BAŞKENT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

TAVSİYE SİSTEMLERİNDE VERİ BÜTÜNLEŞTİRME

EMRAH EKMEKÇİLER

YÜKSEK LİSANS TEZİ
2012

TAVSİYE SİSTEMLERİNDE VERİ BÜTÜNLEŞTİRME

DATA INTEGRATION IN RECOMMENDATION SYSTEMS

EMRAH EKMEKÇİLER

Başkent Üniversitesi
Lisansüstü Eğitim Öğretim ve Sınav Yönetmeliğinin
BİLGİSAYAR Mühendisliği Anabilim Dalı İçin Öngördüğü
YÜKSEK LİSANS TEZİ
olarak hazırlanmıştır.
2012

“Tavsiye Sistemlerinde Veri Bütünleştirme” başlıklı bu çalışma, jürimiz tarafından, 11/01/2012 tarihinde, **BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI 'nda YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Başkan : Prof. Dr. Ziya AKTAŞ

Üye (Danışman) : Doç. Dr. Hasan OĞUL

Üye : Doç. Dr. Hamit ERDEM

ONAY

25/01/2012

Prof. Dr. Emin AKATA

Fen Bilimleri Enstitüsü Müdürü

ÖZ

TAVSİYE SİSTEMLERİNDE VERİ BÜTÜNLEŞTİRME

Emrah Ekmekçiler

Başkent Üniversitesi Fen Bilimleri Enstitüsü

Bilgisayar Mühendisliği Anabilim Dalı

Tavsiye sistemlerinde temel amaç, insanların beğenisini tahmin edip onlara bu beğenileri doğrultusunda tavsiyelerde bulunmaktır. Bu alanda en yaygın olarak kullanılan yöntemler;

- İçerik Tabanlı Filtreleme
- İşbirlikçi Filtreleme Yöntemi (İF)
- Melez Filtreleme Sistemleri
- Demografik Yöntemler
- Bilgi Bazlı Yöntemler

İçerik bazlı filtreleme yöntemleri, üzerinde çalışılan ürünle (şarkı, film, kitap vs.) ilgili bilgiler kullanılarak benzer ürünlerin önerilmesi esasına dayanır. İF yönteminin çalışma prensibi ise diğer kullanıcı ve/veya kullanıcı gruplarının öneri ve öngörülerine dayanır. Bir kullanıcıya tavsiyede bulunabilmek için diğer kullanıcıların benzer içeriklere verdiği değerler kullanılır. Kullanıcılar arasındaki benzerliklere göre diğer kullanıcıya benzer tavsiyelerde bulunulur. Melez sistemler ise iki yöntemin birlikte kullanılmasına dayanır.

Veri bütünleştirme işlemi; İF yönteminde kullanılan, kullanıcı ve bu kullanıcıların öğeler üzerinde yaptığı değerlendirmeler ile oluşturulmuş veri setlerinin iki yönlü olarak birleştirilmesidir. Bu işlemin amacı, dikey oluşturulmuş vektörlerin yatay vektörlerinin sonuna eklenmesi ile daha çok sayıda öz niteliğe sahip yeni vektörler oluşturmaktır.

Bu çalışmada yeni bir melez yöntem önerilmiş ve metaveriler kullanılarak elde edilen içerik bazlı tahminler üzerinde, “bir kullanıcının ürünlere verdiği değerler” bilgisi ile “bir içeriğe verilen bütün kullanıcı değerleri” bilgilerini birleştirererek kullanan bir çift yönlü sınıflandırma yaklaşımı denenmiştir. Oluşturulan yeni gösterim üzerinde güncel iki sınıflandırma algoritması olan “Destek Vektör Makinesı” ve “K-En Yakın Komşu” yöntemleri denenmiş, birbirleriyle ve daha önceki çalışmalarla sonuçları karşılaştırmıştır. Bu çalışmada, veri bütünleştirme yönteminin kullanımı ile tavsiye sistemlerinin daha iyi sonuçlar vermesi amançlanmıştır.

Bu alanda yöntemlerin performanslarını değerlendirirken en çok kullanılan F-ölçütüne göre daha önceki yöntemlerin üstünde bir sınıflandırma yeterliliği elde edilmiştir. Kaynak kullanım performansları açısından değerlendirildiğinde ise “Destek Vektör Makinesı” (DVM), “K-En Yakın Komşu” (KNN) algoritmasına göre daha çok bellek tüketmektedir. Süre olarak da daha uzun sürede sonuç vermektedir. KNN ise CPU kullanımında çok yüksek bir kullanım sergilemektedir.

ANAHTAR SÖZCÜKLER: Tavsiye Sistemi, Veri Bütünleştirme, İşbirlikçi Filtreleme, Destek Vektör Makinası, K - En Yakın Komşu

Danışman: Doç. Dr. Hasan OĞUL, Başkent Üniversitesi, Bilgisayar Mühendisliği Bölümü

ABSTRACT

DATA INTEGRATION IN RECOMMENDATION SYSTEMS

Emrah Ekmekçiler

Başkent University Institute of Science and Engineering

Department of Computer Engineering

The main goal of recommendation systems is to predict the user pleasure and to recommend something in the boundry of its pleasure. In recommendation systems, the most effective techniques are ;

- Content - Based Filtering
- Collaborative Filtering
- Hybrid Filtering Systems
- Demographic Filtering
- Knowledge - Based Filtering

Content - Based Filtering is mainly interested about the similarities of content between the current item (music, movie, book, etc.) and the others. In opposition of Content - Based, in Collaborative Filtering system is interested in interaction of users. The choice of the current user depends on the choices and prevision of the other users or user groups. According to prevision of users, the system recommends similar item to the current user.

Data Integration is a method that uses users evaluation vectors on items. The aim of this operation to get more complex vectors which have more features by joining the horizontal and vertical vectors of datasets.

In this thesis, a new hybrid system is proposed. On the prediction values which are produced by content based filtering using by metadata of items; "Rating of **one user** on **all items**" information and "Ratings of **all users** on **one item**" information is integrated. After this integration; "Support Vector Machine" and "K - Nearest Neighbor", the most popular classification algorithms, are executed on these new datasets. The main goal of this thesis is to increase the forecast perfomance of recommendation systems by data integration.

Finally in the comparison of this technique and other techniques shows us that the data integration technique gives better results according to F- measure. On the other hand, there is also a comparison according to resource usage. "Support Vector Machine" (SVM) uses more memory and takes more time to solve the problem. However "K - Nearest Neighbor" (KNN) uses much more CPU.

KEYWORDS: Recommendation System, Data Integration, Collaborative Filtering, Support Vector Machine, K - Nearest Neighbor

Advisor: Assoc. Prof. Dr. Hasan OĞUL, Başkent University, Department of Computer Engineering

İÇİNDEKİLER

ÇİZELGE DİZİNİ

ŞEKİL DİZİNİ

1. GİRİŞ

Günümüzde, internetin de kişisel olarak hayatımıza girmesi ile birlikte insanlar tarafından ulaşılabilir olan bilgi hacmi günden güne katlanarak büyümektedir. Varolan kitap, film, haber, reklam ve çevrim içi bilgi miktarı şaşırtıcı boyutlardadır. Bu büyüklükte bir veri yığını içerisinden, işe yarar bilgiyi elde edebilmek, gerekli olana ulaşmak, kullanıcının beğenisine göre filtreleme yapmak, ihtiyaçları karşılayacak bilgiye ulaşabilmek bir kullanıcı için oldukça zor bir hal almış durumdadır [1].

Yıllar içinde, insanların daha doğru ve ilgili bilgiye ulaşması için "Bilgi Filtreleme Sistemleri" kavramı ortaya çıkmıştır. Bilgi Filtreleme Sistemlerinin amacı; kullanıcıya otomatik olarak işlenmiş, gereksiz ve istenmeyen verilerden arındırılmış saf bilgiyi sunmaktır. Bu işlemi yaparken de kullanıcıların profillerinin kıyaslanması ve benzer karakterde olan kullanıcılar referans alınarak benzer bilgileri sunmak temel alınmıştır. "Tavsiye Sistemleri" de bu noktada ortaya çıkmıştır.

Bu dönemde tavsiyesi sistemleri hemen hemen her alanda kullanılabilmektedir. Kitap, haber, müzik bu başlıkların en önde gelenlerindendir. Tavsiye sistemlerinde temel amaç, insanların beğenisini tahmin edip onlara bu beğenileri doğrultusunda tavsiyelerde bulunmaktır. Bu amaca ulaşmak için beş yöntem kullanılmaktadır:

- İçerik Bazlı Yöntem(Tavsiye edilecek öğenin temel içerik bilgilerine dayanarak)
- İşbirlikçi Filtreleme Yöntemi (İF - Kullanıcı profillerindeki benzerliklere dayanarak)
- Hibrid Yöntemler("İçerik Bazlı" ve İF yöntemlerin birleşimine dayanarak)
- Demografik Yöntemler

- Bilgi Bazlı Yöntemler

Bu çalışmada amaçlanan; “bir kullanıcının içeriklere verdiği değerler” vektörleri ile “bir içeriğe bütün kullanıcılar tarafından verilen değerler” vektörlerini birleştirerek daha çok öz nitelikli bir vektör elde etmek ve oluşan bu yeni vektörler üzerinde en temel iki sınıflandırma algoritmalası olan “Destek Vektör Makinası” ve “K-En Yakın Komşu” ile sınıflandırma becerilerinin gelişimininin ölçülmesidir.

Tezin geri kalanı şu şekilde organize edilmiştir :

İkinci kısımda, tavsiye sistemleriyle ilgili çalışmalara yer verilmiştir. Tavsiye sistemlerinin içeriği, çeşitleri ve amaçları örneklerle açıklanmaktadır. Sonraki aşamada tavsiye sistemlerinin tipleri, yöntemleri ve tekniklerine değinilmiştir. En sonunda tavsiye sistemlerinin yaşadığı zorluklardan bahsedilmiştir.

Üçüncü kısımda, makine öğrenme ve sınıflandırma algoritmalarına yer verilmiştir. Bu çalışmada kullanılan; uygulaması en kolay olan “en yakın komşu algoritması “ ve “destek vektör makineları” algoritmik yapıları ile anlatılmıştır.

Dördüncü kısım, yapılan çalışmanın detaylandırılmasından oluşmaktadır. Veri bütünleştirme işlemleri ve sınıflandırma algoritmalarının nasıl kullanıldığına dair bilgiler içermektedir.

Beşinci kısım, bu çalışmanın değerledirme bölümünü oluşturmaktadır. Burada sistemde kullanılan veri setlerinin yapısı ve değerlendirmenin hangi ölüçülerde yapıldığına değinilmiştir. Sonraki aşamada ise çıkan sonuçlar ve bu sonuçların belirtilen yöntemlerle değerlendirme aşamasına yer verilmiştir. Son olarak diğer çalışma yöntemleriyle sonuç kıyaslaması yapılarak, sistemin son değerlendirmesi yapılmıştır.

Altıncı kısımda, tezin sonlandırılması ve bu çalışma ile ilgili negatif ve pozitif noktalar ele alınarak, dezavantajları ortadan kaldırmak için neler yapılabileceği ile ilgili tartışma yer almaktadır.

2. TAVSİYE SİSTEMLERİ

Bu bölümde tavsiye sistemleri ile ilgili genel kavramlara ve terminolojiye değinilecektir. Tavsiye sistemlerinde kullanılan farklı teknikler, bu teknikleri kullanan tavsiye sistemleri ve kullanılan algoritmalar açıklanacaktır.

2.1. Giriş

Son yıllarda bilgi hacminde akıl almaz bir büyümeye şahit olduk. İnternet destekli medya ile bilgi akışı inanılmaz bir hıza ulaşmıştır. Bu sayede insanlar her konuda çok sayıda alternatifle karşı karşıya kalmışlardır.

Geçmiş yıllarda herhangi bir ürüne, habere ya da bilgiye ulaşmak için çaba sarf etmek gerekirken günümüzde aynı çabayı belki çok daha fazlasını mevcut bilgi hacmi içinde filtreleme yapmaya harcamak gerekmektedir. Geçmişte seyahat planlamasında, film ya da kitap seçiminde, giyim hatta gıda alışverişlerinde yakın bir arkadaş tavsiyesi, bir uzman makalesi gibi faktörler tercihleri etkilerken; alternatiflerin artması neticesinde kişinin kendisi için en uygun tercihi yapabilmesine yönelik en değerli bilgiler tavsiye sistemleri aracılığıyla kişilere sunulmakta; yani teknoloji gerçekten istek veya ihtiyaç duyduğumuz her türlü ürün, bilgi ve benzeri konuda genel ve özel tercihleri analiz eden, önerilerde bulunan, alternatifleri tercihler doğrultusunda elimine etmek suretiyle bilgi yüküyle boğulmayı engelleyen ve doğru alternatiflere yönlendiren bir araç konumuna gelmiştir. Tavsiye Sisteminin hedefi kullanıcıların ilgilenebileceği her türlü ürün veya öğe için anlamlı öneriler üretmektir [2].

Hepsiburada [3], Amazon [4] gibi siteler e-ticarette tavsiye sistemi kullanan yerli ve yabancı örneklerdir. Türkiye'de Hepsiburada.com da olduğu gibi incelenen ürünle ilgili bilgilerin hemen ardından "Bu ürünü tercih edenlerin satın aldığı diğer ürünler", "Bu ürün en çok aşağıdaki ürünlerle kıyaslandı" başlığı altında site kullanıcılarının kıyasladıkları ürünler ve aynı ürün grubunda en çok satılanların listesi yer almaktadır .

Richard MacManus un “A Guide to Recommender Systems” (Tavsiye Sistemleri Rehberi) isimli makalesinde “Tavsiyeler Kralı” olarak tanımlanan Amazon da ise seçilen ürüne ilişkin bilgilerin hemen ardından “Bu ürünü satın alanların satın aldığı diğer ürünler”, editör yorumunu takiben tüketici yorumları, ayrıca genel olarak tüketiciler tarafından satın alınan aynı ürün grubundaki diğer ürünler, incelenen ürünü inceledikten sonra diğer tüketicilerin satın aldıkları ürünler, kullanıcıların bu ürünle ilgili etiketleri, müşteri tartışma platformu, “Listmania” adı altında ilişkili diğer ürün gruplarından favori listeleri, “So You'd Like to...” başlığı altında farklı ürün gruplarından öneriler yer almaktadır [5].

Bu tür tavsiye motorlarının tasarımı hizmet verilen alana ve kullanılabilir verilerin belirli özelliklerine göre değişkenlik gösterir.

2.2. Tavsiye Sistemlerinde Kullanılan Yöntemler

Tavsiye sistemlerinin tasarımında kullanılan teknikler hizmet verilen alana ve kullanılabilir verilerin belirli özelliklerine göre değişkenlik gösterir. Tavsiye sistemlerinde kullanılan teknikler temel olarak Tablo 2.1’ de göründüğü üzere beş başlık altında toplanabilir [2].

Tavsiyede bulunulacak öğelerin kümesi Ö, tercihleri belli kullanıcıların kümesi K, tavsiyesi şekillendirilecek kullanıcı k, k’nın tercihini tahmin etmek için kullanılacak öğe ise ö olarak belirlenmiştir.

Tablo 2.1 - Tavsiye Sistemi Teknikleri

Kullanılan Teknik	Arka Plan	Girdi	Süreç
İçerik Bazlı	Ö deki öğelerin özellikleri	Ö deki öğeler için k'nın derecelendirmeleri	k'nın derecelendirmelerine davranışına uyan bir sınıflandırıcı oluşturup ö üzerinde kullanılması
İşbirlikçi	K'nın Ö de derecelendirdiği öğelerin değerleri	Ö deki öğeler için u dan alınan derecelendirme değerleri	K içinde k ya benzer özellikte kullanıcıların belirlenmesi ve ö üzerinde derecelendirmelerinden yola çıkarak dış değerleme yapılması
Demografik	K nın demografik bilgileri ve Ö deki öğelerle ilgili derecendirmeleri	k nın demografik bilgileri	Demografik bilgileri k ya benzer olan kullanıcıların tanımlanması ve onların ö üzerinde derecelendirmelerinden yola çıkarak dış değerleme yapılması
Bilgi Bazlı	Ö deki öğelerin özellikleri ve bu öğelerin kullanıcıların ihtiyaçlarını nasıl karşılayacağına ilişkin bilgi	k nın ihtiyaç ve ilgilerinin tanımlanması	k'nın ihtiyacı ö eşleşmesi sonucunun çıkarılması

Tavsiye sistemleri tüketicilerin ilgi ve tercihlerini öğrenen, buna göre önerilerde bulunan, tüketicinin seçim sürecinde daha sağlıklı karar vermesini sağlayan akıllı

mekanizmalardır. Bu nedenle tavsiye sistemleri kullanılabilecek işlenebilir verilere, örtülü ve açık kullanıcı geri bildirimlerine ve hitap edilen alanın özelliklerine göre modellenir. Tavsiye sistemi için girdi (input) kullanılan filtreleme algoritmasına göre değişkenlik gösterir. Çeşitli filtreleme algoritmalarında girdi olarak aşağıdakiler kullanılır:

- Derecelendirme (oylama): nümerik bir gösterge çizelgesi üzerinde örneğin 1 (en kötü) den 10 a (en iyi) kadar kullanıcı tarafından derecelendirme yapılabileceği gibi, yine sıklıkla kullanılan ikili derecelendirme (0 ya da 1 şeklinde) yapılabilir. Ayrıca bu bilgiler kullanıcının alışveriş geçmişi, web kütük bilgileri aracılığıyla da edinilebilir.

- Demografik bilgiler: edinilmesi güç olsa da yaş, cinsiyet, kullanıcıların eğitim durumu ve benzeri gibi kullanıcıdan açık biçimde istenen bilgilerdir.

- İçerik bilgisi: kullanıcı tarafından derecelendirilen öğelerin metinsel olarak analizinden elde edilen bilgilerdir.

Örneğin popüler bir film sitesi olan imdb.com takipçileri filme özel nitelikler oyuncu, yönetmen, konu ve benzeri hakkında detaylı olarak bilgilendirilirken, diğer kullanıcıların yorumlarına, incelenen öğeyi incelemiş ya da izlemiş olan kullanıcıların inceledikleri ya da izledikleri diğer film bilgilerine erişip, inceledikleri filmler için 1 (en kötü) 10 (en iyi) aralığında puanlama yapabilirler. Kullanıcı ve öğeler arasındaki bu etkileşimler kayıt altına alınır. Ayrıca sistemde demografik bilgiler, ayrı ayrı ürünlerin özellikleri, kullanıcı ya da öğenin profilindeki öznitelikler tutulur.

Tavsiye sistemleri en uygun öğe-kullanıcı çiftini eşleştirmek üzere toplanan bu verilerin analizi yöntemleri ve analizde girdi olarak kullanılacak veriler konusunda farklılıklar gösterirler. İşbirlikçi filtreleme sistemleri yalnızca etkileşimlerin tarihçesini analiz ederken, içerik bazlı filtreleme sistemleri profilin özniteliklerini temel alır. Hibrid

(karma) teknik ise işbirlikçi filtreleme ve içerik bazlı filtreleme tekniklerinin bir arada kullanılmasından meydana gelir.

Tavsiye sistemlerinin yaygın olarak kullanıldığı alan doğru tüketici ile doğru ürünün eşleştirilmesinin gerektiği e-ticaret uygulamalarıdır. İnternet üzerinde yapılan satışların büyük bölümü en popüler ürünler üzerinden gerçekleşir. Ancak çok satılan ürünlerin kar marjı düşüktür. Yaygın olarak bilinmeyen ancak kullanıcın beğenmesi muhtemel ürünün tespiti karlılık açısından bakıldığında tavsiye sistemlerinin başarısının ölçütüdür.

Uzun Kuyruk bir perakendecilik kavramı olmakla beraber çok miktarda satılan az ürün yerine az miktarda satılan çok ürünün satılmasını ifade eder. Kavram Chris Anderson tarafından 2004′de Wired'da yayınlanan makalesi ile ortaya çıkmıştır. Bu yazıda Chris Anderson Amazon.com u bu tip bir strateji uygulayan kuruluşlardan biri olarak tanımlamıştır [6].

Bu stratejiyi uygulayan kuruluşların dağıtım ve stok masrafları onların küçük ölçeklerde satılan bulunması zor ürünleri çok miktarda müşteriye satarak hatırı sayılır miktarlarda kar elde etmelerine olanak tanır. İşte bu her bir ürün için az miktarda fakat çok fazla miktarda ürün için geçerli olan satış şekline Uzun Kuyruk denir.

Yapılan analizler gösteriyor ki, Amazon gibi Uzun Kuyruk etkisini başarıyla kullanan firmaların gelirlerinin önemli bir bölümü çok satılan ürünlerden değil, grafiğin uzun kuyruk bölümünden gelmektedir [7].

Tavsiye sistemlerinin e-ticaret uygulamalarında kullanım biçimleri aşağıdaki gibi örneklenebilir:

- Kişisel ana sayfa tavsiyesi: Kullanıcının geçmiş hareketleri göz önüne alınarak genel profiline göre ana sayfa oluşturulması

- Ürün sayfası benzer ürün / yeni ürün tavsiyesi: Kullanıcının anlık amacı göz önüne alınarak yeni ürün tavsiyesi,

- Çapraz ürün satışı: beraber satılan ürünlerin gösterilmesi, toplu indirim uygulanması.

- Kişiselleştirilmiş kampanya / reklam / e-mail tavsiyesi: kullanıcının ilgilenebileceği kampanyaların otomatik oluşturulması ve reklam e-maillerinin kişiye göre hazırlanması. (Tavsiye Sistemleri: Long Tail (Uzun Kuyruk) İle Karlılığı Artırmak (Deniz Oktar -iletken recommendation technologies))

Referans özelliklerin seçimine bağlı olarak tahmin sistemleri temel olarak işbirlikçi filtrelemeyi (collaborative filtering) kullanan sistemler ve içerik bazlı filtrelemeyi (content-based filtering) kullanan sistemler olarak ikiye ayrılır. Her iki yöntemin de avantajları ve dezavantajları bulunmaktadır. Her iki yöntem için de var olan uygulama zorlukları ve dezavantajları elimine etmek için her ikisini de içinde barındıran hibrid (karma) filtreleme tahmin sistemleri kullanılabilir [8].

2.2.1. İçerik bazlı filtreleme yöntemi

İçerik bazlı tavsiye sistemi öğenin içeriği ve kullanıcı tercihleri arsındaki korelasyona göre öğe önerisinde bulunur [9]. İçerik bazlı filtreleme sistemi kullanıcıdan gelen örtülü ve açık geri bildirimlere göre kullanıcı tercihlerini öğrenir. Kullanıcı profili kullanıcı tercihlerini yansıtacak biçimde inşa edilir.

Kullanıcının kendi isteğiyle örneğin siteye girişte doldurmaya zorunlu olduğu formlar aracılığıyla kullanıcının yaş, cinsiyet, eğitim ve benzeri bilgileri, ayrıca “beğendim” ya da “beğenmedim” şeklinde fikir beyanı açık geri bildirimdir. Kullanıcının haberdar olmadığı, davranışlarının izlenmesi sonucu edinilen bilgiler ise örtük bilgilerdir.

Kullanıcının beğenilerinin belirlenmesinde ve kullanıcı profilinin oluşturulmasında otomatik öğrenme ve bilgi toplama algoritmaları kullanılmaktadır. Genellikle kullanıcı ve öğe profilini karakterize etmek için “vektör uzayı modeli” (vector space model – VSM) kullanılmaktadır [10].

“Kişiselleştirilmiş tavsiye sistemi” (Personalized Recommender System – PRES) diğer bir içerik bazlı filtreleme sistemidir [9].

İçerik bazlı sistemin avantajı toplanan örtük bilgilerin sistemin çalıştırılması için yeterli olmasıdır. Ancak bu sistem için de başlangıçta kullanıcı / ürün değerlendirmeleri konusundaki verilerin az olması isabetli tavsiye verilmesi dolayısıyla sistemin güvenilirliği açısından dezavantaj oluşturmaktadır.

2.2.2. İşbirlikçi filtreleme yöntemi

İşbirlikçi filtreleme müzik, film, kitap gibi pek çok alanda uygulanabilir olması; bilinen çeşitli algoritmalarının bulunması ve e-ticaret uygulamalarında yaygın kullanımıyla tavsiye sistemlerinde öneri üretmek için kullanılan en önemli yaklaşımdır.

Temel fikir kullanıcıların ürünleri beğenilerine göre örtük ya da açık olarak derecelendirmek suretiyle değerlendirmeleri; ve geçmişte benzer beğenileri olan kullanıcıların gelecekte de benzer zevklere sahip olacağıdır [11].

İşbirlikçi filtreleme yaklaşımı kullanıcıların ortak özelliklerini yada öğeler arasındaki benzerlikleri temel alan filtreleme sürecidir.

İşbirlikçi filtreleme, haber grupları tarafından süzülmüş dokümanların bir dizi kullanıcıya ulaştırılması amacıyla “Tapestry” adı ile ilk ticari tahmin sistemi olarak ortaya çıkmıştır. Amaç kullanıcıların yoğun bilgi akışı içinde boğulmadan sosyal işbirliğinin arttırılmasıdır. İçerik bazlı filtrelemenin kullanıcıya yapılan önerilerde genel

kullanıcı tabanına erişim ve bilgiden açıkça faydalanma yönünde “işbirlikçi “olmadığı söylenebilir [3].

Tavsiye sisteminde ilk formülasyonlar daha yaygın istatistiki uygulamalar ve otomatik öğrenme (machine learning) literatüründen ziyade korelasyon istatistikleri ve tahmine dayalı modelleme üzerine kuruluydu.

İşbirlikçi filtreleme problemi çözüm kalitesini arttırmaya yönelik boyutluluk azaltma tekniklerinin kullanılmasını sağlayan sınıflandırmayı ortaya koymuştur. Aynı zamanda içerik bazlı filtreleme yöntemi ile işbirlikçi filtreleme ve tavsiye sistemlerinin mimarisiyle ilgili alan bilgisini birleştirmek üzere birçok girişimde bulunulmuştur.

İşbirlikçi filtreleme sistemleri komşuluğa dayalı (neighborhood-based) ve modellemeye dayalı (model-based) yaklaşım olmak üzere genel olarak iki alt başlığa ayrılmıştır. Komşuluğa dayalı yaklaşımda kullanıcı için benzer kullanıcılardan oluşan bir küme belirlenir ve kümedeki kullanıcıların derecelendirme değerlerinin kombinasyonundan faydalanılarak kullanıcı için tahminler yürütülür.

Komşuluğa dayalı modellemede kullanıcıların geçmişte benzer tercihleri varsa gelecekte de benzer tercihlere sahip olacakları; kullanıcı tercihlerinin zaman geçtikçe sabit kalacağı ve süreklilik arz edeceği varsayılır.

Modele dayalı yaklaşım kullanıcıların derecelendirme davranışlarında temel bir yapının olduğunu varsayar ve geçmiş derecelendirmelerine dayanarak tahmin modelleri çıkartır. Kullanıcı derecelendirmeleri için istatistikî model parametrelerini tahmin ederek tavsiyeler sağlar.

İşbirlikçi filtreleme yaklaşımı kullanıcıların ortak noktaları ve öğelerin benzerlikleri üzerinden bilgileri filtreleme sürecidir. Bu yöntemi kullanan tavsiye sistemi hedef

kullanıcıya hedef kullanıcıyla benzer tercihlere sahip diğer kullanıcıların görüşlerine dayalı tavsiyelerde bulunur [12].

İşbirlikçi filtreleme kullanan tavsiye sistemleri birçok sitede doğrudan ya da Amazon, hepsiburada gibi sitelerde karma olarak uygulanmaktadır.

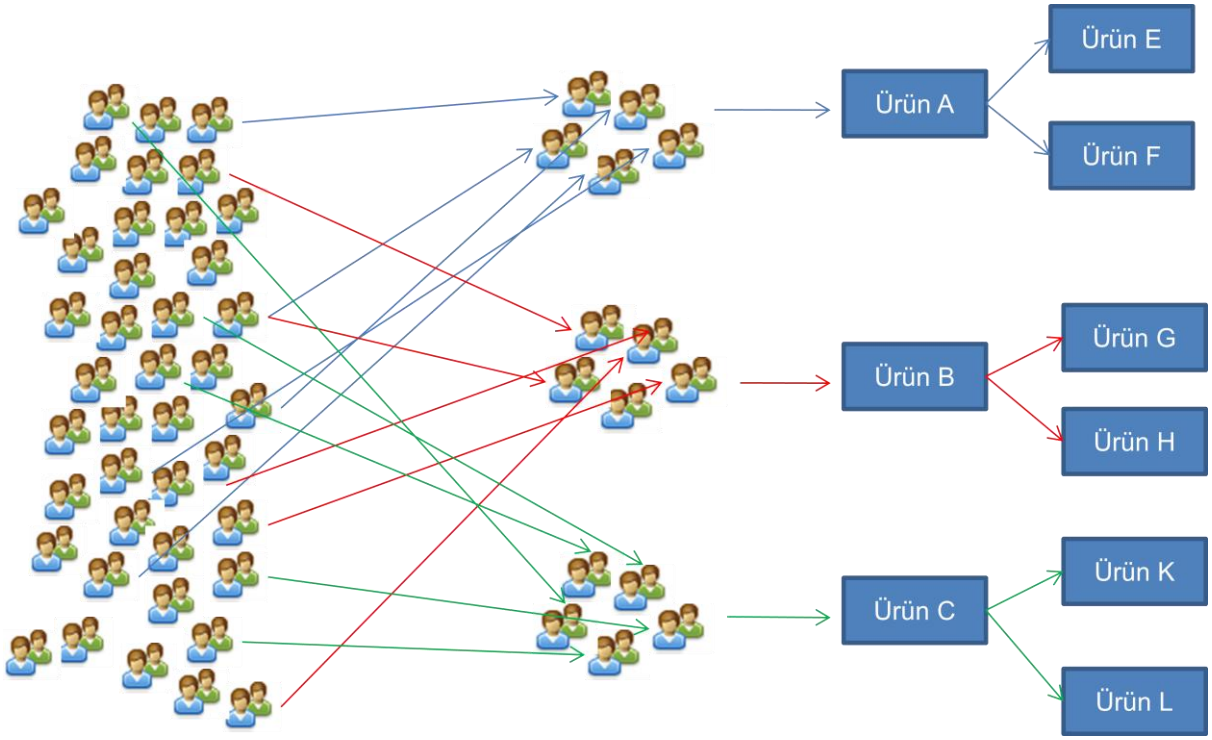
Çeşitli işbirlikçi filtreleme algoritmaları bulunmaktadır. Geleneksel algoritmalar tüm kullanıcı ve tahmin çiftleri arasındaki benzer değerleri belirli bir kullanıcı için hesaplamaktadır. Korelasyon, ortalama kare farkı veya vektör benzerlik dâhil çeşitli benzerlik ölçümleri bulunmaktadır. Bayes ağ modelleri, bağımlılık ağ modelleri veya modelleri kümeleme gibi diğer algoritmalar altta yatan kullanıcı tercihlerinin modelini oluşturur.

Tavsiye sistemleriyle ilgili çalışmalar işbirlikçi filtrelemenin yaygın kullanımını göstermektedir. Genel olarak, işbirlikçi filtreleme kullanan tavsiye sistemleri aşağıdaki adımları takip eder [13]:

- Çok sayıda insanın davranışlarını kaydeder,

- Geçerli kullanıcı için geçmişteki davranışları kişiye benzer olan diğer insanlar bir grup “komşu” (neighbors) seçer,

- Diğer komşuların (other neighbors) davranışlarına bağlı olarak kullanıcın gelecekteki davranışını tahmin eder.



Şekil 2.1 - İşbirlikçi Filtreleme

İşbirlikçi filtreleme yöntemi belirli bir kullanıcıya öneride bulunmak için birçok kullanıcıdan gelen örtülü veya açık değerlendirmelerden faydalanır. Eğer kullanıcı hakkındaki bilgi az ise ya da kullanıcı nadir tercihlere sahipse tavsiye sisteminde yalnızca işbirlikçi filtreleme yönteminin kullanılması tavsiye sisteminin başarısızlığına neden olabilir [12].

Bununla birlikte işbirlikçi filtreleme yöntemi, ürünün içeriği hakkında bilgi yoksa bile öneri yapma kabiliyetine sahiptir.

2.2.3. Demografik yöntemler

Yaş, cinsiyet, eğitim, ülke ve benzeri bilgilerin kullanıcıların gruplanması amacıyla kullanılmasıdır. Bu filtreleme tekniğinde kullanıcının demografik bilgileri mevcut gruplarla karşılaştırılır ve kullanıcıyla en yakın ilgisi olan grup tespit edilir.

Gruplandırma bu teknik için temel unsurdur. Ayrıca öğeler özelliklerine göre ayrılır. Sonunda bu sınıflar karşılaştırılarak en uygun öğe grubu kullanıcıya önerilir.

Demografik filtreleme kullanılırken kullanıcıları mahremiyeti konusuna hassasiyet gösterilmesi gerekmektedir.

Daha doğru sonuçlara ulaşmak için genellikle demografik filtreleme farklı bir filtreleme tekniği ile birlikte kullanılır. Örneğin otomatik müzik öneri sistemlerinde demografik filtreleme ile işbirlikçi filtreleme yöntemlerinin birlikte kullanımı daha tatmin edici sonuçlar alınmasını sağlamıştır.

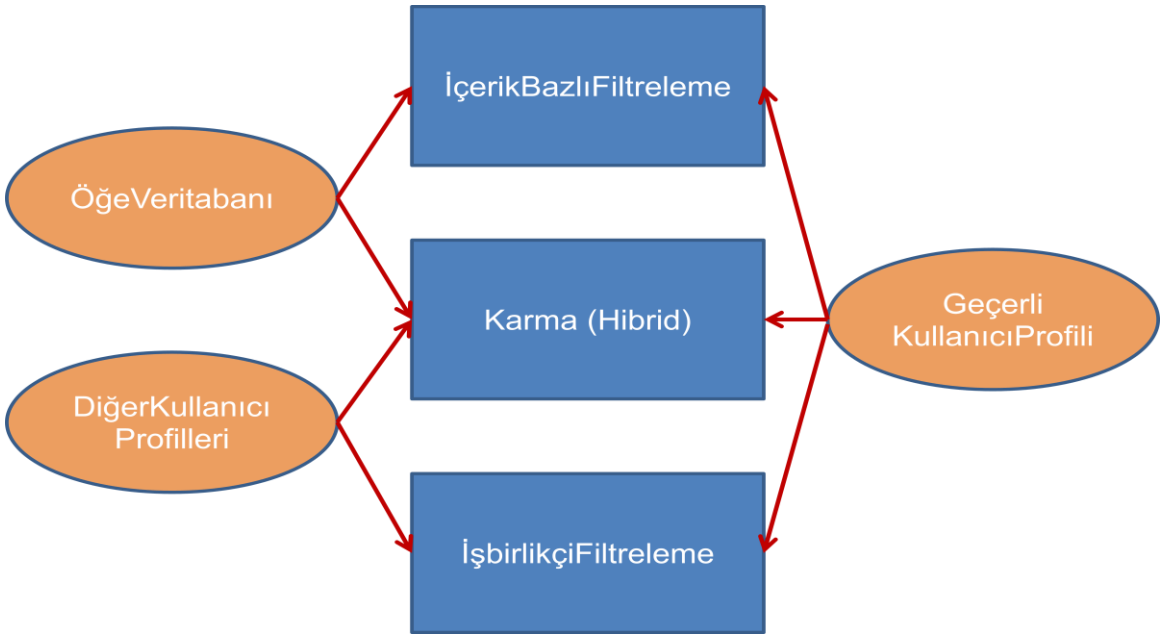
2.2.4. Bilgi bazlı yöntemler

Bilgi bazlı sistemlerde öğeler ve tüm özelliklerinin açık bir biçimde belirtildiği katalog bilgileri, fonksiyonel bilgiler ve kullanıcı bilgileri bulunur. Kullanıcı bilgileri genellikle demografik bilgilerden oluşur.

Bu tür sistemlerin avantajı katalog bilgilerini sağlamada sorun yaşanmayacağından ilk çalıştırma probleminin (cold start problem) yaşanmamasıdır. Ancak bu sistemler önerileri sadece eldeki bilgilerle yapabildiklerinden işbirlikçi sistemler gibi yeni bilgilerle başa çıkamazlar. Ayrıca bilgi azlığı bu sistemler için de dar boğaz oluşturur.

2.2.5. Hibrid (karma) yöntemler

Her tip tavsiye tekniğinin zayıf ve güçlü yönleri bulunmaktadır. Sadece tek tekniğin kullanımıyla üstesinden gelinemeyecek sorunlarla başa çıkmak ve tüm tekniklerin kendi avantajlarından faydalanarak daha isabetli tavsiyelerde bulunmak üzere birden fazla yöntemin bir arada kullanılması karma tekniktir.



Şekil 2.2 - Hibrid tavsiye sistemi

2.3. Tavsiye Sistemlerinde Karşılaşılan Başlıca Problemler

2.3.1. Seyreklik (sparsity)

Sağlanan verilerin seyrekliğidir. Örneğin öğelerin kullanıcılar tarafından değerlendirilmemesi sonucunda az sayıda değere sahip olunmasıdır. Bu aynı

derecelendirme değerlerine sahip kullanıcıların bulunması olasılığında düşüşe neden olduğundan veri seyrekliği işbirlikçi filtreleme yönteminin uygulanabilirliğini dolayısıyla da tavsiyenin güvenilirliği konusunda tereddüt yaratmaktadır [14].

2.3.2. İlk çalıştırma problemi (the cold-start problem)

Öğelerin kullanıcılar tarafından henüz hiçbir değerlendirmeye tabi tutulmamış olması durumunda; hangi kullanıcıların hangi öğeler için ne şekilde değerlendirme yapacağına dair herhangi bir çıkarım yapılamamaktadır.

Kullanıcının ya da ürünün ya da her ikisinin birden yeni olması durumunda hem işbirlikçi filtreleme hem de içerik bazlı filtreleme için kullanıcıların sisteme giriş yapması ve ardından değerlendirme yapmaya başlamasıyla az da olsa sağlanan verilerle tavsiyeler de bulunulabilir [14].

2.3.3. Ölçeklenebilirlik (scalability)

Mevcut kullanıcı ve öğe sayısındaki artış hızla tavsiyede bulunması gereken geleneksel işbirlikçi algoritmaları kullanan sistemler için sorun teşkil etmektedir. Bu nedenle ölçeklenebilirlik hızla büyüyen sistemlerde arzu edilen bir özelliktir [14].

2.3.4. Sahtekârlık (fraud)

Tavsiye sistemlerinin özellikle e-ticarete yönelik sitelerde kullanılmakta ve e-ticarette rekabet hızla artmaktadır. Tavsiye sistemleri karlılık üzerinde önemli bir etkiye sahiptir. Bu nedenle satıcılar kendi ürünleri için olumlu değerlendirmelerle kendi ürünlerinin satışını arttırmaya çalışmakta (itme etkisi - push attacks); rakip ürünleri düşük derecelendirerek rakip ürünlerin çekiciliğini azaltmaya çalışmaktadırlar (nükleer saldırı - nuke attacks) [14].

Tavsiye sistemleri artan kullanıcı ve ürün sayıları, değişen kullanıcı profilleri, artan rekabet düşünüldüğünde sürekli yeniden yapılandırılmalıdır. Ayrıca sözü kullanıcı ve ürün artışının ölçeklenme zorlukları yanında sitenin hafıza ve işlemci ihtiyacı değerlendirilerek donanım sorununu da beraberinde getirdiği söylenebilir. Ayrıca yine yukarıda sözü edilen sahtecilik örnekleri filtreleme sistemlerini yanıltmaktadır.

3. MAKİNE ÖĞRENME VE SINIFLANDIRMA TEKNiKLERİ

3.1. Sınıflandırma

Sınıflandırma işi pek çok insan faaliyetinde ortaya çıkar. Sınıflandırma kavramı en geniş anlamda mevcut bilgilere dayanarak karar alma ve tahmin yürütme işlerini, sınıflandırma işlemi ise daha sonra tekrarlayan yeni durumlarda yargıda bulunmak için kullanılan bir dizi biçimsel yöntemleri barındırır [15].

Veri sınıflandırma verilerin en etkili ve etkin biçimde kullanılabilmesi için kategorizasyonudur. Veri depolamada en temel yaklaşım verinin kritik değerine, veriye erişim ihtiyacındaki sıklığa göre depolanmasıdır. Kritik ve sık kullanılan veriler en hızlı ve muhtemelen daha pahalı medya üzerinde depolanırken, diğerleri daha yavaş ve muhtemelen daha ucuz medya üzerinde depolanabilir. Bu tür bir sınıflandırma depolanmış verinin çoklu amaçlarla (teknik, yasal ya da ekonomik ve benzeri) kullanımını optimize eder.

Sınıflandırma temel olarak üç temel alt başlıkta tanımlanabilir: istatistikî, otomatik öğrenme ve yapay sinir ağı. Bunlar büyük ölçüde farklı mesleki ve akademik grupları kapsar ve farklı konular üzerinde durur. Tüm bu grupların bazı ortak amaçları olabilir ve bu grupların hepsi için sıralanan amaçlarla ortak prosedürler türetilmeye çalışılmıştır:

- kullanıcının karar verme davranışını; sınırları aşmamak koşuluyla ve tutarlılık avantajını sağlayacak şekilde değişken bir ölçüde benzeştirmek,

- çok çeşitli problemleri ele almak, yeterli verinin sağlanması ve genel olması,

- kanıtlanmış başarıya sahip kullanışlı yapıların kullanılması.

3.1.1. İstatistikî yaklaşım

İstatistikî topluluk içinde tanımlanabilen iki temel çalışma aşaması vardır. Birincisi "klasik" aşama Fisher'in doğrusal ayrım üzerindeki ilk çalışmasının türevleri üzerinde çalışır. İkincisi "modern" aşama; modelin pek çoğu her sınıf için yeri geldiğinde sınıflandırma kurallarını oluşturabilecek ortak dağılım özelliklerinin tahminini sağlamaya yönelik daha esnek sınıflardan faydalanır [15].

3.1.2. Otomatik öğrenme

Otomatik öğrenme genellikle, bir dizi örnekten öğrenerek mantıksal ve ikili işlemlere dayanan otomatik hesaplama prosedürlerini kapsar. Kullanıcı tarafından anlaşılacak basitlikte sınıflandırma deyimleri oluşturmayı amaçlar. Karar verme sürecine ışık tutmak için kullanıcın karar verme sürecini tatminkar biçimde taklit etmelidir. İstatistiki yaklaşımlara benzer olarak geçmiş bilgilerden geliştirme aşamasında faydalanılabilir fakat işlemlerde insan müdahalesi yok kabul edilir [15].

3.1.3. Yapay sinir ağları

Yapay sinir ağları alanı insan beynini anlamak ve taklit etmek ile insanlığa duyulan hayranlıktan; daha geniş anlamda konuşma, dil kullanımı gibi insan davranışlarını kopyalamaktan; uygulamadaki ticari, bilimsel ve mühendislik disiplinlerine örüntü tanıma, modelleme ve tahmine kadar değişen farklı pek çok kaynaktan ortaya çıkmıştır. Araştırmacılar için teknolojinin takibi akademik alanda ve sanayi alanında ayrıca bilim ve mühendisliğin çeşitli alanlarında itici güçtür. Yapay sinir ağlarında otomatik öğrenme de olduğu gibi teknolojik ilerlemenin heyecanı zekânın yeniden üretilmesinin zorluğuyla desteklenir [16].

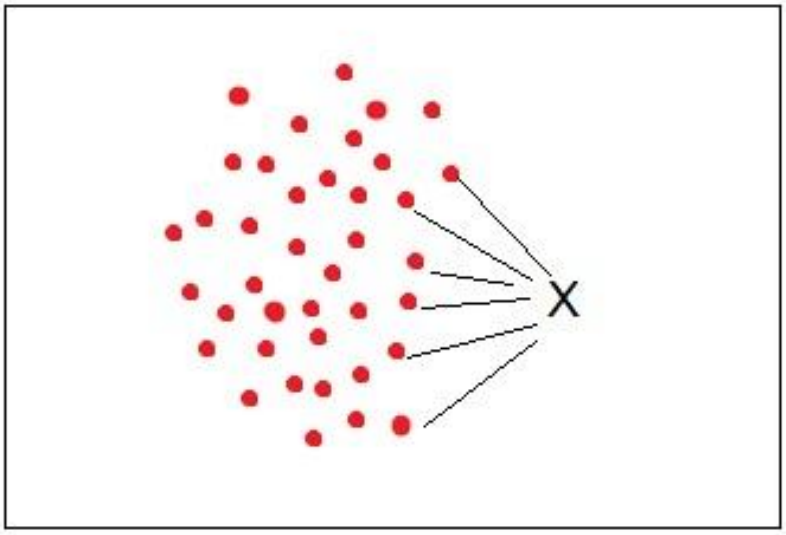
Yapay sinir ağı yaklaşımları bazı istatistikî tekniklerin karmaşıklığıyla insan zekâsını taklit etme amacındaki otomatik öğrenmeyi birleştirir. Daha çok bilinçaltı düzeyinde yapıldığından kullanıcı tarafından net öğrenilmiş kavramlar yaratabilecek eşlik eden bir yeteneği bulunmamaktadır.

Sınıflandırmaya ilişkin üç temel yaklaşım incelendiğinde teknik yapı ve profesyonel geçmiş arasındaki uyum kesin olmamakla birlikte örneğin karar ağaçlarını kullanan tekniğin hem psikolojik araştırmalar tarafından ya da uzman sistemlerden bilgi kazanımı ile güdülenen otomatik öğrenme topluluğu içerisinde hem de doğrusal fonksiyonlara dayalı klasik ayrımcılık tekniklerinin algılanan sınırlamalarına tepki olarak istatistik mesleği içinde paralel geliştirilmiştir. Benzer şekilde istatistikte geliştirilen ileri regresyon teknikleri ve psikoloji, bilgisayar bilimleri ve yapay zekâ geçmişi olan yapay sinir ağı modelleri ile paralellikler görülebilir [15].

3.2. K- En Yakın Komşu (KNN)

K en yakın komşuluk algoritması sorgu vektörünün en yakın k komşuluktaki vektor ile sınıflandırılmasının bir sonucu olan denetlemeli oğrenme algoritmasıdır. Bu algoritma, sınıflandırma problemini çözen denetimlioğrenme algoritmalarından biridir.

Sınıflandırma, yeni bir nesnenin özniteliklerini inceleme ve bu nesneyi önceden tanımlanmış bir sınıfa atamaktır. Burada önemli olan, her bir sınıfın özelliklerinin önceden net bir şekilde belirlenmiş olmasıdır. K en yakın komsu yönteminde; sınıflandırma yapılacak verilerin öğrenmekümesindeki normal davranıs verilerine benzerlikleri hesaplanarak; en yakın olduğu düsünülen k verinin ortalamasıyla, belirlenen eşik değere göre sınıflara atamaları yapılır.



Şekil 3.1 - Knn yöntemi

3.2.1. KNN algoritmasının adımları

- Test kümesindeki her verinin X = {x1, x2, x3, x4, …., xn}, öğrenme kümesindeki verilere D = {d1, d2, d3, d4, …., dm} yakınlığı hesaplanır.

$$d(x_i, d_l) = \sqrt{\left(x_{i_1} - d_{i_1}\right)^2 + \left(x_{i_2} - d_{l_2}\right)^2 + \cdots + \left(x_{i_p} - d_{l_p}\right)^2}$$

- Her verinin öğrenme kümesindeki verilere olan yakınlıkları sıralanıp ilk “k” tanesi alınarak ortalamaları hesaplanır.

- Ortalama değerleri, belirlenen eşik değerinden büyük olanlar normal, küçük olanlar ise anormal olarak sınıflandırılır.

3.2.2. Algoritmanın performansını etkileyen kriterler

Algoritmanın performansını etkileyen bazı kriterler söz konusudur. Bunlar;

- k en yakın komşu sayısı (benzerlik ölçümü için seçilecek komsu sayısı: k),
- Eşik değer
- Benzerlik ölçümü
- Öğrenme kümesindeki normal davranışların yeterli sayıda olması (öğrenmekümesi yeterli çesitlilikte ve sayıda normal davranıs verisi içermiyorsa, test kümesinde yer alan yeni normal davranıs verileri anormal olarak algılanabilir.)

3.2.3. Algoritmanın avantajları

Algoritmanın esas olarak iki önemli avantajı vardır. Bunlar;

- Uygulanabilirliği basit bir algoritmadır.
- Uygulanacak örnek sayısı fazla ise etkilidir.

3.2.4. Algoritmanın dezavantajları

KNN algoritmanısın getirdiği bazı dezavantajlar vardır. Bu dezavantajlar şu şekide sıralanabilir;

- K parametreye ihtiyac duyar.
- Uzaklık bazlı öğrenme algoritması, en iyi sonuçları elde etmek icin, hangi uzaklık tipinin ve hangi niteliğin kullanılacağı konusunda açık değildir.
- Hesaplama maliyeti gerçekten cok yüksektir. Bazı indeksleme metodları ile (örneğin K-D ağacı), bu maliyet azaltılabilir.

3.3. Destek Vektör Makinesi – DVM (Support Vector Machine - SVM)

Destek vektör makineleri istatistiki öğrenme teorilerini temel alan, Vapnik – Chervonenkis teorisine dayanan, kuramsal temelleri güçlü ve oldukça güncel bir algoritmadır [17].

Destek vektör makineleri güçlü düzenlileştirme özelliklerine sahiptir. Bu modelin yeni veriye gore genellenmesi anlamını taşır. Destek vektör makineleri karar sınırlarını tanımlayan karar düzlemini temel alır. Karar düzlemi farklı sınıflara ait nesne kümelerini ayırır.

Destek vektör makineleri popular veri madenciliği tekniklerinden sinir ağları ve radial temelli fonksiyonlara benzer şekilde fonksiyonlar içerir. Ancak sözü edilen algoritmalarda iyi tasarlanmış algoritmalar olsa da destek vektör makinelerinin temelinde düzenlileştirme konusunda daha iyi kurgulanmış bir teorik yaklaşım vardır. Ayrıca genelleme kalitesi ve eğitim kolaylığı göz önüne alındığında destek vektör makineleri yukarıda sözü edilen geleneksel yöntemlerin çok ötesinde bir kapasiteye sahiptir.

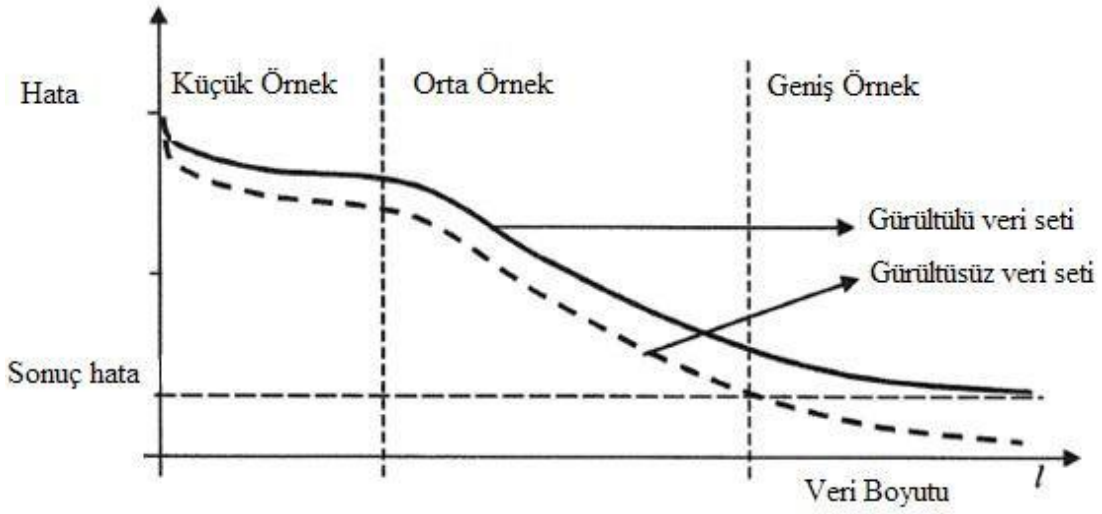
Destek vektör makineleri karmaşık metin ve görüntü sınıflandırma, el yazısı tanıma, yüz tanıma ve gen analizi gibi karmaşık problemleri modelleyebilir. Pek çok özniteliğe sahip veri kümelerinde modelin eğitimi için çok az durum kullanılmış olsa bile destek vektör makineleri verimli bir şekilde çalışır. Destek vektör makinelerinde kullanılacak örnek sayısı önemli bir kriter degildir. Destek vektör makineleri_eğitim esnasında görülmemis verileri de herhangi bir sorun teşkil etmeksizin sınıflandırır. Özniteliklerde sayıca bir üst sınır konulmasına gerek olmadığı halde donanımın dayattığı sınır belirleyicidir. Bu Destek vektör makinelerinin genellestirebilme yeteneğinin göstergesi olarak değerlendirilebilir. Genellestirebilme özelligi Destek vektör makinelerini diğer tekniklere gore daha iyi bir alternatif yapmaktadır [17].

Destek vektör makineleri öğrenme, basit fikirler üzerine kurulma ve pratik uygulamalarda yüksek performans göstermesi bakımından oldukça kullanıslıdır.

3.3.1. Destek vektör makinelerinde öğrenme

Destek vektör makineleri, istatistiksel tekniklerin olasılık dağılımının temel alındığı eğitme algoritması olarak bilinir. Birçok pratik durumda, istatistiksel tekniklerin temelini olusturan dağıtma yasalarının hakkında yeterli bilgi ve dağılım bulunmamaktadır. Bu islem gerçek dünya uygulamalarında ortak olan çok ciddi bir kısıtlamadır [18].

Sahip olduğumuz yüksek boyutlu olan desenleri eğitmek günümüze ait uygulamalarda güçlükle kaydedilir. Öğrenen makine algoritmaları yüksek boyutlu uzaylarda çalıstırabilmeli ve az sayıda veriden öğrenme islemini yapabilmelidir. Boyut indirgeme islemi sağlanırsa veri çiftleri kolay bir sekilde elde edileceginden daha iyi sonuçlar verilmis olur. Klasik istatistiksel tekniklerin temel performansı Sekil 3.2 'de verilmistir. Pratik koşullarda rasgele bir veri takımından alınan küçük örnek boyutu güvenilmezdir, genellikle hata ile sonuçlanır.

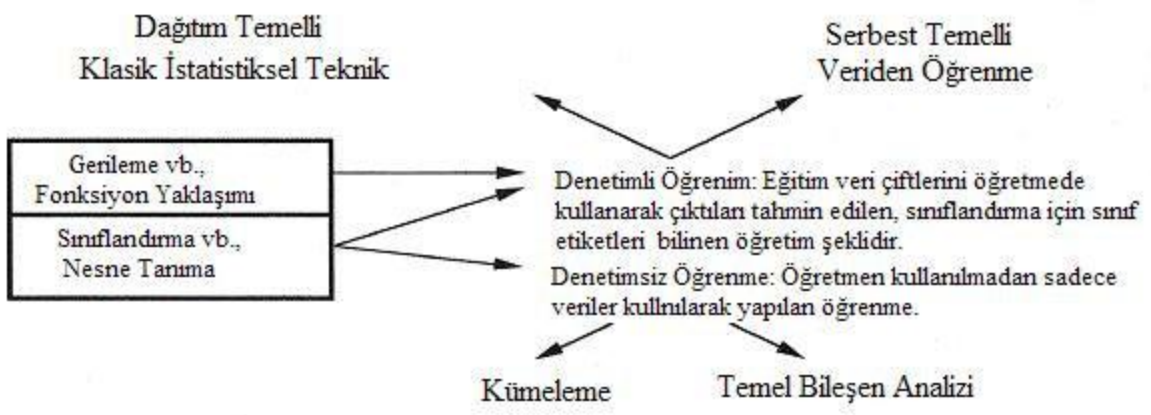


Şekil 3.2 - Klasik istatistiksel teknik performansı

DVM polinomial modellerde, yapay sinir aglarında, bulanık mantıkta ve RTF sınıflandırıcılarda sıkça kullanılan bir metottur. DVM'yi, yapısal risk minimizasyonu (YRM) olarak bulunan yeni ögrenim teknikleri ve VC teorisi temsil eder. Gelistirilen Vapnik teorisi düsük seviyeli model VC boyutu, görünmeyen veriler üzerinde iyi bir genelleme yaparak hatanın düsük olasılıkta olduğunu gösterir. Bu özellik, tüm hesaplayan alana özeldir [18].

Veri çiftlerini eğitmede o kadar iyi model olmasa da genelleme islemini iyi gerçekleştiren modeldir. Bazı kısıtlamalar altında DVM'nin usulüne uydurma teori yapısında, istatistiksel öğrenen teori veya yapısal riskten ziyade minimizasyon türetilebilmesi vardır. Buradan, DVM'nin istatistiksel öğrenen teori ve yapısal risk teorisinin minimizasyonunu çıkaran, yani, öğrenen bir teknik olduğu söylenebilir. Tümevarım prensibini ve VC sınırının teorisini temel alan verilerle bu yaklaşımlarda öğrenme islemi gerçeklesir. En basit desen tanıma görevlerinde, vektör makinelerinin azami kenarla bir sınıflandırıcıyı yaratması için doğrusal ayıran bir yüksek düzlemi kullanır. Bunu gerçeklemek için, öğrenen problem, doğrusal olmayan bir optimizasyon problemi olarak alınır. Verilmis sınıflar uzayının olduğu original giriste doğrusal olarak ayrılamadıgı zaman DVM önce doğrusal olarak, daha yüksek boyutlu bir özellik uzayını orijinal giriş uzayına dönüstürür [19]. Bu dönüsüm, çesitli doğrusal olmayan eşleştirmeleri kullanarak başarılabilir: Polinomial, çok katmalı algılayıcıda olduğu gibi sigmoidal, radyal olarak simetrik görevler Gaussian olduğu esas görevlere sahip olması için RTF eşleştirmeleri olabilir. Doğrusal olmayan dönüşüm adımı yapıldıktan sonra, doğrusal optimal ayrımı bulmak DVM'nin görevidir. Yani, optimizasyon problemini çözmesi, doğrusal ayrılabilir sınıflar için orijinal giris uzayında ayırıcı düzlem hesabı olarak aynı türden olur.

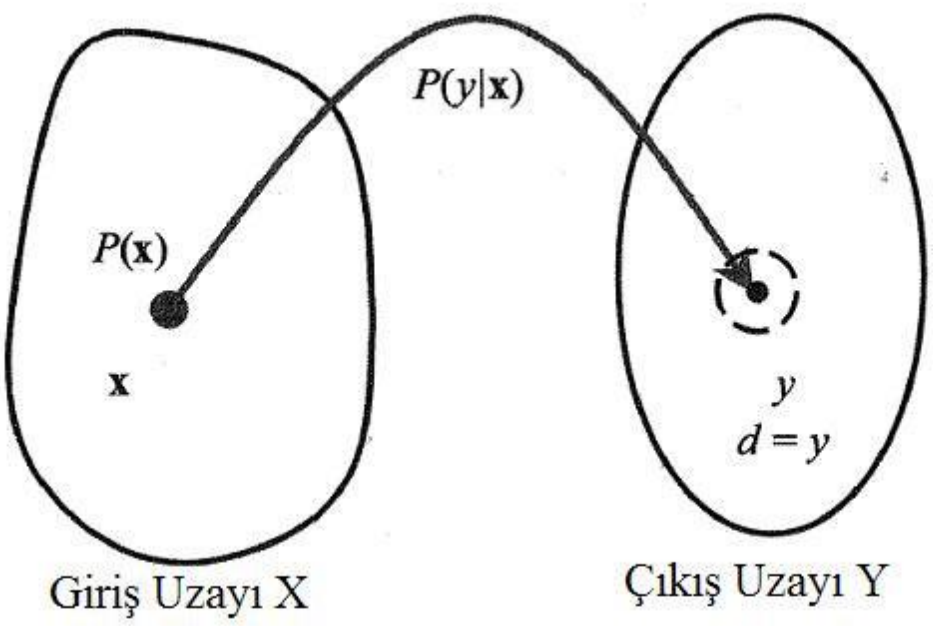
Özellik uzayında sonuç veren yüksek düzlem, azami bir kenar sınıflandırıcısı olduğunda optimal sonucu verir. Standart öğrenme durumu, Sekil 'de gösterilmektedir.



Şekil 3.3 - Rastgele eğitim

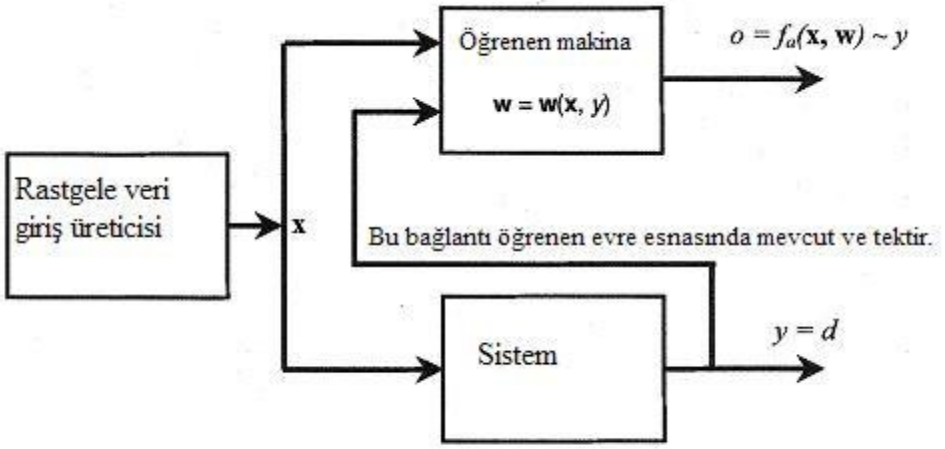
Öğrenme, stokastik bir süreçtir. Sekil 3.3 ’te rasgele eğitim gösterilmektedir. Eğitim takımı, rasgele değişken takımlarından olusturularak, giris değişkeni, rasgele değişken *xi* 'dir.

Giristen (*P xi)* olasılıkla çekilmis, çıktıya ait olan *yi* 'nin, olasılıgı P(*yi*| *xi)* ’dir. Bu özellik eğitim fazı sırasında *di* (istek) tarafından *yi* yanıtını gösterir. Bu yüzden P(di | xi) = P (yi | xi)olarak ifade edilebilir. Y burada, sadece basitlik için kullanılan çıktı değiskeninin yönsüz değeridir. Bütün kökler, temelde aynı vektör çıktısı *y*'den türer. Toplanan *(x,d)* veri noktalarının olasılıkları *P*(*x*, *d*) =*P*(*x*)*P*(*y* | *x*) şeklindedir [20].



Şekil 3.4 - Eğiten veri takımını toplarken öğrenmenin rastgele gösterimi

İstatistiksel ayarda, veri öğrenmede üç temel bilesen vardır. Bunlar; x rasgele girislerin bir üreteci, eğitim sisteminin yanıtları y ve x girisleri ile y sistem çıkıslarını kullanarak öğrenmeye çalısan makinedir. Sekil , çesitli alanlarda ortak olan öğrenimi gösterir. Özellikle, sistem teshisini kontrol eder ve isleme tabi tutmaya çalısır. Kullanılan X ve Y verisinin arasındaki iliskiyi geri dönüste D basarılı sekilde bulması için egiten evre esnasında öğrenen bir makine görevleri ya da bir fonksiyonu ile sınıflandırma görevlerinde veriyi ayırır. Öğrenme isleminin yaklasık fonksiyonel sonucu f (x,w) a 'dır. Bu fonksiyon, yaklasık (veya doğru) temeli olusturmayı, giris ve gerileme veya karar sınırında çıktının arasında bağımlı durumları tahmin eder [20].



Şekil 3.5 - Öğrenen makinenin modeli

Genel olarak "Fonksiyonu tahmin etmek" ismi altında ,girisleri ve buna karsı olusan çıktıları haritaya döken herhangi bir matematige ait yapı kullanılmıstır [21].

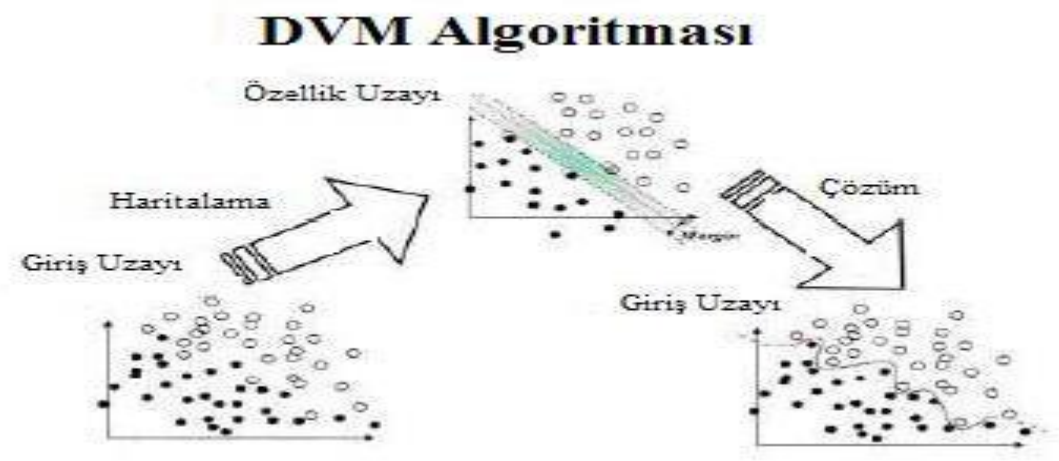
Böylece,"Fonksiyonu tahmin etmek", kabarık bir modelin olduğu bir çok katmanlı perceptron, yapay sinir agı, RTF şebekesi, bir DVM'yi ifade edebilir. Parametrelerin bir takımı *w*, öğrenmenin konusudur vegenellikle bu parametreler, agırlıklar ile çagırılır. Bu parametreler, geometrik fark veya fiziksel anlamlara sahip olur. Görevlerin hipotez uzayı üzerinde H'ye bağlı olan w'nin genellikle olduğu sistemler;

- Gizli ve çıkıs katmanının ağırlıkları çok katmanlı algılayıcılar
- Bulanık alt kümelerin şekil ve pozisyonlarındaki tanımlamalarda kurallar ve parametreler
- Bir polinomial veya Fourier dizisinin katsayıları
- RTF sebekesinin çıktı tabaka agırlıklarına ek olarak Gaussian esas fonksiyonlarının merkezleri ve varyansları veya kovaryansları.

Ana problem, çok küçük giris ve çıktı değişkenlerinin arasında mümkün temeli olusturan görevdir. Erisilebilen eğitim veri kümesi, bagımsız uzayın bazı bilinmeyen olasılık dağılımına göre etiketlenen örnekleri koyar. Takip eden kısımlar, temel fikirler ve küçük numunelerle öğrenmek için gelişmiş öğrenmenin ilk etraflı teorisi olan Vapnik ve Chervonenkis tarafından gelistirilen istatistiksel öğrenen teorinin tekniklerini sunar [20].

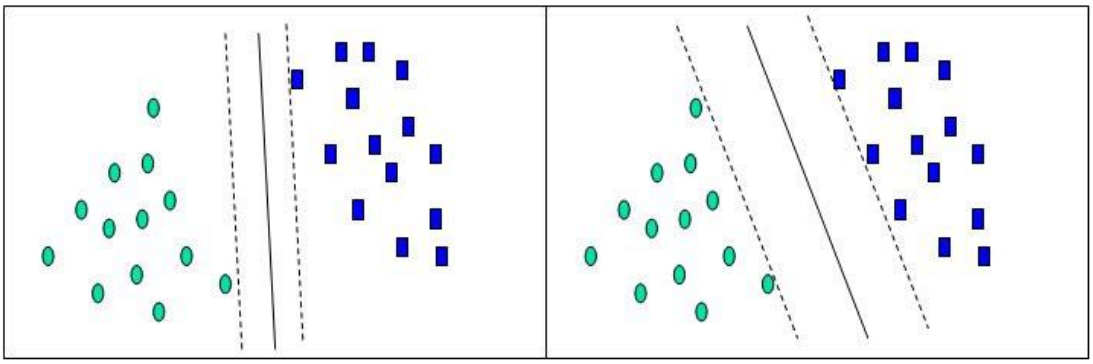
3.3.2. Destek vektör makinesi algoritması

DVM temelinde, öncelik değişkenini bir özellikle çagırıp, çok boyutlu düzlemde kullanılan niteliğe dönüştürür. En uygun temsili seçmenin amacı, özellik seçimi olarak bilinir. Özellikler dogru seçilirse iyi temsiller elde edilerek dogru sonuçlara ulasılabilir. Bir olayı tanımlayan özellik takımı bir vektör ile çagırılır. Bundan dolayı DVM modelinin amacı, hedef değişkeninin bir kategorisiyle olayların vektör kümelerini ayıran optimal askın düzlemibulmaktır. Asırı düzlemin yanındaki vektörler destek vektörleridir [21]. Aşağıdaki şekilde destek vektör algoritmasının genel yapısı görülmektedir.



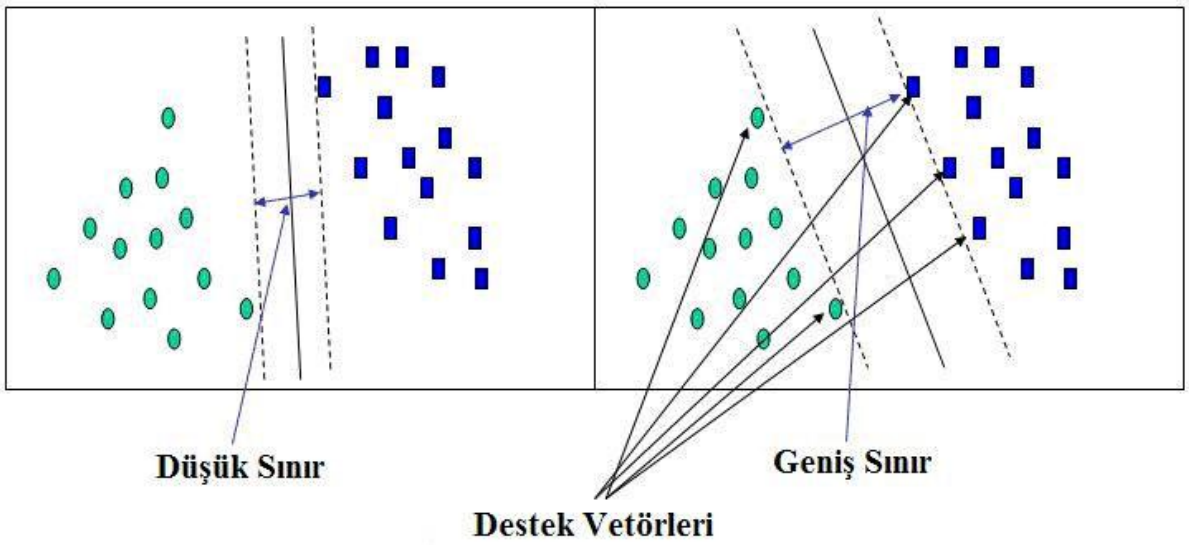
Şekil 3.6 - Dvm algoritma yapısı

N boyutlu aşırıdüzlemi düsünmeden önce basit 2 boyutlu bir örnek üzerindealgoritmanın çalısmasını inceleyelim. Sınıflandırma için 2 kategorili hedef değişkeni ele alalım. Devamlı değerlerle iki öncelik değişkeni olduğunu varsayalım. X ekseninde öncelik değişkenlerinden birini Y ekseninde diğerini kullanarak veri noktalarını oluşturursak asağıdaki şekli elde ederiz. Hedef değişkeninin bir kategorisini dikdörtgenler ile, diğer kategorileri ovallar ile temsil ederiz. Sekil 3.7' de sınıflandırma örnegi görülmektedir [20].



Şekil 3.7 - Sınıflandırma örneği

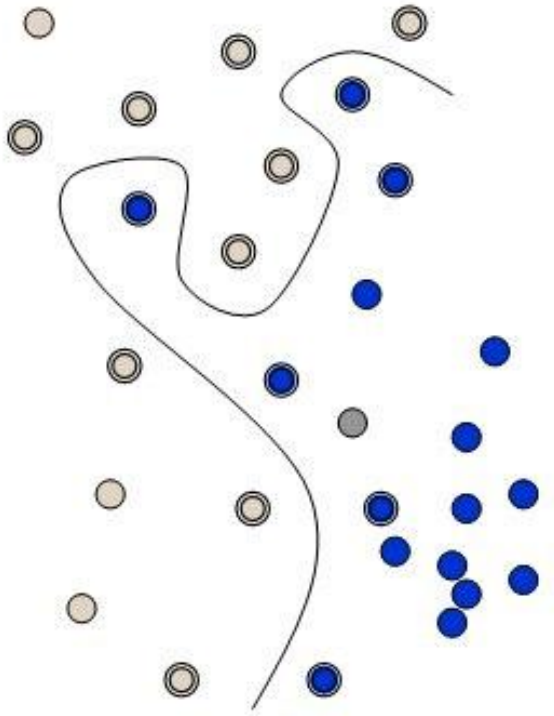
Bu örnekte durumlar tamamen farklı köselerde toplanıp tamamen ayrıstırılmıstır. DVM analizi, olayları ayıran 1-boyutlu askın düzlemi hedef kategorilerini temel alarak bulmaya çalısır. Mümkün çizgilerin sınırsız sayısı vardır. İki aday çizgi yukarıdaki gibi gösterilir. Hangi çizginin, daha iyi olduğunu ve optimal çizgiyi nasıl bulacagımız önemlidir. Noktalı gösterilen çizgiler en yakın vektörler arasında mesafeyi ayıran çizgiye paralel olarak çekilir. Noktalı çizgilerin arasındaki mesafe kenarı çağırır. Kenarın genişliğini zorlayan vektörler, destek vektörleridir. Şekil 3.8' de destek vektörleri ve elde edilen maksimum sınırlar gösterilmistir [21].



Şekil 3.8 - Sınıflandırma örneği (devam)

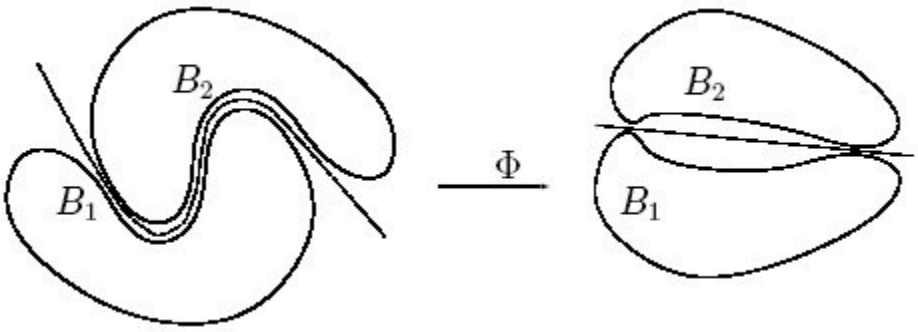
DVM analizi, destek vektörlerinin arasındaki kenarı azami dereceye çıkararak yön verilen çizgileri bulur. Eger bütün analizler, iki öncelik değişkeniyle iki kategori hedef değişkenine dayansaydı noktalar kümesini düz bir çizgi ile bölmüs olurdu. Yüksek boyutlara geçersek; Yukarıdaki örnekte 2 boyut için iki öncelik değişkeni kullanılıp, düzlem üzerinde çözüm bulunmustur. Bu islem 1 boyut için çizgi seklindedir. 3 boyutlu bir örnek ele alınırsa 3. öncelik değişkeni devreye girerek bir küp elde edilmis olur.

Ekstra öncelik değişkenleri eklenildigi gibi, veri noktaları N boyutlu uzayda temsil edilebilir ve (N-1) ayırıcı düzlem, onları ayırabilir. İki grubu bölmek için en basit yol, düz bir çizgi veya N-boyutlu bir düzlemdir. Ama noktalar dogrusal çizgiyle ayrılamayacak sekilde bulunursa bu islem yapılamaz. Şekil 3.9' da doğrusal olmayan sınıflandırma örneği görülmektedir [20].



Şekil 3.9 - Doğrusal olmayan sınıflandırma

Bu durumda doğrusal olmayan bir çizgiye ihtiyaç duyulur [24]. Veriye dogrusal olmayan eğrilerle uymaktansa DVM'yi baska bir uzaya çekirdek fonksiyonu aracılığıyla taşıyarak daha tutarlı bir sekilde ayrım sağlanmıs olunur. Sekil 3.10' da çekirdek fonksiyonlarının üst boyuta taşınması gösterilmektedir.



Şekil 3.10 - Çekirdek fonksiyonları ile üst boyuta taşıma işlemi

4. TAVSİYE SİSTEMLERİNDE VERİ BÜTÜNLEŞTİRME

Bu kısımda yapılan çalışmanın detayları yer almaktadır. Veri setlerinin bütünletirilerek hazırlanma aşaması ve bu hazırlanan veri setlerinin uygulamalara girdi olarak verilme aşamalarından bahsedilecektir.

4.1. Veri Setinin Hazırlanması ve Bütünleştirilmesi

Bu çalışmada kullanılan veri setleri esas olarak bir film tavsiye sistemi olan MovieLens sistemine aittir [25]. Fakat bu verilerin işlenmiş ve düzenlenmiş hali "Content-Boosted Collaborative Filtering using Semantic Similarity Measure" adlı çalışmadan alınmıştır [26].

Kullanılan veri seti;

- 1682 film
- 943 kullanıcı üzerinden
- 100.000 kullanıcı oyu,

şeklindedir.

Bu veri setleri beş set şeklinde ayrılmıştır. Alınan verilerde her kullanıcının bütün filmleri izlemiş olup puanlandırması söz konusu değildir. Uğur Ceylan tarafından yapılan çalışmanın amacı tavsiye sistemleri içerisinde, film nesneleri arasında içerik bazlı anlamsal benzerlikleri bularak işbirlikçi filtreleme yöntemiyle birleştirmektir. Çalışmanın büyük bir kısmı kullanıcı puan - film matrisi üzerindeki verilerin eksik noktalarını otomatik olarak tahmin etme ve doldurma üzerine kurulmuştur.

Öğeler arasındaki benzerlikler kıyaslanırken şu üç farklı ölçüt üzerinden gitmiştir. Bunlar;

- Taksinomi benzerliği
- İlişki benzerliği
- Özellik benzerliği şeklindedir.

Öğeler arası bu benzerlikler ve kullanıcı puan vektörlerinin bulunduğu aktif kullanıcı modeli kullanılarak vektörlerde boş olan noktaları, diğer bir anlatımla kullanıcının henüz puanlamamış olduğu filmleri, tahmin etme çalışması yapılmıştır. Kullanıcılar arasında benzerlikler de göz önünde bulundurulmuştur. Bu benzerlikler bulunurken işbirlikçi filtreleme yöntemi kullanılmıştır. Bu yöntem kullanılırken de benzerlik hesaplamaları k-en yakın komşu algoritması üzerinden yürütülmüştür [26][27].

Uğur Ceylan tarafından yapılan çalışma sayesinde, bu tez çalışmasında kullanılacak olan verilerin eksiksiz olarak tanımlanması sağlanmıştır. Bu şekilde kullanıcıların her bir filmi izlediği ve bu filme puanlama yaptığı varsayılmıştır. Her bir veri seti %80' e %20 oranında parçalanmış, %80' lik kısmı sınıflandırma algoritmalarının eğitilme aşamasında kullanılmıştır. Geriye kalan %20' lik kısmı algoritmalara test verisi olarak dahil edilmiştir. Elde edilen veri setleri şu şekilde oluşmuştur :

u1	i1	i2	i3	i4	i5	i6	i7	i8	…	i1682	sınıf
u2	i1	i2	i3	i4	i5	i6	i7	i8	…	i1682	sınıf
u3	i1	i2	i3	i4	i5	i6	i7	i8	…	i1682	sınıf

…

…

Veri bütünleştirme işleminden önce elde edilen veri seti her kullanıcı için o kullanıcıya ait filmlere verilen puanlar şeklindedir. Veriler bu şekilde sınıflandırma algoritmasına

girdi olarak verilseydi, her kullanıcı bir vektör şeklinde olacak ve bu vektörün öznitelikleri her bir film olacaktı. Her bir özniteliğe ait değerler de o kullanıcının verdiği puanlar olacaktı.

Bu noktada veri bütünleştirme işlemi devreye girmektedir. Veri bütünleştirme işleminin amacı sınıflandırma algoritmalarına sokulan her bir vektörün öznitelik sayısını arttırmak ve bu şekilde sınıflandırma algoritmalarında daha iyi bir sonuç elde edilip edilemeyeceğini değerlendirmektir.

Veri bütünleştirme işlemi adım adım düşünüldüğünde basit olarak aşağıdaki şekilde çalışmaktadır;

1. Bir kullanıcının filmlere verdiği puanlama vektörü ele alınır
2. Bir filme bütün kullanıcılar tarafından verilen puanlama vektörü ele alınır
3. Birinci aşamada elde edilen vektörün sonuna ikinci adımda elde edilen vektör eklenir.

Tablo 4.1 Veri bütünleştirme

	i1	i2	i3	i4	i5	i6	…
u1	v1.1	v1.2	v1.3	v1.4	v1.5	v1.6	…
u2	v2.1	v2.2	v2.3	v2.4	v2.5	v2.6	…
u3	v3.1	v3.2	v3.3	v3.4	v3.5	v3.6	…
u4	v4.1	…	…	…	…	…	…
u5	v5.1	…	…	…	…	…	…
u6	v6.1	…	…	…	…	…	…
u7	v7.1	…	…	…	…	…	…
u8	v8.1	…	…	…	…	…	…
u9	v9.1	…	…	…	…	…	…
…	…	…	…	…	…	…	…

Veri bütünleştirmenin sonunda elde edilen veri setinde her bir vektör aşağıdaki gibi bir şekil almıştır:

i1 i2 i3 i4 i5 i6 i7 i8 i9 i10 i11 i12 i13

i14 … i1674 i1675 i1676 i1677 i1678 i1679 i1681 i1682 u1 u2 u3

u4 u5 u6 u7 … u935 u936 u937 u938 u939 u940 u941 u942

u943 sınıf

Test ve eğitilme kümesinin veri bütünleştirme işleminden geçirilmesinden sonra eğitilme kümesinde verinin çok büyük olmasından dolayı ve eğitilme işleminin kullanılan sınıflandırma algoritmalarında uzun sürmesinden dolayı, bu küme 40.000 vektör ile sınırlandırılmıştır.

4.2. Destek Vektör Makinesi (DVM)

Bütünleştirilmiş veri setleri öncelikle destek vektör makinesine sokulmuştur. DVM algoritması LIBSVM adında bir .NET kütüphanesi kullanılarak hazır halde ele alınmıştır [28]. Veri kümeleri bu kütüphane için tekrar düzenlenmiştir. Vektör şu şekile getirilmiştir:

sınıf 1:değer 2:değer 3:değer 4:değer 5:değer

6:değer 7:değer … 2625:değer

Veri kümeleri yukarıda ki gibi hazırlandıktan sonra algoritmaya doğrudan sokulmuştur.

4.3. K – En Yakın Komşu (KNN)

KNN algoritması hazırlnadıktan sonra eğitilme ve test kümesi üzerinde değişiklik yapılmadan doğrudan algoritmaya sokulmuştur. Fakat işlemin donanımsal olarak sorunlar çıkarması ve oldukça uzun sürmesi sonucu yazılımda değişikliğe gidilmiştir.

KNN' in yapısının uygun olması sonucunda çok izlekli (multi-thread) bir yapı kullanılmıştır. 4 gerçek 4 sanal olmak üzre toplamda 8 çekirdekli bir makinede 7 izlek kullanılmıştır. Bu sebeple test verisi yediye bölünüp 7 ayrı izlek üzerinde çalıştırılmıştır.

5. DEĞERLENDİRME

Bu bölümde kullanılan deneysel veri setleri, çalışma sonunda çıkan sonuçlar, bu sonuçların değerlendirilmesinde kullanılan yöntemler ve bu yöntemler kullanılarak değerlendirilme aşamaları yer almaktadır.

5.1. Veri Kümeleri

Bu çalışmada kullanılan veri setleri esas olarak bir film tavsiye sistemi olan MovieLens sistemine aittir [25]. Fakat bu verilerin işlenmiş ve düzenlenmiş hali "Content-Boosted Collaborative Filtering using Semantic Similarity Measure" adlı çalışmadan alınmıştır [26].

Veri kümeleri, "Veri Setinin Hazırlanması ve Bütünleştirilmesi" kısmında bahsi geçen ön çalışmalardan geçirilip veriler bütünleşik hale getirildikten sonra sınıflandırma algoritmalarına girdi olarak verilmiştir. Veri kümelerinin son hali aşağıda belirtilmiştir.

Eğitici Veri Kümesi : 40.000 vektörden oluşmaktadır. Her bir vektör ;

Sınıf + Bir kullanıcının filmlere verdiği puan + Bir filme kullanıcıların verdiği puan

Test Veri Kümesi : ≅20.000 vektörden oluşmaktadır. Her bir vektör ;

$$S\square n\square f + Bir\ kullan\square c\square n\square n\ filmlere\ verdi\breve{g}i\ puan + Bir\ filme\ kullan\square c\square lar\square n\ verdi\breve{g}i\ puan$$

5.2. Değerlendirme Yöntemleri

Bir tavsiye sistemi değerlendirilirken bir çok yaklaşım kullanılabilmektedir. Bu yaklaşımlardan en temel olanları “zaman / hafıza verimliliği” ve “sonuçların veya kullanıcıların zevklerinin etkinliği” olarak gösterilmektedir [29]**.** Bu tezde ağırlıklı olarak sonuçların etkinliği üzerinde durulmasına rağmen zaman ve kullanılan hafıza ile ilgili de bilgi verilmektedir.

Sonuçların etkinliğinin ölçülmesi için temel olarak "Duyarlılık" (Precision) ve "Anımsama" (Recall) oranları kullanılır. Son adım olarak bu oranlar kullanılarak “F – ölçütü” (F – Measure , F – Score) denilen ölçü birimi hesaplanır.

Duyarlılık ve anımsama oranlarını bulmak için test verisinden elde edilen sınıflandırma sonuçlarıyla, bu verilerin gerçek sınıfları arasındaki doğruluk sayılarının bulunduğu bir tablo kullanılır. Bu tablo “Hata Matrisi” (Confusion Matrix – Table of Confusion) olarak adlandırılmaktadır.

5.2.1. Hata matrisi (confusion matrix – table of confusion)

Tanım olarak bakıldığında hata matrisi iki sınıflandırma arasındaki tutarlılığı ölçmek için kullanılan tablodur [30]. Hata matrisi temel olarak dört alandan oluşur :

- Doğru – Pozitif (TP)
- Doğru – Negatif (TN)
- Yanlış – Pozitif (FP)
- Yanlış – Negatif (FN)

Pozitif ve negatif terimleri sınıflandırma aracının tahmin ettiği değerleri gösterir. Doğru ve yanlış terimleri de tahmin edilen sınıfın doğruluğunu belirtmek için kullanılmaktadır.

TP : Pozitif tahmin edilen sınıfların gerçek sınıfının da pozitif çıkması durumunda miktarının yazılacağı alandır.

TN : Negatif tahmin edilen sınıfların gerçek sınıfının da negatif çıkması durumunda miktarının yazılacağı alandır.

FP : Pozitif tahmin edilen sınıfların gerçek sınıfının, negatif çıkması durumunda miktarının yazılacağı alandır.

FN : Negatif tahmin edilen sınıfların gerçek sınıfının, pozitif çıkması durumunda miktarının yazılacağı alandır.

Yapı olarak hata matrisi şu şekildedir :

Tablo 5.1 - Hata Matrisi

		Tahmin Sınıfı	Tahmin Sınıfı
		Negatif (-)	**Pozitif (+)**
Gerçek Sınıfı	**Negatif (-)**	(TN)	(FP)
Gerçek Sınıfı	**Pozitif (+)**	(FN)	(TP)

5.2.2. Duyarlılık (precision)

Duyarlılık oranı, sınıflandırma algoritmasının bir sınıf için getirdiği doğru sonuçların doğru olarak tahmin ettiği tüm sonuçlara oranı olarak ele alınmaktadır.

Bir örnek üzerinden gidilirse, herhangi bir arama motorunun bir sorgu karşısında, geri getirdiği ilgili belge sayısının tüm belgelere oranıdır.Örneğin arama motoru sorgu karşısında yüz sayfa getirdiyse ve bu sayfaların 85 tanesi konu ile ilgiliyse bu arama motorunun duyarlılık oranı %85' tir [31].

Duyarlılık oranının hata matrisi üzerinden gidildiğinde şu formul ile hesaplanır [30]:

$$Duyarl\square l\square k = \frac{TP}{TP + FP}$$

5.2.3. Anımsama (recall)

Anımsama oranı , sınıflandırma algoritmasının bir sınıf için getirdiği doğru sonuçların doğru ve yanlış olarak tahmin ettiği bütün sonuçlara oranı olarak ele alınmaktadır.

Anımsama oranı bir örnek üzerinden anlatılmak istenirse, tarama yapılan bir veri tabanında tüm ilgili verilerden ne kadarına erişilebildiğini gösterir [33].

Anımsama oranının hata matrisi üzerinden gidildiğinde şu formul ile hesaplanır [32]:

$$An\square msama = \frac{TP}{TP + FN}$$

5.2.4. F – ölçütü (f - measure)

F – ölçütü, verilerin, duyarlılık ve anımsama değerlerinin harmonik ortalamasıdır. Sınıflandırma sonuçlarının değerlendirilmesinde esas alınan birim f – ölçütü olacaktır. Duyarlılık ve anımsama sonuçlarını birlikte ele alması bu ölçütü güçlü hale getirmektedir ve doğruluk derecesi olarak daha anlamlı bir sonuç vermektedir. F – ölçütü şu şekilde hesaplanır :

$$F - \ddot{o}l\ddot{c}\ddot{u}t\ddot{u} = 2 \times \frac{Duyarl\square l\square k \times An\square msama}{Duyarl\square l\square k + An\square msama}$$

F – ölçütü sonuç olarak 1 ile 0 arasında bir sonuç verir. Sonuç ise şu şekilde değerlendirilir :

- **1** en **iyi** sonuç
- **0** en **kötü** sonuç

5.3. Parametreler

Bu çalışmada sınıflandırma işlemi yapılırken Destek Vektör Makinesi (DVM) ve K- En Yakın K – En Yakın Komşu (KNN) algoritmaları kullanılmıştır. Bu algoritmalar kullanılırken başlangıçta girdi olarak parametreler almaktadır.

KNN algoritması için bir “k” değeri belirlenmelidir. Bu değerin belirlenmesinde kesin yöntemler yoktur. Bu nedenle, bu çalışma esnasında “k” parametresi için birden fazla değer için ataması yapılıp tekrar çalıştırılmıştır.

Tablo 5.2 - Knn parametreleri

Parametre	**Değer**
k	5
k	7
k	9

DVM için kullanılan LIBSVM kütüphanesinde varsayılan parametreler üzerinden gidilmiştir. Bu parametreler aşağıdaki Tablo 5.3’ te belirtilmiştir [28].

Tablo 5.3 - Dvm parametreleri

Parametre Adı	Değer
svmType	C_SVC
kernelType	RBF
degree	3
gamma	0
coef0	0
Nu	0.5
cacheSize	40
C	1
Eps	1.00E-03
P	0.1
shrinking	TRUE
probability	FALSE
weights	<int,double>

5.4. Sonuçlar ve Değerlendirilmesi

Bu çalışmada çıkan sonuçların değerlendirilmesinde “Değerlendirme Yöntemleri” başlığı altında bahsi geçen “hata matrisi” üzerinden “anımsama” , “duyarlılık” , “f - ölçütü” kullanılmıştır.

5.4.1. KNN için sonuçlar

Parametre k = 5 değeri için :

Tablo 5.4 - Test 1 veri kümesi için knn hata matrisi (k=5)

		Tahmin Sınıfı	**Tahmin Sınıfı**
		Negatif (-)	**Pozitif (+)**
Gerçek Sınıfı	**Negatif (-)**	(TN)-5269	(FP)-3481
Gerçek Sınıfı	**Pozitif (+)**	(FN)-2758	(TP)- 8456

Tablo 5.5 - Test 1 veri kümesi için knn sonuçları (k=5)

Duyarlılık	0,7083
Anımsama	0,7540
F - ölçütü	0,7305

Tablo 5.6 - Test 2 veri kümesi için knn hata matrisi (k=5)

		Tahmin Sınıfı	**Tahmin Sınıfı**
		Negatif (-)	**Pozitif (+)**
Gerçek Sınıfı	**Negatif (-)**	(TN)-5183	(FP)-3575
Gerçek Sınıfı	**Pozitif (+)**	(FN)-2852	(TP)-8344

Tablo 5.7 - Test 2 veri kümesi için knn sonuçları (k=5)

Duyarlılık	0,7000
Anımsama	0,7452
F - ölçütü	0,7219

Tablo 5.8 - Test 3 veri kümesi için knn hata matrisi (k=5)

		Tahmin Sınıfı	**Tahmin Sınıfı**
		Negatif (-)	**Pozitif (+)**
Gerçek Sınıfı	**Negatif (-)**	(TN)-5539	(FP)-3423
Gerçek Sınıfı	**Pozitif (+)**	(FN)-3006	(TP)-7962

Tablo 5.9 - Test 3 veri kümesi için knn sonuçları (k=5)

Duyarlılık	0,6993
Anımsama	0,7259
F - ölçütü	0,7123

Tablo 5.10 - Test 4 veri kümesi için knn hata matrisi (k=5)

		Tahmin Sınıfı	**Tahmin Sınıfı**
		Negatif (-)	**Pozitif (+)**
Gerçek Sınıfı	**Negatif (-)**	(TN)-5640	(FP)-3413
Gerçek Sınıfı	**Pozitif (+)**	(FN)-2931	(TP)-7941

Tablo 5.11 - Test 4 veri kümesi için knn sonuçları (k=5)

Duyarlılık	0,6994
Anımsama	0,7304
F - ölçütü	0,7145

Tablo 5.12 - Test 5 veri kümesi için knn hata matrisi (k=5)

		Tahmin Sınıfı	**Tahmin Sınıfı**
		Negatif (-)	**Pozitif (+)**
Gerçek Sınıfı	**Negatif (-)**	(TN)-5487	(FP)-3480
Gerçek Sınıfı	**Pozitif (+)**	(FN)-2944	(TP)-7994

Tablo 5.13 - Test 5 veri kümesi için knn sonuçları (k=5)

Duyarlılık	0,6967
Anımsama	0,7308
F - ölçütü	0,7133

Parametre k = 7 değeri için :

Tablo 5.14 - Test 1 veri kümesi için knn hata matrisi (k=7)

		Tahmin Sınıfı	**Tahmin Sınıfı**
		Negatif (-)	**Pozitif (+)**
Gerçek Sınıfı	**Negatif (-)**	(TN)- 5229	(FP)- 3521
Gerçek Sınıfı	**Pozitif (+)**	(FN)- 2616	(TP)- 8598

Tablo 5.15 - Test 1 veri kümesi için knn sonuçları (k=7)

Duyarlılık	0,7094
Anımsama	0,7667
F - ölçütü	0,7369

Tablo 5.16 - Test 2 veri kümesi için knn hata matrisi (k=7)

		Tahmin Sınıfı	**Tahmin Sınıfı**
		Negatif (-)	**Pozitif (+)**
Gerçek Sınıfı	**Negatif (-)**	(TN)- 5199	(FP)- 3559
Gerçek Sınıfı	**Pozitif (+)**	(FN)- 2710	(TP)- 8486

Tablo 5.17 - Test 2 veri kümesi için knn sonuçları (k=7)

Duyarlılık	0,7045
Anımsama	0,7579
F - ölçütü	0,7302

Tablo 5.18 - Test 3 veri kümesi için knn hata matrisi (k=7)

		Tahmin Sınıfı	**Tahmin Sınıfı**
		Negatif (-)	**Pozitif (+)**
Gerçek Sınıfı	**Negatif (-)**	(TN)- 5546	(FP)- 3416
Gerçek Sınıfı	**Pozitif (+)**	(FN)- 2920	(TP)- 8048

Tablo 5.19 - Test 3 veri kümesi için knn sonuçları (k=7)

Duyarlılık	0,7020
Anımsama	0,7337
F - ölçütü	0,7175

Tablo 5.20 - Test 4 veri kümesi için knn hata matrisi (k=7)

		Tahmin Sınıfı	**Tahmin Sınıfı**
		Negatif (-)	**Pozitif (+)**
Gerçek Sınıfı	**Negatif (-)**	(TN)- 5588	(FP)- 3465
Gerçek Sınıfı	**Pozitif (+)**	(FN)- 2777	(TP)- 8095

Tablo 5.21 - Test 4 veri kümesi için knn sonuçları (k=7)

Duyarlılık	0,7002
Anımsama	0,7445
F - ölçütü	0,7217

Tablo 5.22 - Test 5 veri kümesi için knn hata matrisi (k=7)

		Tahmin Sınıfı	**Tahmin Sınıfı**
		Negatif (-)	**Pozitif (+)**
Gerçek Sınıfı	**Negatif (-)**	(TN)- 5431	(FP)- 3536
Gerçek Sınıfı	**Pozitif (+)**	(FN)- 2772	(TP)- 8166

Tablo 5.23 - Test 5 veri kümesi için knn sonuçları (k=7)

Duyarlılık	0,6978
Anımsama	0,7465
F - ölçütü	0,7213

Parametre k = 9 değeri için :

Tablo 5.24 - Test 1 veri kümesi için knn hata matrisi (k=9)

		Tahmin Sınıfı	**Tahmin Sınıfı**
		Negatif (-)	**Pozitif (+)**
Gerçek Sınıfı	**Negatif (-)**	(TN)- 5206	(FP)- 3544
Gerçek Sınıfı	**Pozitif (+)**	(FN)- 2479	(TP)- 8735

Tablo 5.25 - Test 1 veri kümesi için knn sonuçları (k=9)

Duyarlılık	0,7113
Anımsama	0,7789
F - ölçütü	0,7436

Tablo 5.26 - Test 2 veri kümesi için knn hata matrisi (k=9)

		Tahmin Sınıfı	**Tahmin Sınıfı**
		Negatif (-)	**Pozitif (+)**
Gerçek Sınıfı	**Negatif (-)**	(TN)- 5246	(FP)- 3512
Gerçek Sınıfı	**Pozitif (+)**	(FN)- 2620	(TP)- 8576

Tablo 5.27 - Test 2 veri kümesi için knn sonuçları (k=9)

Duyarlılık	0,7094
Anımsama	0,7659
F - ölçütü	0,7366

Tablo 5.28 - Test 3 veri kümesi için knn hata matrisi (k=9)

		Tahmin Sınıfı	**Tahmin Sınıfı**
		Negatif (-)	**Pozitif (+)**
Gerçek Sınıfı	**Negatif (-)**	(TN)- 5543	(FP)- 3419
Gerçek Sınıfı	**Pozitif (+)**	(FN)- 2787	(TP)- 8181

Tablo 5.29 - Test 3 veri kümesi için knn sonuçları (k=9)

Duyarlılık	0,7052
Anımsama	0,7458
F - ölçütü	0,7250

Tablo 5.30 - Test 4 veri kümesi için knn hata matrisi (k=9)

		Tahmin Sınıfı	**Tahmin Sınıfı**
		Negatif (-)	**Pozitif (+)**
Gerçek Sınıfı	**Negatif (-)**	(TN)- 5570	(FP)- 3483
Gerçek Sınıfı	**Pozitif (+)**	(FN)- 2690	(TP)- 8182

Tablo 5.31 - Test 4 veri kümesi için knn sonuçları (k=9)

Duyarlılık	0,7014
Anımsama	0,7525
F - ölçütü	0,7260

Tablo 5.32 - Test 5 veri kümesi için knn hata matrisi (k=9)

		Tahmin Sınıfı	**Tahmin Sınıfı**
		Negatif (-)	**Pozitif (+)**
Gerçek Sınıfı	**Negatif (-)**	(TN)- 5419	(FP)- 3548
Gerçek Sınıfı	**Pozitif (+)**	(FN)- 2649	(TP)- 8289

Tablo 5.33 - Test 5 veri kümesi için knn sonuçları (k=9)

Duyarlılık	0,7002
Anımsama	0,7578
F - ölçütü	0,7279

5.4.2. DVM için sonuçlar

Tablo 5.34 - Test 1 veri kümesi için dvm hata matrisi

		Tahmin Sınıfı	**Tahmin Sınıfı**
		Negatif (-)	**Pozitif (+)**
Gerçek Sınıfı	**Negatif (-)**	(TN)-5139	(FP)-3611
Gerçek Sınıfı	**Pozitif (+)**	(FN)-1973	(TP)-9241

Tablo 5.35 - Test 1 veri kümesi için dvm sonuçları

Duyarlılık	0,7190
Anımsama	0,8240
F - ölçütü	0,7679

Tablo 5.36 - Test 2 veri kümesi için dvm hata matrisi

		Tahmin Sınıfı	**Tahmin Sınıfı**
		Negatif (-)	**Pozitif (+)**
Gerçek Sınıfı	**Negatif (-)**	(TN)-5050	(FP)-3708
Gerçek Sınıfı	**Pozitif (+)**	(FN)-1932	(TP)-9264

Tablo 5.37- Test 2 veri kümesi için dvm sonuçları

Duyarlılık	0,7141
Anımsama	0,8274
F - ölçütü	0,7666

Tablo 5.38 - Test 3 veri kümesi için dvm hata matrisi

		Tahmin Sınıfı	**Tahmin Sınıfı**
		Negatif (-)	**Pozitif (+)**
Gerçek Sınıfı	**Negatif (-)**	(TN)-5317	(FP)-3645
Gerçek Sınıfı	**Pozitif (+)**	(FN)-2134	(TP)-8834

Tablo 5.39 - Test 3 veri kümesi için dvm sonuçları

Duyarlılık	0,7079
Anımsama	0,8054
F - ölçütü	0,7535

Tablo 5.40 - Test 4 veri kümesi için dvm hata matrisi

		Tahmin Sınıfı	**Tahmin Sınıfı**
		Negatif (-)	**Pozitif (+)**
Gerçek Sınıfı	**Negatif (-)**	(TN)-5436	(FP)-3617
Gerçek Sınıfı	**Pozitif (+)**	(FN)-2164	(TP)-8708

Tablo 5.41 - Test 4 veri kümesi için dvm sonuçları

Duyarlılık	0,7065
Anımsama	0,8009
F - ölçütü	0,7507

Tablo 5.42 - Test 5 veri kümesi için dvm hata matrisi

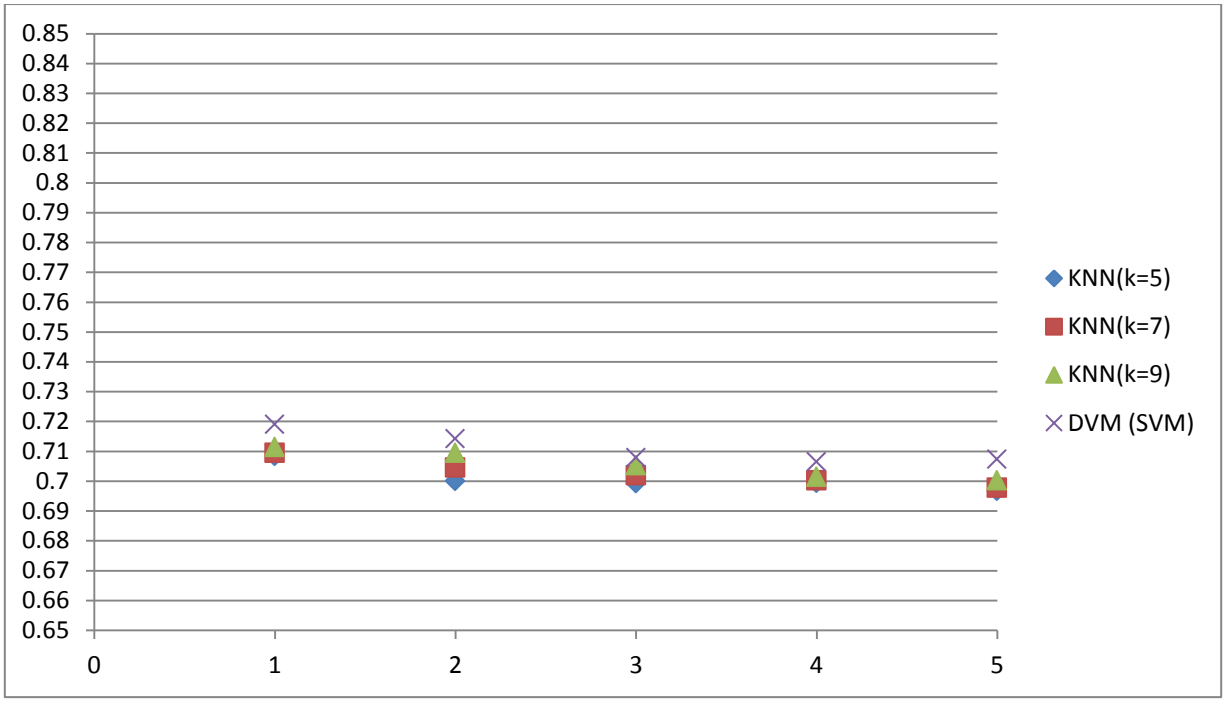
		Tahmin Sınıfı	**Tahmin Sınıfı**
		Negatif (-)	**Pozitif (+)**
Gerçek Sınıfı	**Negatif (-)**	(TN)-5278	(FP)-3689
Gerçek Sınıfı	**Pozitif (+)**	(FN)-2024	(TP)-8914

Tablo 5.43 - Test 5 veri kümesi için dvm sonuçları

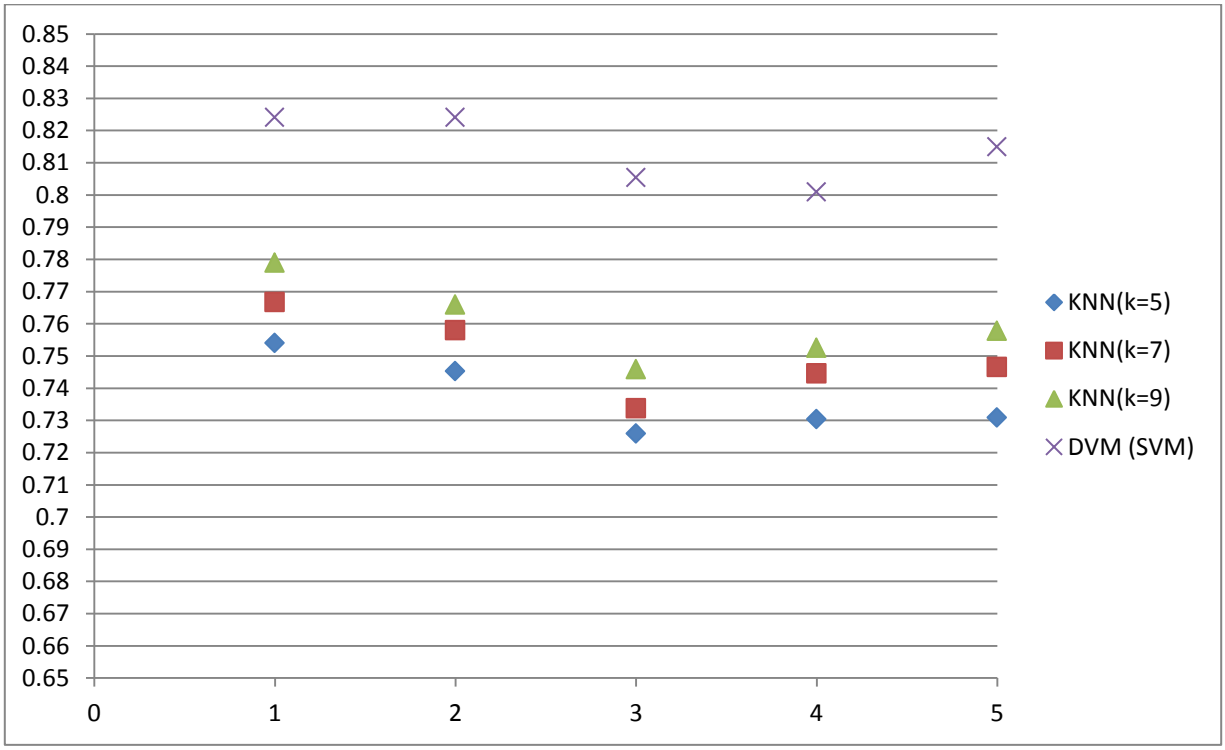
Duyarlılık	0,7072
Anımsama	0,8149
F - ölçütü	0,7573

5.4.3. Sonuç karşılaştırmaları

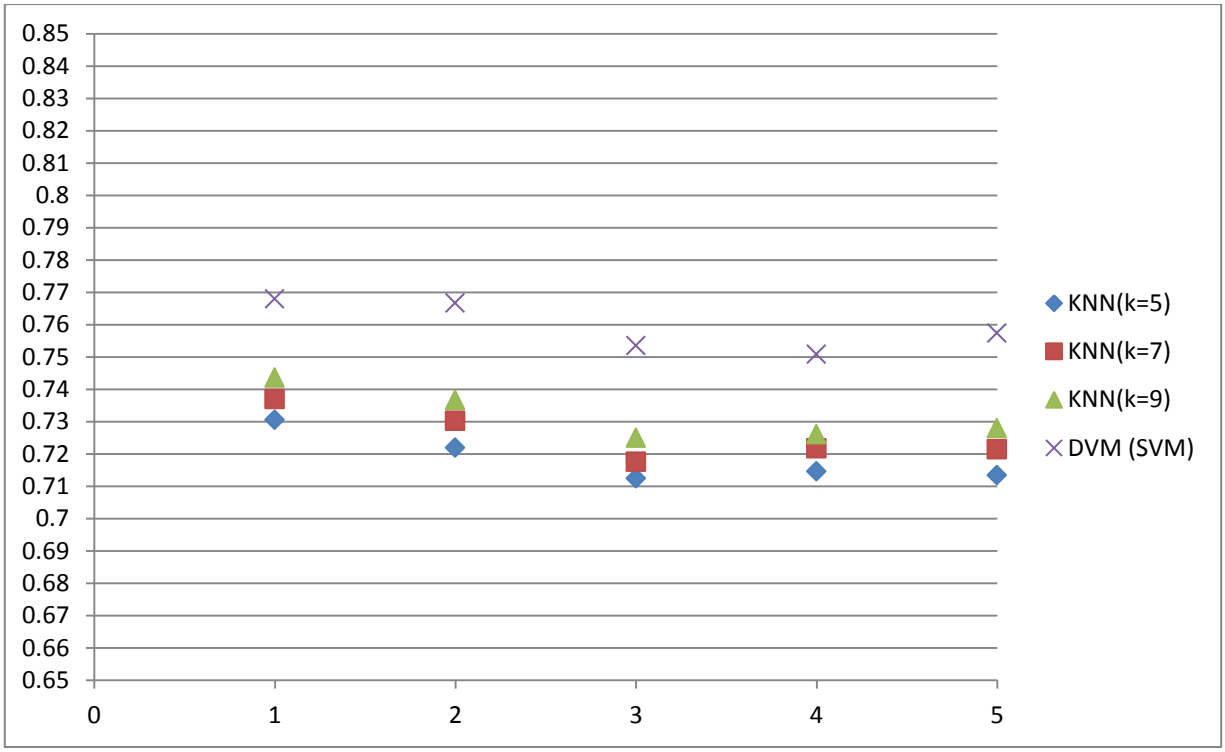
Bu kısımda ölçülen değerlerin karşılaştırma verileri yer almaktadır. Grafiklerde her bir test verisi üzerinden ortaya çıkan duyarlılık, anımsama ve f – ölçütü değerlerinin karşılaştırmaları yapılmıştır.



Şekil 5.1 - Duyarlılık kıyaslaması



Şekil 5.2 - Anımsama kıyaslaması



Şekil 5.3 - F - ölçütü kıyaslaması

Yukarıda karşılaştırılan yöntemlere bakıldığında; DVM, KNN' e göre daha iyi bir performans sergileyerek yüksek başarıda sonuçlar vermiştir.

Kaynak kullanımı açısından kıyaslandığında; DVM 9,5 GB bellek kullanırken, KNN yedi izlekli bir yapıda çalışmasına rağmen 3,5 GB bellek kullanmaktadır. Fakat CPU kullanımlarına bakıldığında DVM %50 civarı bir kullanım içerisindeyken, KNN %80 - %89 arasında bir CPU kullanımı göstermektedir. Bu kaynak kullanımlarıyla beraber DVM bir test kümesini yaklaşık 18 saat gibi bir sürede tamamlarken, KNN 11 saat civarlarında tamamlayabilmektedir.

Sonuç olarak bakıldığında bu çalışmanın esas aldığı ölçütler "Değerlendirme yöntemleri" kısmında da açıklanan duyarlılık, anımsama ve f – ölçütüdür. Çıkan oranlara bakıldığında DVM daha iyi bir sonuç vermiştir. Diğer yaklaşımların gösterdiği performanslar (MovieLens[26], MovieMagician[26], OpenMore[26],

ReMovender[26], CBCF[26], SEMCBF[26], SEMCBCF[26]) ile kıyaslandığında veri bütünleştirme yöntemi ile DVM algoritması kullanıldığında doğru tahmin açısından fayda sağlamıştır. “Yaklaşım – Sonuç” kıyaslamaları Tablo 5.44’ te gösterilmektedir.

Tablo 5.44 - Yaklaşım - sonuç kıyaslaması

Yaklaşım	**Duyarlılık(%)**	**Anımsama(%)**	**F - ölçütü(%)**
MovieLens	66	74	69.8
MovieMagician(öznitelik tabanlı)	61	75	67.3
MovieMagician(tıklama tabanlı)	74	73	73.5
MovieMagician(hibrid)	73	56	63.4
OPENMORE	75.2	73.7	74.4
ReMovender	72	78	74.9
CBCF	60	95.2	73.6
SEMCBF	63.4	92.3	75.2
SEMCBCF	63.7	93.1	75.6
DI - E (SVM)	71.9	82.4	76.8
DI - E (KNN)*	71.1	77.9	74.4

***En iyi sonucu k=9 eşitliğinde vermiştir.

6. TARTIŞMA

Bu tez çalışmasında, tavsiye sistemlerinde kullanılan yöntemlerin işleyişine doğrudan müdahele etmek yerine veri kümeleri üzerinde oynayarak daha verimli bir yapıya sokma işlemi amaçlanmıştır. Matris şeklinde gelen verinin iki yönlü olarak varolan vektörlerini birleştirmek yoluyla daha uzun ve öznitelik açısından daha zengin bir veri kümesi haline getirerek daha iyi sonuç elde edilip edilemediği test edilmiştir. Sınıflandırma algoritmaları olarak, en çok tercih edilenler olarak destek vektör makinesi ve k – en yakın komşu algoritmaları kullanılmıştır.

Gözlemleme ve değerlendirme aşaması, başarı performansı ve kaynak kullanımı olarak iki yönlü olarak ele alınmıştır. Kaynak kullanımı olarak bellek ve CPU kullanımı ile süre göz önünde bulundurulmuştur. Kullanılan iki sınıflandırma yöntemi bu kullanımlar ve süreler üzerinden kıyaslanmıştır.

Başarı performansı değerlendirilirken ölçümlendirme yöntemi olarak hata matrisinden yararlanılmıştır. Hata matrisi üzerinden elde edilen anımsama ve duyarlılık birimleri üzerinden sınıflandırma ölçütü olarak kullanılan f – ölçütü elde edilmiştir. Bu sonuçlar üzerinden öncelikle kullanılan iki yöntem arasında bir kıyaslama yapılmıştır. Bir sonraki adım olarak ise daha önceden kullanılan yaklaşımlarla kıyaslamalar yapılmıştır.

Sonuçlara bakıldığında ise veri bütünleştirildikten sonra; iki yöntem arasında başarı performans açısından DVM, KNN' e nazaran daha başarılı bir sonuç elde etmiştir. Fakat kaynak kullanımı açısından DVM daha fazla bellek kullanırken KNN karşıt olarak CPU kullanımında DVM' den daha yorucu bir algoritma olduğunu göstermiştir.

Diğer yaklaşımlar ile kıyaslandığında, veri bütünleştirme f – ölçütü üzerinden daha başarılı bir sonuç verdiği görülmüştür. Uğur Ceylan' ın yaptığı çalışmada kullanıcı puan - film matrisinde tek yönlü olarak kullanılan vektör yapısının aksine bu

çalışmada dikey ve yatay olan vektörler bütünleştirilerek daha uzun ve ve öznitelik sayısı açısından daha yoğun bir vektör elde edilmiştir.

Veri bütünleştirme yaklaşımı ile diğer yaklaşımlar arası kıyaslamaya tekrar bakıldığında bu yaklaşımın daha iyi sonuç verdiği görülmektedir. Yapılan çalışmalar sonucunda eğitici verinin tamamının kullanılması doğrultusunda aradaki kıyaslamanın da daha açık olacağı öngörülmektedir.

7. KAYNAKLAR LISTESI

1. Cohen, J., Special Issue on Information Filtering, Communications of the ACM, vol. 35, no: 12, s. 26-28, 1992.
2. Melville, P., Sindhwani, V., Recommender Systems, IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, 2002
3. www.hepsiburada.com
4. www.amazon.com
5. MacManus R., A Guide to Recommender Systems, http://www.readwriteweb.com/archives/recommender_systems.php, 2009
6. Anderson C.,The Long Tail: Why the Future of Business Is Selling Less of More, New York, NY: Hyperion, 2006
7. Oktar D., Tavsiye Sistemleri: Long Tail (Uzun Kuyruk) İle Karlılığı Artırmak, http://www.webrazzi.com/2009/09/09/tavsiye-sistemleri-long-tail-uzun-kuyruk-ile-karliligi-artirmak/, 2009
8. Karaman, H., A content based movie recommendation system empowered by collaborative missing data prediction, Orta Doğu Teknik Üniversitesi, M.Sc. tezi, 2010
9. van Meteren, R., van Someren M., Using content-based filtering for recommendation, vol. 4203/2006, Citeseer, 2000
10. Huang, Shiu-li, Comparision of Utility-Based Recommendation Methods, PACIS (Pacific Asia Conference on Information systems), 2008
11. Jannach D.,Zanker M., Felfernig A.,Recommender systems : an introduction, Cambridge U.P.,, 2010
12. Popescul, A. , Ungar, L., Pennock, D., Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments, 17th Conference in Uncertainty in Artificial Intelligence, University of Washington, Seattle, Washington, USA, 2001

13. Mukherjee, R., Sajja, N., Sen, S., A Movie Recommendation System - An Application of Voting Theory in User Modeling, User Modeling and User-Adapted Interaction , Kluwer Academic Publishers Hingham, MA, USA,2003
14. Öztürk, G., A hybrid video recommendation system based on a graph-based algorithm ,Orta Doğu Teknik Üniversitesi, M.Sc. tezi, 2010
15. Michie, D. , Spiegelhalter, D.J. , Taylor , C.C., Machine Learning Neural and Statistical Classification , Prentice Hall ,1994
16. Tsaptsinos, D., Mirzai, A., and Jervis, B. ,Comparison of machine learning paradigms in a classification task. In Rzevski, G., editor, Applications of artificial intelligence in engineering V: proceedings of the fifth international conference, Berlin. Springer-Verlag, 1990
17. Support vector machines, http://docs.oracle.com/cd/B28359_01/datamine.111/b28129/algo_svm.htm, 2011
18. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P., Numerical Recipes, Cambridge University Press, 3. Edition 2007
19. Crisitianini ,N., Taylor, J. S., An introduction to Support Vector Machines and other kernel Based learning methods, Cambridge University Press, 2000
20. Karagülle, F., Destek vektör makinelerini kullanarak yüz bulma,Trakya Üniversitesi, M.Sc. tezi, 2008
21. Antonini, G., Popovici, V., Independent Component Analysis and Support Vector Machine for Face Feature Extraction, Swiss Federal Institute of Technology Lausanne, AVBPA'03 Proceedings of the 4th international conference on Audio- and video-based biometric person authentication, 2003
22. Marli, A. H., Trends and Contreversies Support vector machines, IEEE, vol. 8/1998, s. 18-21 1998
23. Burges, C. J. C., A Tutorial on Support Vector Machines for Pattern Recognition, Data Mining and Knowledge Discovery , vol.2, s. 121-167 ,1998
24. Automatic Face and Gesture Recognition, Proceedings. Fourth IEEE International Conference,2000
25. http://www.grouplens.org

26. Ceylan, U., Content-Boosted Collaborative Filtering using Semantic Similarity Measure, 7th International Conference on Web Information Systems and Technologies (WEBIST'11),Noordwijkerhout, Netherlands, 2011
27. Herlocker , J. L., Konstan, J. A., Riedl, J. An algorithmic framework for performing collaborative filtering, New York, 1999
28. Chang ,C.-C., Lin C.-J., LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology,vol. 2, s. 1-27,2011
29. van Rijsbergen, C. J., "Information Retrieval". London; Boston. Butterworth, 2nd Edition, 1979
30. Lewis, H. G., Brown, M., A generalized confusion matrix for assessing area estimates from remotely sensed data, Remote Sensing, vol. 22, s. 3223-3235, 2001
31. http://www.bilisimsozlugu.net/
32. Olson, D. L., Delen, D. Advanced Data Mining Techniques, Springer, 1 edition ,February 1, 2008
33. Alkan, N., Precision and Recall Ratios in Evaluating the Quality of Literature Searches, Türk Kütüphaneciliği, vol. 4, s. 254-265,1994