized Kernels Between Labeled Graphs. In Proceedings of the International Conference on Machine Learning, 2003.

[15] Kriege, N., Mutzel, P., Subgraph Matching Kernels for Attributed Graphs. In Proceedings of the International Conference on Machine Learning, 2012.

[16] Li, B., Zhu, X., Chi, L., Zhang, C., Nested Subtree Hash Kernels for Large-Scale Graph Classification over Streams. In Mohammed Javeed Zaki, Arno Siebes, Jeffrey Xu Yu, Bart Goethals, Geoffrey I. Webb, and Xindong Wu, editors, ICDM, pages 399–408. IEEE Computer Society, 2012.

[17] Mahé, P., & Vert, J. P., Graph kernels based on tree patterns for molecules. Machine Learning, 75:3–35, 2009

[18] Schietgat, L., Costa, F., Ramon, J., De Raedt, L., Effective feature construction by maximum common subgraph sampling. Machine Learning, 83:137–161, 2011.

[19] Shervashidze, N., Schweitzer, P., Van Leeuwen, E. J., Mehlhorn, K., Borgwardt, K. M., Weisfeiler-Lehman Graph Kernels. Journal of Machine Learning Research, 12:2539–2561, 2011.

[20] Shervashidze, N., Petri, T., Mehlhorn, K., Borgwardt, K. M., Vishwanathan, S. V. N., Efficient graphlet kernels for large graph comparison. Journal of Machine Learning Research - Proceedings Track, 5:488–495, 2009.

[21] Neumann, M., Patricia, N., Garnett, R., Kersting, K., Efficient Graph Kernels by Randomization. In Peter A. Flach, Tijl De Bie, and Nello Cristianini, editors, ECML/PKDD (1), volume 7523 of Lecture Notes in Computer Science, pages 378–393. Springer, 2012.

[22] Cortes, C., Vapnik, V., Support-vector networks. Machine learning 20.3 1995: 273-297.

[23] Dantzig, G.B., Application of the simplex method to a transportation problem. Activity analysis of production and allocation 13 (1951): 359-373.

[24] Ling, H., Okada, K., An efficient earth mover's distance algorithm for robust histogram comparison. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2007, 29.5: 840-853.

[25] Tan, H. K., & Ngo, C. W. Common pattern discovery using earth mover's distance and local flow maximization. In: Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on. IEEE, 2005. p. 1222-1229.

[26] Conte, D., Foggia, P., Sansone, C., & Vento, M. Thirty years of graph matching in pattern recognition. International journal of pattern recognition and artificial intelligence, 2004, 18.03: 265-298.

[27] Conte, D., Foggia, P., Sansone, C., & Vento, M.. Graph matching applications in pattern recognition and image processing. In: Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on. IEEE, 2003. p. II-21-4 vol. 3.

[28] Alpaydin, E. Introduction to Machine Learning. MIT Press, 2014.

[29] Domingos, P. A few useful things to know about machine learning. Communications of the ACM, 2012, 55.10: 78-87.

[30] Murphy, K. P. Machine learning: a probabilistic perspective. MIT press, 2012.

[31] Ayodele, T. O. Types of machine learning algorithms. INTECH Open Access Publisher, 2010.

[32] Ayhan, S., Erdoğmuş, Ş. AYHAN, Sevgi; ERDOĞMUŞ, Şenol. Destek vektör makineleriyle sınıflandırma problemlerinin çözümü için çekirdek fonksiyonu seçimi. Eskişehir Osmangazi Üniversitesi İktisadi ve İdari Bilimler Dergisi, 2014, 9.1.

[33] Markowetz, F. Support Vector Machines in Bioinformatics. Master's thesis, Mathematics Department, University of Heidelberg, 2001.

[34] Polikar R., Pattern Recognition, In Wiley Encyclopedia of Biomedical Engineering, Ed. Akay, M., New York, NY: Wiley., 2006.

[35] "Minimum Spanning Trees" erişim adresi: http://algs4.cs.princeton.edu/43mst/, erişim tarihi: 15 Haziran 2015.

[36] "Support Vector Machines" erişim adresi: http://www.statsoft.com/Textbook/Support-Vector-Machines erişim tarihi: 1 Temmuz 2015.

[37] Yakut, E., Elmas, B., Yavuz, S. Yapay Sinir Ağları ve DestekVektör Makineleri Yöntemleriyle Borsa Endeksi Tahmini. Suleyman Demirel University Journal of Faculty of Economics & Administrative Sciences, 2014, 19.2.

[38] Scholkopf B., Smola, A. J. Learning with Kernels, Support Vector Machines, Regularization, Optimization and Beyond. MIT University Press, 2002.

[39] Vishwanathan, S. V. N., Schraudolph, N. N., Kondor, R., & Borgwardt, K. M. Graph kernels. The Journal of Machine Learning Research, 2010, 11: 1201-1242.

[40] Karagülle, F, 2008, Destek Vektör Makinelerini Kullanarak Yüz Bulma, Yüksek Lisans Tezi, Trakya Üniversitesi, Fen Bilimleri Enstitüsü, Edirne.

[41] Foggia, P., Percannella, G., & Vento, M. Graph matching and learning in pattern recognition in the last 10 years. International Journal of Pattern Recognition and Artificial Intelligence, 2014, 28.01: 1450001.

[42] Conte, D., Ramel, J. Y., Sidere, N., Luqman, M. M., Gaüzère, B., Gibert, J., Vento, M. A comparison of explicit and implicit graph embedding methods for pattern recognition. In: Graph-Based Representations in Pattern Recognition. Springer Berlin Heidelberg, 2013. p. 81-90.

[43] "k-Nearest Neighbors" erişim adresi: http://www.statsoft.com/textbook/k-nearest-neighbors erişim tarihi: 1 Temmuz 2015.

[44] Gallagher, B. Matching structure and semantics: A survey on graph-based pattern matching. AAAI FS, 2006, 6: 45-53.

[45] Karagül, K., İstanbul menkul kıymetler borsası'nda işlem gören firmaların destek vektör makineleri kullanılarak sınıflandırılması, Pamukkale University Journal of Engineering Sciences, 2014, 20.5.

ÖZGEÇMİŞ

Kişisel Bilgiler

Soyadı, adı : GÖKÇER, YUNUS

Uyruğu : T.C.

Doğum tarihi ve yeri : 03.05.1990 Denizli

Medeni hali : Bekar

Telefon : 0 534 913 6763

e-mail : gokceryunus@gmail.com

Eğitim

Derece	Eğitim Birimi	Mezuniyet tarihi
Lisans	İ.D. Bilkent Üniversitesi/Bilgisayar Müh.	2013

İş Deneyimi

Yıl	Yer	Görev
2014-2015	LST Yazılım	Yazılım Geliştirme

Yabancı Dil

İngilizce, Almanca

Yayınlar

Y.Gokcer, M.Fatih Demirci, M.Tan. Graph-based Pattern Recognition for Chemical Molecule Matching. 6th International Conference on Bioinformatics Models, Methods, and Algorithms. Lisbon, Portugal, 12-15 January, 2015.