GİZLİLİĞİ KORUYAN ÇOKLU-ÖLÇÜTLÜ ORTAK FİLTRELEME

Alper YARGIÇ

DOKTORA TEZİ

Bilgisayar Mühendisliği Anabilim Dalı
Danışman: Dr. Öğr. Üyesi Alper BİLGE

Eskişehir
Anadolu Üniversitesi
Fen Bilimleri Enstitüsü
Şubat 2019

JÜRİ VE ENSTİTÜ ONAYI

Alper Yargıç'ın "Gizliliği koruyan çoklu-ölçütlü ortak filtreleme" başlıklı tezi 05/02/2019 tarihinde aşağıdaki jüri tarafından değerlendirilerek "Anadolu Üniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliği'nin ilgili maddeleri uyarınca, Bilgisayar Mühendisliği Anabilim dalında Doktora tezi olarak kabul edilmiştir.

Jüri Üyeleri	**Unvanı Adı Soyadı**	**İmza**
Üye (Tez Danışmanı)	: Dr. Öğr. Üyesi Alper BİLGE	
Üye	: Prof. Dr. Yaşar HOŞÇAN	
Üye	: Doç. Dr. Cihan KALELİ	
Üye	: Dr. Öğr. Üyesi Mehmet KOÇ	
Üye	: Dr. Öğr. Üyesi Murat OKKALIOĞLU	

Prof.Dr. Ersin YÜCEL
Fen Bilimleri Enstitüsü Müdürü

ÖZET

GİZLİLİĞİ KORUYAN ÇOKLU-ÖLÇÜTLÜ ORTAK FİLTRELEME

Alper YARGIÇ

Bilgisayar Mühendisliği Anabilim Dalı

Anadolu Üniversitesi, Fen Bilimleri Enstitüsü, Şubat, 2019

Danışman: Dr. Öğr. Üyesi Alper BİLGE

Gizliliği koruyan ortak filtreleme sistemleri, tek-ölçüt değerlerinde bulunan gizlilik tehditlerini ortadan kaldırmaya odaklanır ve çoklu-ölçütlü tercih alanındaki gizlilik riskleri göz ardı edilir. Çoklu-ölçütlü tercih verilerinin yapısı, bireylerin neden bir öğenin kullanıcı tarafından tercih edildiğini anlama olanağı sağlamasına rağmen, bireyleri daha ciddi gizlilik tehditlerine maruz bırakır. Bu nedenle, bu sistemler esnek ve her bir alt-ölçütün yapısı ile uyum sağlayan akıllı koruma mekanizmalarına ihtiyaç duyar.

Bu tezde, çoklu-ölçütlü öneri sistemleri açısından mevcut gizlilik ihlali koşulları değerlendirilmekte ve bu hizmetlerin maruz kaldığı tehditler kapsamlı bir şekilde tartışılmaktadır. Bu tür tehditleri hafifletmek için, çoklu-ölçütlü ortak filtreleme sistemleri için rastgele karıştırma yöntemine dayalı gizlilik koruma yaklaşımları ve geleneksel tek-ölçütlü sistemlerde etkin bir şekilde kullanılan gizlilik koruma yöntemleri çoklu-ölçütlü derecelendirmelere uyarlanmaktadır. Öneri doğruluğunu arttırmak için, orijinal çoklu-ölçütlü tercihlerin bozulmasından kaynaklanan doğruluk kayıplarını azaltan entropi tabanlı rastgelelik belirleme prosedürünü uyarlayan yeni bir gizlilik koruma protokolü sunulmuştur. Ek olarak, olağandışı kullanıcı derecelendirmelerinin tahmin doğruluğu üzerindeki olumsuz etkilerini azaltmak için yeni bir veri karıştırma yaklaşımı sunulmuştur. Önerilen gizlilik koruma programlarının, farklı mahremiyet seviyelerinde hem kullanıcı gizlilik seviyeleri hem de tahmin doğruluğu üzerindeki etkilerini göstermek için yaklaşımlar Yahoo!Movies çoklu-ölçütlü tercih veri setinin üç alt grubunda deneysel olarak değerlendirilmiştir. Elde edilen deneysel sonuçlara göre, önerilen yaklaşımlar geleneksel gizlilik koruma senaryosunun sağladığı gizlilik seviyesini korurken önemli ölçüde başarımı yüksek öneriler üretebilmektedir.

Anahtar Kelimeler: İşbirlikçi filtreleme, Çoklu-ölçüt, Gizlilik

ABSTRACT

PRIVACY-PRESERVING MULTI-CRITERIA COLLABORATIVE FILTERING

Alper YARGIÇ

Department of Computer Engineering

Anadolu University, Graduate School of Sciences, February, 2019

Supervisor: Assist. Prof. Dr. Alper BİLGE

Privacy-preserving collaborative filtering systems focus on eliminating the privacy threats inherent in single preference values, and the privacy risks in the multi-criteria preference domain are disregarded. The structure of multi-criteria preference data exposes individuals to more severe privacy threats although it provides the opportunity to understand why an item is preferred by the user. Therefore, these systems require intelligent protection mechanisms that are flexible and adapting to the structure of each sub-criterion.

In this dissertation, existing privacy violation conditions from the perspective of multi-criteria recommender systems are evaluated and threats exposed by such services are discussed comprehensively. In order to alleviate such threats, randomized perturbation-based privacy-preserving approaches for multi-criteria collaborative filtering systems and the privacy protection methods efficiently used in traditional single-criterion systems are adapted onto multi-criteria ratings. To increase the prediction accuracy, a novel privacy-preserving protocol by adapting an entropy-based randomness determination procedure is introduced that can recover accuracy losses resulting from perturbation of original multi-criteria preferences. In addition, a novel data perturbation approach was introduced to mitigate the adverse effects of unusual user ratings on prediction accuracy. The proposed schemes are experimentally evaluated on three subsets of Yahoo!Movies multi-criteria preference dataset to demonstrate the effects of the proposed privacy-preserving schemes on both user privacy levels and prediction accuracy for differing sparsity rates. According to the obtained experimental outcomes, the proposed schemes can produce significantly more accurate predictions while maintaining an identical level of privacy provided by the traditional privacy protection scenario.

Keywords: Collaborative filtering, Multi-criteria, Privacy

TEŞEKKÜR

Tez çalışmam süresince yardımını ve desteğini esirgemeyen, bilgi birikimlerini paylaşarak bana yol gösteren danışmanım Dr. Öğr. Üyesi Alper BİLGE'ye teşekkürü bir borç bilirim.

Çalışmalarım sırasında bilgi ve görüşlerinden yararlandığım ayrıca tez komitemde bulunan Doç. Dr. Cihan KALELİ, Dr. Öğr. Üyesi Mehmet KOÇ, Prof. Dr. Yaşar HOŞÇAN ve Dr. Öğr. Üyesi Murat OKKALIOĞLU'na katkılarından dolayı teşekkür ederim.

Tez çalışmalarım sırasında benden yardımlarını esirgemeyen Araş. Gör. Dr. Burcu YILMAZEL ve Öğr. Gör. Emre KAÇMAZ'a teşekkür ederim.

Bu tezin bir kısmı 215E335 proje numarasıyla TÜBİTAK tarafından desteklenmiştir. TÜBİTAK'a maddi katkılarından dolayı teşekkür ederim.

Yaşamım boyunca benden hiçbir zaman sevgisini ve desteğini esirgemeyen, her konuda bana güvenen ve destek veren annem Ömür Yargıç'a, babam Fatih Yargıç'a ve abim Semih Yargıç'a teşekkür ederim.

Ayrıca, her zaman desteği ile yanımda olan değerli eşim Adife Şeyda YARGIÇ'a ve kızım Yağmur Gökçe YARGIÇ'a sonsuz teşekkür ederim.

Alper YARGIÇ

Şubat, 2019

05/02/2019

ETİK İLKE VE KURALLARA UYGUNLUK BEYANNAMESİ

Bu tezin bana ait, özgün bir çalışma olduğunu; çalışmamın hazırlık, veri toplama, analiz ve bilgilerin sunumu olmak üzere tüm aşamalarında bilimsel etik ilke ve kurallara uygun davrandığımı; bu çalışma kapsamında elde edilen tüm veri ve bilgiler için kaynak gösterdiğimi ve bu kaynaklara kaynakçada yer verdiğimi; bu çalışmanın Anadolu Üniversitesi tarafından kullanılan "bilimsel intihal tespit programı"yla tarandığını ve hiçbir şekilde "intihal içermediğini" beyan ederim. Herhangi bir zamanda, çalışmamla ilgili yaptığım bu beyana aykırı bir durumun saptanması durumunda, ortaya çıkacak tüm ahlaki ve hukuki sonuçları kabul ettiğimi bildiririm.

Alper YARGIÇ

İÇİNDEKİLER

Sayfa

Sayfa

Sayfa

TABLOLAR DİZİNİ

Sayfa

Sayfa

ŞEKİLLER DİZİNİ

Sayfa

SİMGELER VE KISALTMALAR DİZİNİ

$\bar{\cdot}$	:	Ortalama Benzerlik
$\lfloor \cdot \rfloor$	:	En Kötü Durum Benzerlik
σ	:	Sigma
σ_u	:	Standart Sapma
β	:	Beta
μ	:	Ortalama
a	:	Aktif Kullanıcı
ds	:	Gerçek Derecelendirme Sayısı
G	:	Gerçek Oy Değerleri
GB	:	Genel Beğeni Ölçütü
GBS_{max}	:	Genel beğeni değeri ve en yüksek entopi
$GKOF$	:	Gizliliği Koruyan Ortak Filtreleme
$GKOF^k$	:	Gizliliği Koruyan Çoklu-Ölçütlü Ortak Filtreleme
H	:	Shannon Entropisi
k	:	Ölçüt sayısı
m	:	Ürün Sayısı
MAE	:	Ortalama Mutlak Hata
mvr	:	Çoklu Değişkenli Regresyon
n	:	Kullanıcı Sayısı
$\mathcal{N}$	:	Normal
OF	:	Ortak Filtreleme
OF^k	:	Çoklu-Ölçütlü Ortak Filtreleme
$ÖS$	:	Öneri Sistemleri
PK	:	Pearson Korelasyonu
q	:	Hedef Ürün
R	:	Rastgele Sayı
r_0	:	Genel Derecelendirme Ölçütü
RD	:	Rastgele Doldurma
RK	:	Rastgele Karıştırma
RKT	:	Rastgele Karıştırma Teknikleri

S_{max}		En Yüksek Entropi
SBF	:	Sıra Dışılık Belirleme Fonksiyonu
SK	:	Sıra Dışı Kullanıcılar
SO	:	Sıra Dışı Oy Değerlerinin
TD	:	Tüm Derecelendirmeler
u	:	Kullanıcı
$\mathcal{U}$	:	Uniform
YM	:	Yahoo!Movies
Z	:	z-Skoru

1. GİRİŞ

İnternete erişim yaygın hale geldikçe, insanların eğilimleri günlük rutinlerini geleneksel yollardan ziyade çevrimiçi hizmetler üzerinden gerçekleştirmeye doğru evrilmektedir. Bunun nedeni, çevrimiçi hizmet sağlayıcılar tarafından sağlanan sınırsız miktardaki veriye kolay ve hızlı bir şekilde ulaşabilmesidir. Günümüzde bireyler; haberleri takip etmek, alışveriş yapmak, film izlemek, müzik dinlemek ya da konaklama rezervasyonu yapmak gibi günlük işlemleri gerçekleştirmek için çevrimiçi servislerden yararlanmaktadır (Ricci, Rokach ve Shapira, 2015).

İnternet kullanımı ve kullanıcı sayısındaki hızlı büyüme ile birlikte depolanması ve işlenmesi gereken veri miktarı da artmaktadır. Bu artış, aşırı bilgi yükleme problemi olarak adlandırılan soruna neden olarak, bireylerin karar verme sürecinin daha da karmaşık bir hal almasına neden olmaktadır (Ricci, Rokach ve Shapira, 2015). Bireyler günlük yaşamlarında, sezgisel olarak yakın çevresindeki kişilerin tecrübelerinden faydalanıp, karar verme sürecinde diğer kişilerin izlenimlerini referans alma yoluna giderler. Öneri Sistemleri (*ÖS*), bireylerin günlük hayatlarında istemsiz olarak gerçekleştirdikleri sözel tavsiyeye göre karar verme sürecini taklit edip otomatikleştiren ve onların aşırı bilgi yükleme sorunuyla başa çıkmasına yardımcı olan yazılım araçlarıdır (Adomavicius ve Tuzhilin, 2005). Özetle; *ÖS* bireylerin İnternet üzerinde ulaşmaya çalıştıkları ilgili ve değerli bilgiyi bulmaya yardımcı olmak için doğal sosyal rehberlik sürecini otomatikleştiren etkili araçlardır (Resnick ve Varian, 1997; Su ve Khoshgoftaar, 2009). *ÖS* hakkında son yıllarda yapılan çalışmalarda; otomatik öneri üretme sürecinde içerik tabanlı, Ortak Filtreleme (*OF*), bilgi tabanlı ve bu yöntemlerin melezlendiği tekniklerden oluşan birçok yaklaşım önerilmektedir (Burke, 2002; Adomavicius ve Tuzhilin, 2005; Ekstrand vd., 2011; Bobadilla vd., 2013).

1.1. Ortak Filtreleme

ÖS içerisinde *OF* teknikleri, otomatik öneri üretme sürecinde oldukça başarılı sonuçlar elde edilmesine olanak sağlayan ve birçok e-ticaret uygulamasında yaygın olarak kullanılan bir bilgi filtreleme tekniğidir (Adomavicius ve Tuzhilin, 2005). Bu kavram ilk olarak Tapestry projesi olarak isimlendirilen e-posta filtreleme servisinin geliştiricileri Goldberg vd. (1992) tarafından ortaya konmuştur. Sonrasında, Amazon, YouTube, Last.fm, Yahoo!Movies, TripAdvisor gibi kullanıcılarının geçmiş beğenilerine

göre önerilere ihtiyaç duyduğu birçok servis ile bütünleşerek kullanımı gün geçtikçe daha da yaygınlaşmıştır.

OF sistemleri; kullanıcılarının daha önceden deneyimledikleri çeşitli hizmet ya da ürünlere verdikleri derecelendirme değerlerinden yola çıkarak, kullanıcıların ilgilendikleri ancak daha önceden deneyimlemedikleri ürün/hizmetler hakkında tahmin üretilmesini sağlayan etkili yazılım araçlarıdır. *OF* sistemlerinin öneri üretme sürecinde referans aldığı temel yaklaşım, geçmişte benzer beğeni profillerine sahip kullanıcıların gelecekte de benzer beğenilere sahip olacağı varsayımıdır. *OF* sistemleri bu varsayımdan yola çıkarak kullanıcılarının servis sağlayıcı tarafından sunulan ürün ya da hizmet hakkındaki derecelendirme verilerini toplayıp, elde edilen kullanıcı verilerini analiz ederek üretilecek tahminler için yararlı bilgiler çıkarmaktadır (Goldberg vd., 2001). *OF* sistemleri tarafından kullanılan algoritmalar, bir kullanıcı tarafından ilgilendiği bir ürün/hizmet ile ilgili öneri talep edildiğinde bu işlemi *kullanıcı* × *ürün* matrisi olarak isimlendirilen *n* tane kullanıcının *m* tane ürün hakkındaki tercih derecelendirmelerini içeren veri tabanları üzerinde gerçekleştirmektedir. Genel olarak bu veri tabanları büyük miktardaki kullanıcı ve ürün sayısına sahip olmakla birlikte oldukça seyrek derecelendirme sayılarına sahiptir. Örneğin TripAdvisor, 390 milyon aylık ziyaretçiye sahip olan otel, 7 milyona yakın konaklama, restoran ve turistik yer hakkında 435 milyon inceleme ve görüşün toplandığı bir veri tabanına sahiptir. Kullanıcılara ait tercih verileri temel olarak iki farklı yaklaşım ile elde edilmektedir. Bunlardan ilki kullanıcı tarafından sisteme açıkça beyan edilen nümerik ürün beğeni değerleri ya da ürünlere ait metinsel kullanıcı yorumlarıdır. İkinci yaklaşımda ise servis sağlayıcı tarafından kullanıcının geçmişte ziyaret ettiği sayfalar, sayfada geçirilen zaman, kullanıcının aktif olduğu zaman dilimi ve süresi gibi etkinlik geçmişini kullanarak dolaylı olarak elde edilen verilerdir (Ricci, Rokach ve Shapira, 2015).

OF sistemlerinde öneri üretme sürecinde kullanılan yaklaşımlar temel olarak hafıza tabanlı ve model tabanlı yaklaşımlar olarak iki ana başlık altında toplanmaktadır (Herlocker vd., 2004; Bobadilla vd., 2013). Hafıza tabanlı yöntemler kullanıcının geçmiş tercihlerine dayalı derecelendirmelerini içeren bütün *kullanıcı* × *ürün* matrisi üzerinde işlem yapmaktadır ve kullanıcılar arasındaki korelasyonları referans alarak öneri üretme işlemini gerçekleştirmektedir (Herlocker vd., 1999; Sarwar vd., 2001). Model tabanlı yaklaşımlar tüm *kullanıcı* × *ürün* matrisi yerine bu veri setinden elde edilen model

prototipi üzerinde öneri üretme işlemini gerçekleştirmektedir (O'Connor ve Herlocker, 1999; Goldberg vd., 2001).

Hafıza tabanlı yaklaşımlar *OF* sistemlerinin ortaya çıktığı günden günümüze yaygın bir şekilde kullanılan, yüksek seviyede uygulanabilirliğe sahip yaklaşımlardır (Goldberg vd., 1992; Resnick vd., 1994; Aggarwal, 2016). Bu yaklaşımda komşuluk belirleme işlemi kullanıcı tabanlı ve ürün tabanlı olarak iki farklı teknik kullanılarak gerçekleştirilmektedir. Kullanıcı tabanlı yaklaşımda; öneri isteyen aktif kullanıcıya en benzer oy verme profiline sahip kullanıcılar belirlenmekte ve belirlenen kullanıcıların oy değerleri kullanıcı benzerlik katsayıları ile ağırlıklandırılarak öneri üretme işlemine dahil edilmesi temeline dayanmaktadır (Herlocker vd., 2004; Aggarwal, 2016). Öte yandan, ürün tabanlı yaklaşımlar kullanıcı tabanlı yaklaşımın aksine, ürünler arasındaki benzerliklere dayalı olarak öneri üretme işlemini gerçekleştirmektedir (Sarwar vd., 2001; Linden, Smith ve York, 2003; Deshpande ve Karypis, 2007). Benzerlik tabanlı yaklaşımda komşular arası benzerlikleri elde etmek için Pearson Korelasyonu (*PK*) (Resnick vd., 1994), kosinüs benzerliği (Sarwar vd., 2001), uzaklık tabanlı ve entropi tabanlı (Herlocker, Konstan ve Riedl, 2002) benzerlik metrikleri yaygın olarak kullanılmaktadır.

Üretilen önerilerin doğrudan bütün kullanıcı derecelendirmeleri üzerinden elde edildiği hafıza tabanlı yaklaşımların aksine model tabanlı yaklaşımlar *kullanıcı* × *ürün* matrisini tahmin modelini oluşturmak için kullanmaktadır (Ning, Desrosiers ve Karypis, 2015). En yaygın olarak kullanılan modeller arasında, Bayes sınıflandırıcıları (Breese, Heckerman ve Kadie, 1998; Cho, Hong ve Park, 2007), sinir ağları (Ingoo, Kyong ve Tae, 2003), bulanık sistemler (Yager, 2003), genetik algoritmalar (Ho, Fong ve Yan, 2007; Gao ve Li, 2008), matris faktörizasyonu (Luo, Xia ve Zhu, 2012), destek vektör makineleri (Grcar vd., 2012) ve tekil değer ayrışımı (Takács vd., 2008; Takács vd., 2009) gibi yöntemler bulunmaktadır (Bobadilla vd., 2013).

Hafıza-tabanlı ve model-tabanlı yaklaşımların kendine özgü avantaj ve dezavantajları bulunmaktadır. Örneğin, model tabanlı yaklaşımlar çevrimiçi işlem süresi olarak hafıza tabanlı yaklaşımlara göre daha iyi bir performans sunarken, üretilen öneri doğrulukları bakımından kıyaslandığında hafıza tabanlı yaklaşımlara göre hata değeri yüksek öneriler üretmektedir. Ayrıca, model tabanlı yaklaşımlarda yeni ürün ve kullanıcıların sisteme dâhil edilmesi model oluşturma süreci nedeniyle, hafıza tabanlı yaklaşımlara kıyasla daha zordur. Bu nedenle araştırmacılar, her iki yaklaşımın da

avantajlı taraflarının alınıp melezlendiği hibrit yaklaşımları sunmaktadır (Bobadilla vd., 2013).

1.2. Çoklu-Ölçütlü Ortak Filtreleme

Geleneksel *OF* sistemlerinde kullanıcılar bir ürün/hizmeti bütün yönleriyle değerlendirir ve deneyimledikleri ürün/hizmete ait beğeni miktarını derecelendirmek için tek bir genel beğeni değerini servis sağlayıcı sisteme sunarlar. Bu sistemler halen birçok *ÖS* tarafından başarıyla kullanılsa da, kullanıcıların bir ürün/hizmeti hangi nedenlerden dolayı sevdiği veya sevmediği bilgisini tek-ölçütlü *OF* sistemleri ile elde etmek mümkün değildir. *ÖS* kullanıcı profillerini daha iyi analiz ederek onların özel beğenilerine özgü daha geniş bir spektrumda tavsiyelerde bulunmak ve daha doğru öneriler üretebilmek için Çoklu-Ölçütlü Ortak Filtreleme (OF^k) algoritmaları kullanmaktadır (Adomavicius ve Tuzhilin, 2005; Adomavicius ve Kwon, 2007). OF^k sistemlerinde bir kullanıcının belirli bir ürün/hizmeti derecelendirirken, ürün/hizmet hakkında sadece genel beğeni derecelendirmesi vermek yerine, bu ürün/hizmete ait birden çok ölçüt için ayrı ayrı puanlama yapması istenmektedir. Böylece, OF^k sistemleri kullanıcılarından kişisel beğenileri ile ilgili daha fazla veri toplamaktadır. Bunun sonucunda kullanıcı beğeni profilleri daha ayrıntılı bir şekilde analiz edilerek üretilen öneri doğruluklarının arttırılması hedeflenmektedir (Jannach, Karakaya ve Gedikli, 2012).

ÖS tarafından kullanılan güncel yaklaşımlar, belirli bir öğenin ayırt edici alt özellikleri üzerinde çoklu-ölçütlü derecelendirmelerin kullanılmasını teşvik etmektedir (Adomavicius ve Kwon, 2007; Adomavicius ve Kwon, 2015). Örneğin, ünlü bir film öneri sistemi olan Yahoo!Movies web servisi; bir filmin yönetmenlik, oyunculuk, senaryo ve görsellik alt-ölçütleriyle birlikte kullanıcının film hakkındaki toplam beğeni değerini ifade etmek için kullanılan genel beğeni ölçütünü kayıt altına almaktadır. Başka bir örnekte ünlü restoran rehberi Zagat.com; öneri üretirken kullanıcı eğilimlerini daha hassas bir şekilde analiz etmek için yiyecek kalitesi, hizmet, dekor ve maliyet ölçütleriyle ilgili kullanıcı derecelendirmeleri toplamaktadır. Çoklu-ölçütlü kullanıcı verileri üzerinde çalışan OF^k yöntemleri ilk olarak Adomavicius ve Kwon (2007) tarafından önerilmiştir. OF^k sistemleri üzerinde yapılan çalışmalarda elde edilen ampirik sonuçlar, çoklu-ölçüt kullanımının tahmin doğruluğunu artırma potansiyeline sahip olduğu gösterilmektedir. Araştırmacılar ayrıca, farklı amaçlar için toplu kullanıcı verileri üzerinde çalışan çeşitli OF^k sistemleri örnekleri sunmaktadır. Bu doğrultuda Naak, Hage ve Aimeur (2009)

çoklu-ölçütlere dayalı olarak akademisyenlere makaleler tavsiye etmek için bir OF^k sistemi önermektedir. Fuchs ve Zanker (2012) tarafından yapılan çalışmada ise, TripAdvisor.com tarafından toplanan çoklu-ölçütlü kullanıcı verileri analiz edilmekte ve yapılan bu çalışma turizm sektöründe tavsiyeler isteyen kullanıcılara rehberlik etmektedir. OF^k sistemlerinin geçmişine dair yapılan çalışmalar Adomavicius ve Kwon (2015) tarafından ayrıntılı bir şekilde analiz edilmektedir.

OF^k yaklaşımları geleneksel *OF* sistemlerinde de olduğu gibi genellikle hafıza tabanlı ve model tabanlı yaklaşımlar olarak iki ana başlık altında gruplandırılmaktadır (Adomavicius ve Kwon, 2015). Geleneksel hafıza tabanlı *OF* sistemlerinde kullanılan benzerlik hesaplama tekniklerinin çoklu-ölçütlü veri setine uyarlandığı çalışmalarda, benzerlik değerleri ya her ölçüt için ayrı ayrı hesaplanıp bir araya getirilmekte ya da ölçütler için çok boyutlu benzerlik ölçüm teknikleri kullanılmaktadır. Bu yaklaşımlarda, her bir alt-ölçüt için korelasyon tabanlı ya da kosinüs tabanlı benzerlik hesaplama yaklaşımları kullanılarak, sistemden öneri talebinde bulunan aktif kullanıcı ile diğer kullanıcılar arasındaki benzerlik hesaplanmaktadır. Her bir ölçüt için elde edilen benzerlik katsayıları ortalaması veya elde edilen en kötü benzerlik katsayısı kullanılarak (Adomavicius ve Kwon 2007) ya da Tang ve McCalla (2009) tarafından önerildiği gibi her bir ölçüt için önerilen benzerlik katsayılarının ağırlıklı toplamı alınarak nihai benzerlik değeri elde edilmektedir. Diğer yaklaşımlarda, kullanıcı benzerliklerini elde etmek için Manhattan, Öklid ve Chebyshev gibi çok boyutlu uzaklık ölçüm metriklerine başvurulmuştur (Adomavicius ve Kwon 2007). Kullanıcı benzerlikleri elde edildikten sonra, geleneksel *OF* sistemlerinde kullanılan öneri üretme yaklaşımına benzer bir şekilde aktif kullanıcı için öneri üretme süreci gerçekleştirilmektedir. OF^k sistemlerinde geleneksel hafıza tabanlı yaklaşımlara ek olarak bulanık sistemlerin benzerlik elde etme sürecine dâhil edildiği çalışmalar da mevcuttur. Bunlara örnek olarak Nilashi, Ibrahim ve Ithnin (2014) tarafından bulanık tabanlı Öklid mesafesi ve bulanık tabanlı ortalama benzerlik yaklaşımlarının kullanılması önerilmektedir.

OF^k sistemlerinde model tabanlı yaklaşımlar, gerçek *kullanıcı* × *ürün* matrisini direkt olarak öneri üretme işleminde kullanmak yerine veri seti üzerinden elde edilen tahmin modelini kullanarak öneri üretme işlemini gerçekleştirmektedir. Çoklu-ölçütlü veri setlerinde model tabanlı yaklaşımlarda yaygın olarak; basit birleştirme fonksiyonları (Adomavicius ve Kwon 2007), olasılıksal modeller (Sahoo vd., 2012; Zhang vd., 2009), destek vektör regresyonu (Fan ve Xu, 2013; Jannach, Karakaya ve Gedikli, 2012;

Samatthiyadikun, Takasu ve Maneeroj, 2013) ve çoklu değişkenli tekil değer ayrıştırması (Li, Wang ve Geng, 2008) gibi yaklaşımlar kullanılmaktadır.

Birleştirme fonksiyonu yaklaşımlarında, benzerlik tabanlı sezgisel yaklaşımların aksine genel derecelendirme ölçütü r_0, diğer alt-ölçütlerden farklı olarak ele alınıp alt-ölçütlerin toplamını ifade eden ölçüt olarak varsayılmaktadır (Adomavicius ve Kwon 2007). Bu varsayım göz önünde bulundurularak, birleştirme fonksiyonu temelli yaklaşımlarda r_0 ve diğer alt-ölçütlerin sahip olduğu derecelendirme değerleri arasındaki ilişkiyi ifade eden birleştirme fonksiyonunun ($r_0 = f(r_1, \dots, r_k)$) bulunması hedeflenmektedir. Alt-ölçütlerin r_0 ölçütünü temsil etmekteki önem derecesini belirlemek için Adomavicius ve Kwon (2007) doğrusal regresyon kullanılmasını önermektedir. Diğer birkaç çalışma genel toplama fonksiyonu yaklaşımını takip etmektedir. Ancak geleneksel doğrusal en küçük kareler regresyon yöntemini kullanmak yerine, regresyon tabanlı derecelendirme toplama işlevlerini öğrenmek için destek vektör regresyonunun kullanıldığı çalışmalar da bulunmaktadır (Fan ve Xu, 2013; Jannach, Karakaya ve Gedikli, 2012; Samatthiyadikun, Takasu ve Maneeroj, 2013).

Bazı *OF*^{k} yaklaşımları, veri madenciliğinde ve makine öğrenmesinde giderek daha popüler hale gelen olasılıksal modelleme algoritmalarını benimsemektedir. Örneğin; Sahoo vd. (2012), Si ve Jin (2003) tarafından geliştirilen esnek karışım modeli çoklu-ölçütlü veri setleri üzerinde kullanılmaktadır. Bir başka olasılık modelleme yaklaşımında Zhang vd. (2009), tek-ölçütlü *OF* sistemleri için kullanılan olasılıksal gizli semantik analiz yaklaşımını çoklu-ölçütlü veri setlerine uyarlamıştır. Buna ek olarak Zhang vd. (2009); kullanıcıların ölçüt derecelendirme modellerini elde etmek için çok değişkenli normal dağılım ve doğrusal normal regresyon modelli olasılık modelleme yaklaşımlarını kullanmaktadır.

1.3. *OF*^{k} Sistemlerinin Zayıflıkları

OF^k sistemleri, kullanıcılara ait beğeni profillerini daha iyi kişiselleştirme ve analiz etme konusunda geleneksel *OF* sistemlerine göre daha başarılı sonuçlar elde etmesine rağmen halen birçok çözümlenmemiş problemle karşı karşıyadır. *OF*^{k} sistemleri yapısı gereği bir öğenin ayırt edici alt unsurları üzerinde çoklu-ölçütlü derecelendirmelerin toplanmasını teşvik etmektedir. Bunun sonucunda öneri doğruluğunun artması beklenirken, kullanılan çoklu-ölçütlü *kullanıcı* × *ürün* matrisi nedeniyle kullanıcılar daha ciddi mahremiyet ihlalleri ile karşı karşıya gelmektedir. Bu nedenle, kullanıcı profilleri

çoklu-ölçütlü veri setine özgü bir şekilde analiz edilmeli ve sistem tarafından üretilen önerinin doğruluğu ile sistem tarafından sağlanan veri gizliliği arasında bir denge kurulmalıdır.

1.3.1. Gizlilik

Bireyler öneri servislerini kullanırken kişisel bilgilerini ve ürün tercih bilgilerini servis sağlayıcılar ile paylaşmak mecburiyetinde kaldıkları için kişisel mahremiyetlerinin ihlal edildiğini düşünmektedir (Yargıç ve Bilge, 2017). Bu endişe neticesinde bireyler kişisel güvenliklerini sağlamak adına bazen *OF* servislerine yanlış beğeni değerleri beyan etmekte bazen de bu tür hizmetleri kullanmayı tamamıyla reddetmektedir (Berkovsky vd., 2007). Kullanıcıların kişisel mahremiyet kaygılarını gidermek için servis sağlayıcılar tarafından gizlilik koruma protokolleri geliştirilmeli ve ortaya konan bu yöntemler kullanıcı mahremiyetini sağlarken aynı zamanda da öneri üretme kalitesinden ödün vermemelidir (Canny, 2002).

Gizlilik kavramının kesin bir tanımı ve ölçeği olmaması nedeniyle öneri sistemleri alanında analiz edilmesi oldukça güçtür. Bunun temel nedeni mevcut gizlilik anlayışının büyük ölçüde farklı disiplinlere (hukuk, psikoloji, bilgi sistemleri vd.) ve farklı bireysel ihtiyaçlara bağlı olmasıdır (Yargıç ve Bilge, 2017). Farklı disiplinler altında yapılan çalışmalar, mahremiyet konusunda farklı öncelikler atfetmektedir. Bireylerin gizlilik konusunda *OF* sistemleri için öncelikleri; tecrübe edilen ürünlere vermiş oldukları kişisel derecelendirme değerlerinin ve derecelendirilmiş ya da derecelendirilmemiş öğeler kümesinin üçüncü şahıslara ifşa edilmemesidir (Polat ve Du, 2005). Kullanıcı mahremiyetinin daha ciddi olarak ihlal edilebileceği OF^k sistemlerinde bireylerin gerçek tercihlerini beyan etmelerini ve OF^k sisteminin daha kaliteli öneriler üretebilmesini sağlamak için kullanıcıların gizlilik konusundaki endişeleri giderilmelidir.

1.3.2. Doğruluk

OF^k sistemlerinin ortaya çıkışındaki temel motivasyon kullanıcı profillerini daha iyi analiz ederek kullanıcılarına daha doğru ve kişisel tercihleri ile uyumlu öneriler sunmaktır. Bu amaç doğrultusunda Bölüm 1.2'de özetlenen maskelenmemiş orijinal kullanıcı verileri üzerinde öneri kalitesini arttırmaya yönelik literatürde birçok çalışma mevcuttur (Su ve Khoshgoftaar, 2009; Bobadilla vd., 2013; Ning, Desrosiers ve Karypis, 2015). Ancak OF^k sistemi tarafından üretilen önerinin kalitesi kullanılan öneri üretme

tekniğinin yanı sıra üzerinde çalıştığı veri seti ve o veri setini oluşturan kullanıcıların oy verme eğilimleriyle de ilişkilidir (Adomavicius ve Kwon, 2015; Bilge ve Yargıç, 2017). Literatürde yaygın olarak kullanılan çoklu-ölçütlü veri setlerinde kullanıcıların, sistemdeki ürün sayısına oranla oy verdiği ürün/hizmet sayısı oldukça düşüktür (Jannach, Karakaya ve Gedikli, 2012; Nilashi vd., 2015). Günlük hayatta gerçek kullanıcılar tarafından kullanılan *ÖS*'nin sahip olduğu veri setlerindeki derecelendirme seyrekliği, kullanıcılar arasındaki benzerliklerin keşfedilmesini olumsuz olarak etkilemekle birlikte öneri kalitesini de düşürmektedir. Buna ek olarak, veri setini oluşturan kullanıcıların bir kısmının sıra dışı olarak nitelendirilen, genel kullanıcı profillerinden farklı eğilimler sergileyen oy verme alışkanlıkları *OF* sistemi tarafından üretilen önerilerin doğruluğunu olumsuz olarak etkilemektedir. Bu nedenle, *OF* sistemlerinde kullanıcı profilleri analiz edilip sıra dışı kullanıcı profilleri belirlenmeli ve bu kullanıcılara özgü veri maskeleme prosedürleri uygulanmalıdır.

1.4. Sunulan Çözümler ile İlgili Gelişmeler

OF^k sistemlerinde kullanıcı veri gizliliği ve sıra dışı oy değerlerinin öneri doğruluğuna etkisi araştırmacılar tarafından incelenmekte ve bahsedilen bu zayıflıkları hafifleten çeşitli çözümler sunulmaktadır. Bu bölümde, araştırmacılar tarafından sunulan çözümler listelenmektedir.

1.4.1. Gizliliği koruyan ortak filtreleme

OF sistemlerinin kullanımının yaygınlaşması ile birlikte, kullanıcı mahremiyetine getirdiği ek mahremiyet riskleri ve bunları engellemeye yönelik geliştirilen teknikler araştırmacılar tarafından incelenmektedir.

OF uygulamalarında gizlilik önlemleri oluşturmak için Canny (2002a, 2002b) dağıtık ortamlardaki kullanıcı profillerini gizlemek amacıyla kriptografik tekniklerden faydalanmaktadır. Geçtiğimiz yirmi yıl boyunca, Gizliliği Koruyan Ortak Filtreleme (*GKOF*) sistemlerinin geliştirilmesiyle ilgili birçok çalışma yapılmış ve farklı yaklaşımlar sunulmuştur. Bu yaklaşımlardan bazıları; kriptografi temelli yaklaşımlar, anonimleştirme yöntemleri, diferansiyel gizlilik, veri gizleme ve karıştırma temelli yöntemler ile veri gizleme işlemini gerçekleştirmektedir (Wei, Tian ve Shen, 2018).

ÖS için kriptografik teknikler genellikle dağıtık uygulamalarda kullanılan gizlilik koruma teknolojilerdir (Li vd., 2017a; Erkin vd., 2012; Badsha, Yi ve Khalil, 2016). Li

vd. (2016), hesaplama aşamaları sırasında kullanıcılara eşzamanlı olarak ihtiyaç duymadan, ürün benzerliklerini aşamalı olarak hesaplayan, güvenli çok-partili hesaplama protokolü sunmuştur. Shmueli ve Tassa (2017), hem kullanıcı derecelendirmelerini hem de derecelendirilen öğelerin sıralamasını paylaşan bir aracı ile servis sağlayıcının şifrelenmiş verilerini güvenli protokoller üzerinden paylaşan bir dağıtık öneri yaklaşımı sunmaktadır. Bir başka yaklaşımda Li vd. (2017a), öneri üretme sürecinde kullanıcı gizliliğinin korunmasına odaklanan grup tabanlı bir *ÖS* önermektedir. Zou vd. (2015) tarafından önerilen yarı-dağıtık *GKOF* yaklaşımında, ürün benzerliklerinin hesaplanması olasılıksal bir çıkarım problemi olarak modellenmiştir ve belirli bir zaman diliminde ağa yalnızca belirli bir kullanıcı alt-grubu dâhil edilerek kullanıcıların özel derecelendirmelerinin açıklanması engellenmiştir.

Anonimleştirme yaklaşımı, bireysel kullanıcıyla derecelendirme profili arasındaki ilişkiyi gizlemek için kullanılan diğer popüler yaklaşımlardan biridir. Casino vd. (2015), kullanıcı gizliliğini korumak için birleştirme tabanlı bir *k*-anonimleştirme maskeleme yaklaşımı önermektedir. Chen ve Huang (2012), *ÖS* gibi büyük ölçekli ve seyrek veri setleri için anonimlik elde ederek gizliliği koruyan kümelenme-temelli bir yaklaşım önermektedir. Ayrıca Wei, Tian ve Shen (2018) mevcut anonimleştirme yöntemlerini geliştirmek için (p, l, α)-çeşitlilik metodunu önermektedir. Burada, p kullanıcının derecelendirme vektörü hakkında önceki bilgileri temsil ederken, l ve α kullanıcıların gizlilik seviyesini arttırmak için kullanılan çeşitlilik değişkenleridir.

Diferansiyel gizlilik, bir kullanıcıya ait özel derecelendirme tercihleri öneri üretme işleminde kullanıldığında, ortaya çıkabilecek en yüksek düzeydeki gizlilik kaybının ölçülmesi için *OF* sistemlerinde faydalanılan diğer etkili tekniktir. Diferansiyel gizlilik kavramı ilk olarak Dwork vd. (2006) tarafından ortaya konmuş olup, gerçek kullanıcı derecelendirmelerini rastgele hale getirmek için özel bir kovaryans matrisi kullanan McSherry ve Mironov (2009) tarafından öneri sistemleri alanına dâhil edilmiştir. Benzer bir yaklaşımda Guerraoui vd., (2015), değiştirilmiş profilleri ortaya çıkarmak için kullanılan uzaklık tabanlı bir diferansiyel gizlilik yaklaşımı önermektedir. Başka bir yaklaşımda Shen ve Jin (2014), teorik olarak gizliliği garanti etmek için oluşturulan maskeleme verisinin kalibrasyonunu sağlayan, tercih profillerine örneğe dayalı gürültü verisi ekleyen bir mekanizma önermiştir. Ayrıca, Hou vd. (2018) tıbbi *ÖS* için gizli komşu seçimi ve komşuluk-temelli diferansiyel gizlilik için bir metodoloji önermektedir. Li vd. (2017b), kullanıcı düzeyinde gizliliği garantilemek yerine ürüne özel gizlilik

gereksinimlerini vurgulayan yerel-kümelenme tabanlı kişiselleştirilmiş bir diferansiyel gizlilik protokolü sunmaktadır.

Veri gizleme ve Rastgele Karıştırma Teknikleri (*RKT*) gibi veri modifikasyonlarına dayalı yaklaşımlar, *OF* sistemlerinde gizlilik sorunlarını ele almak için yaygın olarak kullanılan diğer gizlilik koruma mekanizmalarıdır. Gizleme teknikleri, kullanıcıların kişisel mahremiyetini korumak için kullanıcılara ait gerçek derecelendirme verisini; önceden tanımlanmış sabit değerlerle, derecelendirme aralığında belirlenmiş rassal değerlerle ya da örnek olarak normal dağılım gibi belirli dağılımlar aracılığıyla belirlenen rassal sayılarla gerçek kullanıcı derecelendirmelerini maskeleyerek kişisel mahremiyeti koruyan sistemlerdir. Bununla birlikte; yapılan gizleme işlemi sonrasında *OF* sistemi tarafından öneri üretmek amacıyla gerekli bilgileri hala erişilebilir halde tutan gizlilik koruma teknikleridir (Berkovsky vd., 2007; Berkovsky, Kuflik ve Ricci, 2012; Bilge vd., 2013). Bu doğrultuda, Parameswaran ve Blough (2007) merkezi sunucu tabanlı *OF* uygulamaları için permütasyon tabanlı veri gizleme yöntemini kullanmaktadır. Başka bir yaklaşımda (Boutet vd., 2016) dağıtık öneri sistemlerinde kullanıcı profillerinin gizlenmiş sürümlerinin kullanılması önerilmektedir. Geleneksel tek seviyeli gizlemeye dayalı yaklaşımlara ek olarak Elmisery ve Botvich (2017) daha güvenilir hedef kullanıcıları tespit etmek için rastgele seçilmiş gizleme seviyelerinin kullanılmasını önermektedir.

Alternatif olarak *RKT*; kullanıcıya ait derecelendirmeleri öneri sunucusuna göndermeden önce gerçek kullanıcı derecelendirmelerini deforme ederek kullanıcı gizliliğini sağlamaktadır (Polat ve Du, 2003; Polat ve Du, 2005a; Polat ve Du, 2005b). Bu doğrultuda, Bilge ve Polat (2013), rastgele veri karıştırma yaklaşımlarına dayanan ölçeklenebilir bir *GKOF* yaklaşımı önermektedir. Başka bir çalışmada, Gong (2011) öneri süreçlerinde çoklu veri depoları üzerinden dağıtılan profillerin kullanıcı gizliliğini sağlamak için *RKT* ve gizliliği sağlanmış çok taraflı hesaplama yöntemlerini birleştirmektedir. Polatidis vd. (2017) rastgele değerler üzerinde çoklu gizlilik seviyelerine göre belirlenen değişken değer aralıkları içerisinde *RKT* ile üretilen sayılar aracılığıyla gerçek kullanıcı derecelendirmelerini maskeleyen gizlilik koruma protokollerini önermektedir. Bu çalışmalara ek olarak, Liu vd. (2017) mevcut *RKT* tabanlı koruma mekanizmalarına göre daha yüksek seviyede gizlilik elde etmek için *RKT* ve diferansiyel gizlilik metodunun birleştirildiği melez bir gizlilik koruma yaklaşımı önermektedir.

1.4.2. Sıra dışı kullanıcı ve derecelendirme problemi

Sıra dışı kullanıcı olarak tanımlanan ve sisteme kayıtlı diğer kullanıcılar ile genel anlamda uyumsuz derecelendirme geçmişine sahip kullanıcılar *OF* sistemleri tarafından yanlış önerilere maruz kalmaktadır (Claypool vd., 1999). Dahası, sıra dışı derecelendirme geçmişine sahip kullanıcılar diğer kullanıcıların komşuluklarına dâhil olarak bütün sistemin öneri üretme kalitesini olumsuz olarak etkilemektedir (Ghazanfar ve Prügel-Bennett, 2014).

Sıra dışı kullanıcı problemini hafifletmek için içerik tabanlı yaklaşımlar ve semantik web madenciliği aracılığı ile elde edilen semantik verileri birleştirerek öneri kalitesi bakımından daha güvenli modeller oluşturulabilmektedir (Kim vd., 2011; Moreno vd., 2016). Ancak önerilen yöntemler kullanıcı oy değerleri dışında ek bilgi gerektiren uygulamalar olduğu için sisteme ek maliyet getirmektedir. Ghorbani ve Novin (2016) daha kolay bir yaklaşım ortaya koyarak kümeleme yöntemlerinin uygulanmasını önermektedir. Ghazanfar ve Prügel-Bennett (2014) tarafından gerçekleştirilen çalışmada, kümeleme yöntemleri ayrıntılı olarak ele alınmakla birlikte, sıra dışı kullanıcı profillerinin *k*-ortalama kümeleme kullanılarak tespit edildiği ve kullanıcı profillerine dayalı önerilerin sunulduğu bir yaklaşım önerilmektedir.

Kümeleme tabanlı yöntemlerin dışında Gras, Brun ve Boyer (2016), kullanıcı derecelendirmelerinin dağılımına dayalı olarak oy dağılımlarında ortaya çıkan sapmaları sıra dışı kullanıcıları etiketlemek için kullanmaktadır. Buna ek olarak, üretilen önerilerde ortaya çıkan beklenmedik seviyelerdeki hataları da sıra dışı kullanıcıları belirlemek için kullanmaktadır. Ancak, öneri doğruluğunda ortaya çıkan hata değerleri yalnızca bir kullanıcının sıra dışı olup olmadığını değerlendirmek için yeterli değildir. Bunun nedeni, sıra dışı oy verme eğiliminin tek başına büyük öneri hatalarına yol açan tek neden olmamasıdır. Diğer bir deyişle, büyük tahmin hatalarıyla ilişkili bir kullanıcının sadece hata oranları referans alınarak sıra dışı olarak nitelendirilmesi her zaman için doğru değildir (Sánchez-Moreno vd., 2016). Bir diğer yaklaşımda, sıra dışı olarak tanımlanan kullanıcıların diğer birçok kullanıcıyla düşük korelasyon göstermeleri ve çok az sayıda kullanıcının komşuluğuna girebilmelerinden yararlanılarak, bu kullanıcıları belirleme işlemi gerçekleştirilmektedir (Claypool vd., 1999). Kullanıcı benzerliklerinden yararlanılarak sıra dışı kullanıcı profillerinin belirlendiği diğer bir yaklaşımda ise; kullanıcıların birbirleri ile olan benzerlik ilişkileri istatistiksel olarak analiz edilip, aykırı saplamalara neden olan kullanıcı profilleri sıra dışı olarak sınıflandırılmaktadır (Zheng,

Agnani ve Singh, 2017b). Ancak, kullanıcı benzerliğine dayalı ikili korelasyonlar kullanılarak yapılan sınıflandırmalar bazı dezavantajlara sahiptir. Bunlardan ilki; eğer kullanıcılar arasında ortak derecelendirilmiş ürünler yoksa kullanıcı tabanlı benzerlik değerinin hesaplanması mümkün değildir. Diğer problem ise iki kullanıcı tarafından ortak olarak derecelendirilen öğelerin sayısı sınırlı olduğu durumlarda, kullanıcı benzerlik değerinin gerçeği yansıtıp yansıtmadığının garantisi verilemez.

1.5. Amaç ve Katkılar

OF sistemleri ile ilgili yapılan güncel çalışmalar, belirli bir öğenin ayırt edici alt özellikleri üzerinde çoklu-ölçütlü tercihlerin kullanılmasını teşvik etmektedir (Adomavicius ve Kwon, 2015). Bu durum, kullanıcıların bir öğeye ilişkin genel beğeni seviyesinin alt-ölçüt derecelendirmelerinden bağımsız olarak kabul edilemeyeceğinin de göstergesidir. Genel derecelendirme ve alt-ölçütler arasındaki bu ilişki tek-ölçütlü *OF* sistemlerine göre daha ciddi gizlilik risklerini de beraberinde getirmektedir. Bu tez çalışmasının amacı; çoklu-ölçütlü veri seti kullanımı ile ortaya çıkan gizlilik ihlallerini hafifletirken aynı zamanda öneri doğruluğundan ödün vermeyen gizliliği koruyan yaklaşımlar geliştirmektir.

OF sistemlerinde kişisel tercihlerin toplanması nedeniyle kullanıcıların maruz kaldıkları gizlilik riskleri literatürde tartışılmaktadır. Ancak, bu çalışmalar geleneksel tek-ölçütlü tercih değeri kullanan *OF* sistemlerinin maruz kaldığı tehditler üzerinde durmakta ve çoklu-ölçütlü tercih verisi alanında ortaya çıkabilecek gizlilik risklerinin değerlendirilmesinde yetersiz kalmaktadır. Bu eksikliği gidermek amacıyla; çoklu ölçütlü veri setlerinde kullanıcı verisine doğrudan ya da dolaylı yöntemlerle sahip olan kötü niyetli kişilerin neden olabileceği gizlilik ihlalleri ayrıntılı bir şekilde tanımlanmaktadır.

Geleneksel *GKOF* sistemleri, tekil tercih değerlerinin doğasında bulunan gizlilik tehditlerini ortadan kaldırmaya odaklanmaktadır ve çoklu-ölçütlü tercih verileri alanındaki gizlilik riskleri göz ardı edilmektedir. OF^k sistemlerinde ortaya çıkan bu eksikliği ortadan kaldırmak için *RKT* temelli gizlilik koruyucu yaklaşımlar sunulmaktadır. Geleneksel tek-ölçütlü *OF* sistemlerinde verimli bir şekilde kullanılan *RKT* temelli gizlilik koruma yöntemleri çoklu-ölçütlü derecelendirme verilerine adapte edilmiştir. Böylece OF^k sistemleri için Rastgele Karıştırma (*RK*) ve Rastgele Doldurma (*RD*) yaklaşımları ile elde edilen referans gizlilik ve doğruluk seviyeleri elde edilmiştir.

Geleneksel veri maskeleme yaklaşımlarının çoklu-ölçütlü veri setlerinde kullanılmasıyla ortaya çıkan en büyük dezavantaj, veri mahremiyetini sağlarken önerilen tahminlerin doğruluğunda büyük kayıplara neden olmasıdır. Bu nedenle, elde edilen gizlilik seviyeleri ile tahmin doğruluğu arasında bir denge kurmak çok önemlidir. Çoklu-ölçütlü veri alanında, tüm alt-ölçütler kullanıcılar için aynı düzeyde önem seviyesine sahip değildir ve ölçütler arasında bir bağımlılık söz konusudur. Her bir alt-ölçüte ait tercih vektörünün genel beğenme üzerinde farklı etkileri olduğu iddia edilebilir. Bu yaklaşım farklı önem derecesine sahip ölçütlerin farklı gizlilik seviyelerinde maskelenmesi temeline dayanır. Bu amaçla, maskeleme işlemi sırasında ortaya çıkacak öneri doğruluk kayıplarını hafifletmeyi hedefleyen, kullanıcı ve kullanıcının ölçütleri değerlendirme alışkanlıklarına göre maskeleme işleminde kullanılacak gizlilik parametrelerini belirleyen entropi tabanlı yeni gizlilik-koruma protokolleri geliştirilmiştir. Böylece, geleneksel veri maskeleme yaklaşımı ile elde edilen gizlilik seviyeleri muhafaza edilirken aynı zamanda öneri üretme doğruluğunda ortaya çıkan kayıplar hafifletilmiştir.

GKOF sistemlerinde öneri doğruluğunda ortaya çıkan kayıpların diğer bir nedeni de kullanıcıların sıra dışı oy verme eğilimleridir. Geleneksel *GKOF* sistemlerinde, veri maskeleme işlemi istemci tarafında gerçekleştirilen bir süreçtir ve sunucu sadece maskelenmiş *kullanıcı* × *ürün* matrisine sahiptir. Maskelenmiş veri setleri üzerinde geleneksel kullanıcı tabanlı sıra dışılık belirleme stratejileri yetersiz kaldığından, bu sistemin yerine aktif kullanıcıya ait derecelendirme vektörü üzerinde ürün tabanlı sıra dışılık belirleme yaklaşımları geliştirilmiştir. Bu yaklaşımda, kullanıcı derecelendirme vektörü sunucuya ulaşmadan istemci tarafında, aktif kullanıcının oy verme eğilimi analiz edilip, kullanıcının oy verme eğilimine göre standart dışı oy değerleri ile derecelendirdiği ürün listesi belirlenmekte ve bu derecelendirmelere özgü olarak yeni veri maskeleme yaklaşımı uygulanarak öneri doğruluğunda gözlemlenen kayıp hafifletilmektedir.

2. GENEL BİLGİLER

Bu bölümde *OF* sistemlerinde öneri üretme süreci ve kullanıcı veri gizliliğini koruma mekanizmalarına dair genel bilgiler açıklanmaktadır. Öncelikle; *OF* ve OF^k sistemlerinde herhangi bir veri gizleme işlemine tabi tutulmamış ham kullanıcı verileri üzerinde öneri üretme süreci tanımlanmıştır. Sonrasında, kullanıcı gizliliğini korumaya yönelik uygulanan veri koruma mekanizmaları tanımlandıktan sonra *RK* ve *RD* yöntemleri geleneksel *OF* sistemleri üzerinde açıklanmıştır. Son olarak, deneysel olarak kullanılan veri setleri ve değerlendirme ölçütleri açıklanmıştır.

2.1. Ortak Filtreleme Sistemlerinde Öneri Üretme Süreci

OF sistemleri, *m* kullanıcıdan elde edilen *n* tane ürün derecelendirmesinden oluşan *kullanıcı* × *ürün* matrisi üzerinde öneri üretme işlemini gerçekleştirmektedir. *OF* sistemlerinde öneri oluşturma süreci boyunca, bir aktif kullanıcı (a), geçmişte derecelendirmiş olduğu öğelerine ait *kullanıcı* × *ürün* vektörünü servis sağlayıcı ile paylaştıktan sonra bir hedef ürün (q) için öneri talep eder. *OF* sisteminin öneri oluşturma sürecinde, kullanıcılar geçmişte derecelendirmiş olduğu ürün listesini sistem ile paylaştıktan sonra bir ürün için öneri talep eder. Komşuluğa dayalı hafıza tabanlı *OF* sistemlerinde öneri üretme işlemi en genel haliyle iki adımlı bir süreçten meydana gelmektedir. Bunlar;

(i) öneri isteyen kullanıcı ile sistemdeki diğer kullanıcılar arasındaki benzerliklerin hesaplanması ve

(ii) benzer beğeni eğilimlerine sahip kullanıcıların geçmişteki derecelendirme tercihlerine dayalı olarak öneri istenen ürün için bir tahmin üretilmesidir.

2.1.1. *OF* sistemlerinde öneri üretme süreci

Geleneksel *OF* sistemlerinde kullanıcılara ait ürün beğeni değerleri sadece genel derecelendirme değeri olarak isimlendirilen tek bir ölçüt aracılığıyla hesaplanmaktadır. Geleneksel tek-ölçütlü bir *OF* sisteminde kullanıcı derecelendirmelerinin kaydedildiği temsili *kullanıcı* × *ürün* örnek matrisi Tablo 2.1'de gösterilmekte ve kullanıcılar $\{u_1, u_2, \dots, u_n\}$, ürünler $\{i_1, i_2, \dots, i_m\}$ olarak ifade edilmektedir. Bu tabloda a olarak örneklenen u_1 kullanıcısı q olarak i_5 ürünü hakkında *OF* sisteminden öneri üretilmesini istemektedir.

Tablo 2.1. *Tek-ölçütlü kullanıcı-ürün matrisi*

	i_1	i_2	i_3	i_4	i_5
u_1	5	7	5	7	**?**
u_2	5	7	5	7	8
u_3	5	7	5	7	8
u_4	6	6	6	6	5
u_5	6	6	6	6	5
⋮	⋮	⋮	⋮	⋮	⋮

Komşuluk tabanlı geleneksel *OF* yaklaşımlarında, sistem ilk olarak *a* kullanıcısı ile kendisine en benzer derecelendirme profiline sahip komşuları elde etmelidir. Kullanıcı benzerlikleri birçok farklı metodoloji ile elde edilebilirken *PK* katsayısı, kullanıcı benzerliklerini belirlemek için yaygın olarak kullanılan (Herlocker vd., 2004) ve öneri üretme doğruluğu bakımından literatürde kabul görmüş bir yöntemdir (Choi ve Suh, 2013). *PK* katsayısı ile *a* kullanıcısının diğer sistem kullanıcıları arasındaki benzerlik oranını elde etmek için Denklem 2.1 kullanılmaktadır.

$$PK(a,u) = \frac{\left(\sum_{i \in I}\left(r_{a,i} - \overline{r_a}\right) \times \left(r_{u,i} - \overline{r_u}\right)\right)}{\left(\sqrt{\sum_{i \in I}\left(r_{a,i} - \overline{r_a}\right)^2}\sqrt{\sum_{i \in I}\left(r_{u,i} - \overline{r_u}\right)^2}\right)} \quad \textbf{(2.1)}$$

Burada *I*, *a* ve *u* kullanıcıları arasında ortak olarak derecelendirilmiş ürün kümesini temsil ederken, $i \in I$ için *a* ve *u* kullanıcılarına ait derecelendirme değerleri $r_{a,i}$ ve $r_{u,i}$ ile ifade edilmektedir. *a* ve *u* kullanıcısının ortalama oy değerleri sırasıyla $\overline{r_a}$ ve $\overline{r_u}$ ile gösterilmektedir.

Denklem 2.1 aracılığıyla *a* kullanıcısının diğer sistem kullanıcıları arasındaki profil benzerlikleri (*PK(a,u)*) elde edildikten sonra, komşular benzerlik değerlerine göre sıralanıp en benzer *N* tane kullanıcı üzerinden öneri talebinde bulunulan q için öneri üretme işlemi Denklem 2.2 ile gerçekleştirilmektedir.

$$P_{a,q} = \frac{\sum_{u \in N}\left(r_{u,q} - \overline{r_u}\right) PK(a,u)}{\sum_{u \in N} PK(a,u)} \quad \textbf{(2.2)}$$

Burada *N*, *a*'nın en yakın komşuluk kümesini temsil etmektedir.

2.1.2. OF^k sistemlerinde öneri üretme süreci

Geleneksel *OF* sistemleri, tecrübe edilen ürünlerin her biri için tek bir tercih değerine ihtiyaç duysa da, kullanıcılara ait beğeni profillerini daha iyi kişiselleştirmek ve

analiz etmek için ortaya konan yeni yaklaşımlar, bir öğenin ayırt edici alt unsurları üzerinde çoklu-ölçütlü derecelendirmelerin toplanmasını teşvik etmektedir (Jin ve Si, 2004; Nilashi vd., 2015). Çoklu-ölçüte dayalı beğeni değerlerinin toplanması, yalnızca tercih edilen ürünlerin birbiriyle nasıl ilişkilendirildiğini değil, aynı zamanda kullanıcının bir ürünü neden tercih ettiği, sevdiği ya da sevmediği bilgisi ile birlikte kullanıcılar arasındaki gizli ilgileşimleri keşfetmeye de yardımcı olmaktadır. Bir OF^k sisteminde kullanıcı derecelendirmelerinin kaydedildiği temsili *kullanıcı* × *ürün* matrisi Tablo 2.2'de gösterilmektedir. Yahoo!Movies çevrimiçi hizmeti gibi çoklu-ölçütlü tercih değerlerinin kullanıldığı bir film sitesi örnek alındığında; veri setinde yer alan nümerik değerler bir sinema filmi için senaryo, oyunculuk, yönetmenlik ve görsel efektler gibi alt-ölçütler ile birlikte her bir ürün için genel beğeni derecesini temsil eden diğer dört alt-ölçütün aritmetik ortalaması olarak gösterilen genel beğeni ölçütünü içermektedir.

Tablo 2.2. *Çoklu-ölçütlü kullanıcı-ürün matrisi*

	i_1	i_2	i_3	i_4	i_5
u_1	$5_{2,2,8,8}$	$7_{5,5,9,9}$	$5_{2,2,8,8}$	$7_{5,5,9,9}$	**?**
u_2	$5_{8,8,2,2}$	$7_{9,9,5,5}$	$5_{8,8,2,2}$	$7_{9,9,5,5}$	$8_{9,9,7,7}$
u_3	$5_{8,8,2,2}$	$7_{9,9,5,5}$	$5_{8,8,2,2}$	$7_{9,9,5,5}$	$8_{8,8,8,8}$
u_4	$6_{3,3,9,9}$	$6_{4,4,8,8}$	$6_{3,3,9,9}$	$6_{4,4,8,8}$	$5_{6,4,6,4}$
u_5	$6_{3,3,9,9}$	$6_{4,4,8,8}$	$6_{3,3,9,9}$	$6_{4,4,8,8}$	$5_{5,5,5,5}$
⋮	⋮	⋮	⋮	⋮	⋮

Tablo 2.2'de a olarak belirlenen u_1 kullanıcısı q olarak i_5 ürünü için OF^k sisteminden öneri üretilmesini talep etmektedir. Bu veri setinde geleneksel *OF* sisteminde kullanılan yaklaşım referans alınırsa; u_1 kullanıcısına en yakın beğeni profiline sahip ve i_5 ürünü için oy vermiş kullanıcıları bularak öneri değeri tahmin edilmelidir. u_1 kullanıcısı genel beğeni değeri olarak u_2 ve u_3 kullanıcıları ile aynı değere sahip olsa da aslında alt-ölçütler bazında birbirlerinden çok farklı profillere sahiptir. Bu örnekte olduğu gibi, tek-ölçüte dayalı kullanıcı tercihleri, öneri üretme işleminde yetersiz kalabilir. u_4 ve u_5 kullanıcıları ele alındığında genel beğeni değerleri diğer iki kullanıcı kadar eşleşmese de alt-ölçütlere göre değerlendirildiğinde u_1 kullanıcısı ile daha benzer beğenilere sahip oldukları görülmektedir. Özetle, geleneksel tek-ölçütlü *OF* sistemi, bu kullanıcının i_5 ürünü için beğeni derecesini 8 olarak tahmin edecek iken, OF^k kullanımı ile ilgili ürün için üretilen tahmin 5 olacaktır. Örnekte görüldüğü üzere, genel derecelendirme perspektifinden bakıldığında benzer beğenilere sahip oldukları düşünülen

kullanıcıların, çoklu-ölçüt göz önünde bulundurulduğunda ilgili ürünleri farklı nedenlerle beğendikleri ortaya çıkmaktadır.

OF^k sistemleri, çoğu zaman genel bir derecelendirmeyle birlikte, değişen alt-ölçütler üzerinden detaylı kullanıcı derecelendirmelerine sahiptir. Sonuç olarak, iki kullanıcı arasındaki genel benzerlik değeri, alt-ölçütler ve genel ölçüt derecelendirmelerinden elde edilen benzerliklere dayanmaktadır. OF^k sistemlerinde öneri üretme sürecinde alt-ölçütler üzerinden genel beğeni ölçütünü elde etmek amacıyla Adomavicius ve Kwon (2007) benzerlik tabanlı ve birleştirme fonksiyonu tabanlı iki temel şema ortaya koymuştur.

Benzerlik tabanlı yaklaşımlarımda *a* ve diğer kullanıcılar arasındaki genel benzerlik değerini tahmin etmek için, alt-ölçütlerden elde edilen bireysel benzerlik değerleri iki farklı şekilde ele alınmıştır. Bunlar;

(i) Ortalama benzerlik ($\bar{\cdot}$) yaklaşımında, genel derecelendirme değerinin her bir alt-ölçüt ile eşit seviyede değerli olduğu varsayımından yola çıkılmakta ve alt-ölçütlere ait benzerlik katsayılarının aritmetik ortalaması Denklem 2.3'e göre elde edilerek nihai benzerlik katsayısı elde edilmektedir. Burada k sistemdeki toplam alt-ölçüt sayısını ifade ederken, $sim_i(a,u)$ fonksiyonu *a* kullanıcısının *i.* alt-ölçüt için *u* kullanıcısı ile benzerlik katsayısını ifade etmektedir.

$$\overline{sim(a,u)} = \frac{1}{k}\sum_{i=1}^{k} sim_i(a,u) \qquad \textbf{(2.3)}$$

(ii) En kötü benzerlik ($\lfloor\cdot\rfloor$) yaklaşımda; *a* kullanıcısı ile *u* kullanıcısı arasındaki benzerlik elde edilmek için Denklem 2.4'e göre *k* tane alt-ölçüt içerisindeki en kötü benzerlik katsayısı belirlenip nihai benzerlik değeri olarak kabul edilmektedir.

$$\lfloor sim(a,u) \rfloor = \min_{i=1,\dots,k}(sim_i(a,u)) \qquad \textbf{(2.4)}$$

Birleştirme fonksiyonu tabanlı yaklaşımlar sezgisel olarak genel ölçüt ile alt-ölçütler arasında bir ilişki olduğu varsayımına dayanmaktadır. Bu nedenle genel derecelendirme ölçütü (r_0), alt-ölçütlere bağımlı olarak $r_0 = f(r_1, \dots, r_k)$ fonksiyonu ile tanımlanır. Birleştirme fonksiyonu tabanlı yaklaşımlarda, belirli bir *a* kullanıcının *q* ürünü için r_0 ölçütüne öneri üretirken ilk olarak bu ölçütü ifade eden f fonksiyonu

belirlenmelidir. Bu fonksiyonu elde ederken r_0 ile diğer alt ölçütler $(r_1, \ldots, r_k)$ arasındaki gizli korelasyonu otomatik olarak belirleyebilmek için Bölüm 1.2'de ifade edildiği gibi makine öğrenmesi ya da istatistiksel yöntemlere sıklıkla başvurulmaktadır. Adomavicius ve Kwon (2007), doğrusal regresyon teknikleri kullanarak alt-ölçütlerin r_0 ile ilişkisini derecelendirip ve elde edilen değerleri her bir alt-ölçüte ağırlık olarak tayin edip genel derecelendirme değerini belirlemeyi önermiştir. Benzer diğer bir yaklaşımda (Jannach, Karakaya ve Gedikli, 2012) destek vektör regresyonu ve ağırlıklandırılmış destek vektör regresyonu ile alt-ölçütlerin r_0 ile ilişkisini derecelendirerek kullanıcı ve ürün tabanlı öneriler üretilmiştir.

2.2. Gizliliği Koruyan Ortak Filtreleme Sistemleri

Bireyler, çevrimiçi hizmetleri kullanırken genellikle mahremiyetlerinin ihlal edildiğini düşünmektedir (Yargıç ve Bilge, 2017). Bunun bir sonucu olarak kullanıcılar bazen çevrimiçi hizmetleri kullanmayı tamamıyla reddederken bazı durumlarda sisteme yanlış bilgi beyan ederek kişisel olarak mahremiyet kaygılarını hafifletmek isteyebilirler. Bunun sonucunda *OF* sistemi tarafından üretilen önerilerin doğruluğu olumsuz olarak etkilenir. Bununla birlikte, *GKOF* sistemleri tarafından sunulan gizlilik koruma protokolleri genellikle kullanıcı profilinde bozulmalara neden olur. Bunun sonucunda öneri doğruluğunda azalmaya neden olmaktadır. Bu nedenle, üretilen önerilerin doğruluğu ve kullanıcı mahremiyetini korumak için önerilen gizlilik koruma yaklaşımları etkili bir gizlilik koruma sağlarken aynı zamanda da öneri doğruluğundan minimum düzeyde ödün vermelidir.

GKOF sistemlerinde ortaya konulan gizlilik koruma şemalarının kullanıcı mahremiyetini sağlamak için iki temel hedefi vardır. Bunlar:

(i) kullanıcıların, ürünler için beğenilerini ifade etmekte kullandığı gerçel derecelendirme değerlerini gizlemek ve

(ii) oy verilen ürün listesini gizlemektir (Polat ve Du, 2005a).

Bu hedef doğrultusunda, kullanıcıların gerçek oy değerlerini maskelemek için *RK* yöntemi, kullanıcıların oy verdiği ürün listesini gizlemek için *RD* yöntemi geleneksel *GKOF* sistemlerinde yaygın olarak kullanılmaktadır (Bilge ve Polat, 2013; Bilge vd., 2013).

2.2.1. *RK* ve *RD* yöntemleri

RK yöntemi *GKOF* sistemlerinde yaygın olarak kullanılan, kullanıcının geçmişte tecrübe ettiği ürünlere ait derecelendirme listesi üzerine, tercih edilen prosedür doğrultusunda üretilen sayılardan oluşan maskeleme vektörünü ekleyerek gerçek derecelendirme değerlerinin gizlendiği veri maskeleme yöntemidir (Agrawal ve Srikant, 2000; Polat ve Du, 2005a; Bilge vd., 2013). *RK* yönteminin altında yatan ana fikir, istemci tarafında kullanıcının kendi derecelendirme listesini maskeleyerek sunucuya ham derecelendirme listesi yerine maskelenmiş veri listesini göndermesidir. Böylece, servis sağlayıcının gerçek kullanıcı derecelendirmelerine erişimi engellenirken aynı zamanda da maskelenmiş veri üzerinden servis sağlayıcının öneri üretebilmesine olanak sağlamaktadır (Polat ve Du, 2003; Polat ve Du 2005a).

RK yöntemi kullanıcıya ait Gerçek Oy Değerleri (G) vektörünü maskelemek için G vektörü üzerine ortalaması sıfıra eşit olacak şekilde rastgele üretilmiş Rastgele Sayı (R) vektörü eklenmektedir. Böylelikle, orijinal kullanıcı oy değerlerini $G + R$ ile değiştirerek G vektöründe bulunun gerçek kullanıcı derecelendirmelerinin gizlemeyi hedeflemektedir.

RK prosedürü ile oluşturulan maskeleme vektörü, servis sağlayıcı tarafından belirlenen ve en yüksek gizlilik seviyesini ifade eden σ_{max} katsayısı aracılığıyla kullanıcının ihtiyaç duyduğu gizlilik seviyesine göre üretilmektedir. Kullanıcı gizlilik seviyesini belirlemek için kullanılan σ katsayısı $(0, \sigma_{max}]$ değer aralığı içerisinden kullanıcı tarafından belirlenmektedir. Veri maskeleme işleminde kullanılacak R vektörü Üniform ($\mathcal{U}$) ya da Normal ($\mathcal{N}$) dağılım gibi farklı dağılımlar kullanılarak üretilebilmektedir (Bilge vd., 2013). $\mathcal{U}$ dağılıma göre üretilen rastgele sayı değerleri $[-\sqrt{3}\sigma, \sqrt{3}\sigma]$ değer aralığında üretilmektedir. $\mathcal{N}$ dağılıma göre üretilen rassal sayılar, ortalaması (μ) sıfıra eşit olacak şekilde kullanıcı ürün derecelendirme vektörünün standart sapmasına (σ_u) bağlı olarak üretilmektedir.

Gizliliği koruma politikalarının ikinci temel amacı; kullanıcıların gerçek oy değerlerini gizlemenin yanı sıra, kullanıcıların oy verdiği ürün listesinin gizlenmesidir. Kullanıcı derecelendirme vektörü üzerine eklenecek sahte doldurma verisinin miktarı, servis sağlayıcı tarafından belirlenen ve en yüksek gizlilik seviyesini ifade eden β_{max} katsayısı aracılığıyla kullanıcının ihtiyaç duyduğu gizlilik seviyesine göre üretilmektedir. Bu amaçla uygulanan *RD* yönteminde kullanıcıdan, ihtiyaç duyulan gizlilik düzeyine göre β katsayısını belirlemesi istenmektedir. β katsayısı kullanılarak kullanıcının

derecelendirilmemiş ürün listesinden rastgele olarak seçilmiş $\%\beta$ kadarı belirlenip seçilen derecelendirilmemiş boş hücrelere *RK* yöntemi ile üretilen sayı değerlerinin eklenir. Böylece kullanıcın gerçekte oy verdiği ürün listesi gizlenmektedir. *RK* ve *RD* işlemlerinin geleneksel *OF* sistemleri için temel adımları Prosedür 2.1'de ifade edilmektedir.

Prosedür 2.1. *GKOF sistemlerinde veri maskeleme prosedürü* (Bilge ve Polat, 2013)

Require: Kullanıcı × ürün vektörü (G), σ_{max}, β_{max}

z-skoru değerini hesapla $(\rightarrow \boldsymbol{Z})$

1 : $\bar{G} \leftarrow MEAN(G)$

2 : $\sigma_G \leftarrow STD(G)$

3 : **for** all items in G $(i \leftarrow 1\ to\ m)$ **do**

4 : $\quad Z_i = (G_i - \bar{G}_i)/\sigma_{G_i}$

5 : **end for**

Gizlilik parametrelerinin belirlenmesi

6 : $\beta \leftarrow (0, \beta_{max}]$

7 : $\sigma \leftarrow (0, \sigma_{max}]$

8 : $\alpha \leftarrow \sqrt{3}\sigma$

9 : $e \leftarrow |E|$ ► # oy verilmemiş ürün

10 : $g \leftarrow |G|$ ► # gerçek kullanıcı oyları

11 : $F \leftarrow e \times \beta\%$ ► # doldurulacak hücre sayısı

Dağılımı belirle ve rastgele sayı türet

12 : $dist \leftarrow RANDOM\ (uniform,\ normal)$

13 : $R \leftarrow dist(g + F;\ \mu = 0, \sigma|\ \alpha)$

z-skoru değerlerini maskele $(\rightarrow \boldsymbol{Z'})$:

14 : **for** all items in G $(i \leftarrow 1\ to\ m)$ **do**

15 : $\quad Z'_i = (Z_i + R_i)$

16 : **end for**

17 : **return** $\boldsymbol{Z'}$

Prosedür 2.1'de temel adımları gösterilen *GKOF* sistemleri için veri gizleme sürecinde yapılan işlemler aşağıdaki gibi özetlenebilir:

(i) *Standardizasyon işlemi (z-*skoru *normalleştirme):* Veri seti içerisindeki kullanıcı oylarının, kullanıcı profiline aykırı olarak olağan dışı derecelendirme değerlerine sahip olması sistem tarafından üretilen önerilerin kalitesinde olumsuz bir etkiye sahiptir (Bilge ve Yargıç, 2017). Ayrıca ham kullanıcı verileri ile üretilen öneriler ve normalleştirme sürecinden geçirilen kullanıcı verileri ile üretilen önerilerde doğruluk bakımından önemli farklar bulunmaktadır. Gerçek kullanıcı derecelendirmeleri de z-skoru normalleştirme işlemine tabi tutularak ortalaması sıfıra eşitlenir. Buna ek olarak, veri maskeleme işleminde kullanılacak maskeleme verisi ortalaması sıfıra eşit olacak şekilde üretilmektedir. Böylece hem kullanıcı hem de maskeleme vektörünün ortalaması sıfıra eşitlenerek her kullanıcı için aynı ortalamaya sahip maskeleme vektörleri üretilir. Böylece, maskeleme işlemi ile ortaya çıkabilecek sapmaların ortadan kaldırılması sağlanır. Bu amaçla Denklem 2.5'te gösterilen z-skoru normalleştirme tekniği kullanılmıştır, böylece kullanıcı derecelendirmeleri normalleştirilmiş ve kullanıcı derecelendirmeleri arasındaki olağan dışı oylara ait sapmalar ortadan kaldırılarak bu derecelendirmelerin etkisi azaltılmıştır (Herlocker vd., 1999).

$$z_{ui} = \frac{r_{ui} - \overline{r_u}}{\sigma_u} \quad \textbf{(2.5)}$$

Burada, u kullanıcısının i ürünü için orijinal derecelendirme değeri r_{ui} olarak ifade edilmiştir. $\overline{r_u}$ ve σ_u girdi setinin yani u kullanıcısına ait gerçek derecelendirme değerlerinin sırasıyla ortalamasını ve standart sapmasını ifade etmektedir.

(ii) *Kişisel gizlilik parametrelerinin belirlenmesi:* Kullanıcının sunucu tarafından belirlenen σ_{max}ve β_{max} katsayılarına göre, ihtiyaç duyduğu gizlilik seviyesine göre kendisine ait σ değerini $(0, \sigma_{max}]$, β değerini $(0, \beta_{max}]$ değer aralığı içerisinden belirlemesi.

(iii) *Rastgele sayıların üretilmesi:* Belirlenen rastgele sayı üretme aralığı ve boş hücre oranına göre $\mathcal{N}$ ya da $\mathcal{U}$ dağılıma göre rastgele sayı üretme işlemi sonucunda R vektörünün elde edilmesi.

(iv) *Maskelenmiş derecelendirme vektörünün oluşturulması:* Kullanıcı, oy verdiği ürünlere ait tercih vektörünü sunucuya göndermeden önce *z*-skoru normalleştirme işlemine tabi tutulan gerçek oylara ait *G* vektörü üzerine *R* vektörünü ekleme işlemi.

2.3. Veri Setleri ve Değerlendirme Ölçütleri

Bu bölümde, literatürde OF^k sistemleri üzerinde yapılan çalışmalarda yaygın olarak kullanılan Yahoo!Movies web-servisinden derlenen (YM) veri seti kullanılmaktadır. YM veri seti, çoklu-ölçütlü kullanıcı derecelendirmelerine dayalı olarak genel beğeni ölçütü ile birlikte; senaryo, oyunculuk, yönetmenlik ve görsel efektler olmak üzere dört alt-ölçüte ait derecelendirmelerden meydana gelmektedir. Bu beş ölçüte ait kaydedilen derecelendirmeler, *F* en düşük ve *A+* en yüksek skor olmak üzere, 13 derecelendirme değerinden oluşmaktadır [A+, A, A-, B+, B, B-, C+, C, C–, D+, D, D-, F]. Beğeni dereceleri üzerinde işlem yapabilmek için bu değerler 1 *F*'yi ve 13 *A+*'yı temsil edecek şekilde sayısal derecelendirmelere dönüştürülerek kullanılmıştır.

Tablo 2.3. *Veri setlerinin özellikleri*

	YM5	*YM10*	*YM20*
Kullanıcı Sayısı	4,377	1,293	202
Ürün Sayısı	2,565	1,164	247
Derecelendirme Sayısı	63,027	34,846	8,157
Seyreklik Oranı	%99,44	%97,69	%83,65

Ham haliyle YM veri seti, mevcut öğelerinin yalnızca %0,02'sinin derecelendirildiği son derece seyrek bir veri setidir. Bu nedenle yapılan çalışmada, YM veri setinin farklı seyreklik oranlarına sahip *YM5*, *YM10* ve *YM20* olarak isimlendirilen üç farklı alt kümesi kullanılmıştır (Adomavicius ve Kwon, 2007; Jannach vd., 2012). Veri setini isimlendirmek için kullanılan sayısal değerler, veri setlerindeki ürünler ve kullanıcılar için minimum miktardaki oy sayısını ifade etmektedir. Örneğin *YM5* veri setinde, her kullanıcı en az 5 ürünü derecelendirmiştir ve veri setindeki her ürün en az 5 farklı kullanıcı tarafından derecelendirilmiştir. Benzer yaklaşım *YM10* ve *YM20*'de de izlenerek farklı boyutlarda ve seyreklik seviyelerinde veri setleri elde edilmiştir. Elde edilen veri setlerinin sahip oldukları derecelendirme, kullanıcı ve ürün sayıları ile birlikte seyreklik seviyeleri Tablo 2.3'te verilmiştir.

Bu bölümde, *OF* sistemi tarafından üretilen önerilerin kalitesini ölçeklendirmek için literatürde yaygın olarak kullanılan yöntemlerden biri olan Ortalama Mutlak Hata (*MAE*) kullanılmaktadır (Cantador vd., 2015). Bu metrik gerçek kullanıcı derecelendirme değerleri ile *OF* sistemi tarafından üretilen öneriler arasındaki sayısal farkın negatif ya da pozitif olmasını göz ardı ederek toplam hatanın ortalama büyüklüğünü Denklem 2.6 aracılığı ile hesaplamaktadır.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|g_i - p_i| \quad \textbf{(2.6)}$$

Burada *g* gerçek kullanıcı oyunu, *p* üretilen tahmin değerini, *n* ise toplamda üretilen öneri sayısını ifade etmektedir. Önerilen tahminlerde ortaya çıkan her bir fark toplam hataya eşit miktarda etki ettiği için mutlak hataların ortalama değeri nihai hata değeri olarak ele alınmıştır.

3. ÇOKLU-ÖLÇÜTLÜ ORTAK FİLTRELEME SİSTEMLERİNDE GİZLİLİK RİSKLERİ

Bireyler çevrimiçi hizmetleri kullanırken mahremiyetlerinin ihlal edildiğini düşünmektedir. Bunun bir sonucu olarak kullanıcılar bazen sisteme yanlış bilgi beyan ederlerken bazen de bu hizmetleri kullanmayı tamamıyla reddetme eğiliminde olabilirler. Kullanıcıların sistem hakkındaki mahremiyet kaygılarını hafifletmek için bu tür sistemlerin neden olabileceği mahremiyet ihlalleri ayrıntılı bir şekilde analiz edilmeli ve servis sağlayıcı tarafından mahremiyeti koruma mekanizmaları geliştirilmelidir. Bu bağlamda ortaya konan koruma mekanizmaları kullanıcı mahremiyetini korurken, sistemin ürettiği öneri doğruluğundan ödün verilmemesine özen göstermelidir. *OF* sistemlerinde kişisel tercihlerin toplanması ve kayıt altında tutulması nedeniyle kullanıcıların maruz kaldıkları gizlilik riskleri literatürde tartışılmaktadır. Ancak, bu tür çalışmalar, geleneksel tek-ölçütlü tercih değeri kullanan *OF* sistemlerinin maruz kaldığı tehditler üzerinde durmakta ve çoklu-ölçütlü tercih verisi alanında ortaya çıkabilecek gizlilik risklerinin değerlendirilmesinde yetersiz kalmaktadır.

3.1. Giriş

OF sistemlerinde üretilen önerilerin doğruluğu, kullanılan tercih verilerinin kalitesi ile yakından ilişkili olduğu tartışılmaz bir gerçektir. Çoklu-ölçütlü derecelendirmelerin kullanımı ile birlikte, daha kişiselleştirilmiş kullanıcı profilleri oluşturmak mümkün hale gelmektedir ve sonucunda kişiselleştirilmiş önerilerin kalitesini arttırmaktadır. Ancak, çoklu-ölçütlü tercih verisi kullanımı ile bireyler daha ciddi mahremiyet riskleri ile karşı karşıya kalabilmektedir. Bu nedenle, bireyleri çoklu-ölçütlü öneri sistemlerinin kullanımına yönlendirmek, sağlanan tahminlerin doğruluğu ile kullanıcı mahremiyeti arasında bir denge kurmaya dayanmaktadır. Bu bölümde, mevcut gizlilik ihlalleri çoklu-ölçütlü *OF* sistemleri perspektifinden değerlendirilip, bu tür hizmetlerin maruz kaldığı tehditler tartışılmaktadır.

Bireylerin bilgi gizliliği konusundaki endişeleri, öneri hizmeti sunan e-ticaret sistemlerinden faydalanmalarının önüne geçen önemli bir faktördür. Dijital formattaki kişisel bilgiler, öneri sistemleri tarafından kişiselleştirilmiş öneriler sunmak için kullanılmaktadır. Fakat toplanma amacının yanı sıra bu bilgiler; çalınma, kopyalanma, değiştirilme ve istismar edilme gibi ciddi risklere neden olabilirler. Bireyler kişisel bilgilerini dijital ortamda paylaşırken mevcut risklerin farkında olmalı ve getirdiği

faydanın yanı sıra, bir risk teşkil edebileceğini de unutmamalıdır. Bu tür hizmetleri kullanmaktan kaçınan bireylerin mahremiyet kaygılarını gidermek için, mahremiyet kavramı genel olarak ele alınmalı ve bilgi gizliliği konusu kapsamlı olarak incelenmelidir.

Öneri sistemleri tarafından üretilen tahminlerin doğruluğu ile sistem kullanıcılarının ürün tercihleri için kullandığı oylama verisinin kalitesi yakından ilişkilidir. Geleneksel öneri sistemleri tarafından kullanılan tek-ölçüte nazaran çoklu-ölçüt kullanımı; dinamik kullanıcı profilleri oluşturmak için daha ayrıntılı bir içerik sunar. Böylece daha yüksek seviyede kişiselleştirilmiş profiller oluşturarak daha kaliteli öneriler üretmeye yardımcı olmaktadır. Bununla birlikte, kullanılan çoklu-ölçüt yapısı kullanıcıları daha ciddi mahremiyet sorunları ile karşı karşıya getirmektedir. Bu nedenle, bireylerin çoklu-ölçütlü öneri sistemlerini kullanmalarını sağlamak, aslında sistem tarafından üretilen önerinin doğruluğu ile sistem tarafından sağlanan veri gizliliği arasında denge kurmaya dayanmaktadır.

3.2. Gizlilik Kavramının Genel Tanımı

Bilgi gizliliğiyle ilgili kaygılar, bireylerin e-ticaret servislerinde *OF* sistemlerinin sağladığı hizmetlerden yararlanmasını engelleyen önemli bir faktördür. Dijital formdaki kişisel bilgiler, bu tür sistemlerin kullanıcıları için kişiselleştirilmiş öneriler sağlamak amacıyla kullanılmaktadır. Ancak aynı zamanda, toplanma amacından farklı olarak kopyalanması, çalınması, değiştirilmesi ve istismar edilmesi gibi ciddi risklere de tabidir. Bu bağlamda, bireyler ve hizmet sağlayıcılar kişisel bilgilerin paylaşılmasının sonuçlarından haberdar olmalı ve bu bilgiler her iki taraf için ayrı ayrı ele alınmalıdır (Malhotra, Kim ve Agarwal, 2004). Bu hizmetleri kullanmaktan kaçınan kişilerin gizlilik endişelerini gidermek için, genel gizlilik kavramı bağlamında bilgi gizliliğini kapsamlı bir şekilde analiz etmek çok önemlidir.

Gizlilik kavramı, tam olarak anlamak ve tanımlamak için hala zor bir kavramdır. Kavramsal karmaşıklığın temel nedeni, mevcut gizlilik anlayışının büyük ölçüde farklı disiplinlere bağlı olmasıdır. Bu disiplinlerdeki çalışmalar, mahremiyet konusunda farklı öncelikler atfetmektedir. Gizlilik kavramı hemen hemen bütün sosyal bilim alanlarında keşfedilen bir konudur; özellikle hukuk (Parent, 1983), ekonomi (Waldo, Lin ve Lynette, 2010), psikoloji (Schoeman, 1984; Parker, 1973), pazarlama (Goodwin, 1991) ve bilgi yönetimi (Westin, 1968; Culnan, 1993; Dinev vd., 2013) alanlarında tanımlanmıştır. Gizlilik kavramı; hukuk disiplininde bir "hak" olarak tanımlanırken (Warren ve Brandeis,

1890), felsefe ve psikoloji disiplinlerinde "sınırlı erişim" veya "tecrit" halidir (Schoeman, 1984), sosyal bilimler ve bilgi sistemlerinde "kontrol" paradigmasıdır (Westin, 1968; Culnan, 1993). Disipline bağlı tanımlardaki farklılık gizlik kavramını anlamayı zorlaştırmaktadır. Bilgi gizliliği, bireylerin veya hizmet sağlayıcıların gizli bilgilerini ne zaman, nasıl, hangi bağlamda ve ne ölçüde kullanabileceği kapsamında gizlilik konusu ile ilgilidir (Westin, 1968). Bilgi gizliliği kavramsal olarak kendini açıklayıcı olsa da bunun gerçek yaşam sınırları endüstriyel sektörler, kültürler ve düzenleyici yasalar bağlamında farklılık göstermektedir (Malhotra vd., 2004). Bilgi gizliliği ile ilgili endişeler ise bireylerin gelenekleri, ruhsal eğilimleri ve zihniyetiyle yakından ilişkilidir. Gizlilik olgusu ile ilgili verilen temel örnekte, "bir kapıyı çalmadan açmanın farklı toplumlarda bir gizlilik ihlali olarak algılandığı" açıklanmaktadır (Moore, 2008). Ayrıca, gizlilik endişeleri, bireyin yaşı, cinsiyeti ve içinde bulundukları ortam ile ilişkilidir ve onları farklı davranışlar sergilemeye yönlendirir (Ackerman, 1999; Spiekermann, Grossklags ve Berendt, 2001).

3.3. Ortak Filtreleme Sistemleri Açısından Gizliliğin Tanımı

OF sistemleri kullanıcılarına kişiselleştirilmiş tahminler oluşturmak için; kullanıcı oyları, demografik veriler ve kullanıcıların davranışsal geçmişi (örneğin; ziyaret edilen sayfalar, sayfa tıklanma sayısı ve bir sayfada geçirilen zaman vb.) gibi tercihlere ilişkin ayrıntılı kişisel verileri izler ve depolar. Bireyler için mahremiyet riski taşıyan bu bilgileri doğrudan ya da dolaylı olarak elde eden *OF* sistemlerinin bu bilgileri amacı dışında kullanılmayacağının veya sızdırılmayacağının garantisi yoktur. Kişisel bilgiler bir kez dijital ortamda paylaşıldıktan sonra, kim tarafından veya hangi amaçla kullanıldığını kontrol etmek neredeyse imkânsız hale gelir. Maalesef bazı kullanıcılar, kişisel bilgilerin toplanmasından doğabilecek gizlilik ihlalleri konusunda bilgisizdir. Gross ve Acquisti (2005), bazı bireylerin kişisel bilgilerini çevrimiçi hizmetlerle paylaşırken varsayılan gizlilik ayarlarını bile değiştirmediklerini ortaya koymaktadır. Bununla birlikte, bireylerin birçoğu mahremiyetin kendileri için önemli olduğunu ve kişisel bilgilerinin ifşa edilmesiyle ilgili endişelerini dile getirmiştir (Spiekermann vd., 2001; Golbeck, 2016). Bireyler her ne kadar bu tür hizmetler konusunda çoğunlukla pragmatik davransalar da, bu kişiler öncelikli olarak öneri sistemlerini kullanmaya ve gizlilik endişeleri nedeniyle veri paylaşmaya isteksizdirler. Ancak, bireyler bu tür hizmetleri kullanmaya başladıklarında, geçmiş önyargılarını hızla terk ederler (Tüfekçi, 2008). Bireylerin bu tür

kaygılarını gidermek için, *OF* sistemleri, hangi tür verilerin toplanacağı, hangi amaçlarla kullanılacağı, kiminle paylaşılacağı ve ne kadar süre saklanacağı hakkında bilgi işleme politikaları açıklamalıdır. Hizmet sağlayıcılar kişisel bilgilerin kullanımıyla ilgili gizlilik bildirimlerini yayınlasalar da, denge yine de hizmet sağlayıcının lehine değişmektedir. Kişisel Verilerin Korunmasına ilişkin olarak Ekonomik İşbirliği ve Kalkınma Örgütü (OECD) rehberi, veri toplayıcılarının gizlilik ihlallerini hafifletmek için adil bilgi uygulamalarının temel ilkelerini tanımlamıştır. Bu ilkeler, herhangi bir veri toplayıcısının gizlilik ihlaline yol açabileceği veri toplama ve kullanım sınırlama, veri kalitesi, amaç belirtimi, güvenlik önlemleri, açıklık, bireysel katılım ve hesap verebilirlik uygulamalarını düzenlemektedir (Friedman vd., 2015).

OF sistemleri açısından gizli kabul edilen veriler; bireylerin tecrübe ettikleri ürünlere verdikleri özel tercih değerleri ve derecelendirilmiş ürün listesidir (Polat ve Du, 2005a; Polat ve Du, 2005b). Bir kullanıcının ürün beğenilerine ait gerçek düşünceleri ile satın alınan ya da beğenilen ürün kümesinin açığa çıkması, kullanıcının kişisel hayatı ve eğilimleri hakkında bazı çıkarımlar yapabilme olasılığını ortaya çıkarmaktadır.

Genel olarak, *OF* sistemlerinde bir bireyin özel tercih bilgilerine diğer sistem kullanıcılarının veya kötü niyetli kullanıcıların erişemeyeceği varsayılsa bile, bazı özel verileri elde etmek için o veriye doğrudan erişim gerekli değildir. Bir bireyin kişisel bilgileri, kötü niyetli bir kullanıcının sahip olduğu diğer veriler kullanılarak elde edilebilir. Örneğin, dinlenen müzik veya izlenilen film, ürüne ilgi duyan bireyin yaşı ve cinsiyeti ile ilişkilidir (Chaabane, Acs ve Kaafar, 2012; Shyong, Frankowski ve Riedl, 2006). Bir bireye ilişkin yaş ve cinsiyet bilgilerinin açıklanması, mahremiyetin önemli bir ihlali olarak değerlendirilemeyebilir. Yine de, bu tür bir ihlal yoluyla ek bilgiler elde edilebilir. Shyong, Frankowski ve Riedl (2006) bir bireyin kimliğinin sadece posta kodu, yaş ve cinsiyet bilgisi kullanılarak %87 oranında doğrulukla belirlenebileceğini ifade etmiştir. Westin ve Maurici'ye (1998) göre, kullanıcıların demografik farklılıkları çevrimiçi gizlilik konusundaki tutumlarını değiştirmektedir. Bir kullanıcıya ait cinsiyet bilgisi bireyler açısından özellikle kadın kullanıcılar için gizli olarak nitelendirilebilecek bir bilgidir. Tek-ölçütlü geleneksel *OF* sistemlerinde bile kullanıcı derecelendirmelerine dayalı olarak bu bilgiyi ortaya koyma yeteneği, kadın internet kullanıcılarının büyük bir kısmının mahremiyetleri konusunda daha fazla endişe duymasına neden olmaktadır (Mekovec ve Vrcek, 2011, Weinsberg vd., 2012). Chaabane, Acs ve Kaafar (2012), bir kullanıcının tercih ettiği müzik türünün veya izlediği filmin, o kullanıcının yaş ve

cinsiyeti ile ilişki olduğunu ifade etmiştir. Bu nedenle, kullanıcılara ait bilgi verebilecek kritik ürünlere yönelik derecelendirmeleri değerlendirerek, bir kullanıcının cinsiyet bilgisini ve yaş aralığını bulmak mümkündür.

3.4. Çoklu- Ölçütlü Ortak Filtreleme Sistemlerinde Gizlilik

OF sistemlerinde bireylerin demografik bilgileri, yaşam tarzı özellikleri, alışveriş alışkanlıkları, finansal durumu, yaş, cinsiyet, isim, adres ve sosyal güvenlik numarası gibi kişisel tanımlayıcılar mahrem bilgiler olarak tanımlanmaktadır (Phelps vd., 2000). Bir bireyin mahrem bilgilerinin, belirli bir kullanıcının *OF* sistemine gönderdiği derecelendirmeleri kullanarak çıkarılabileceğine dair bazı kanıtlar vardır (Jeckmans vd., 2013). Dolayısıyla *OF* sistemlerinin, gizli verilerin nasıl ve ne ölçüde kullanıldığına karar verme hakkı olarak tanımlanan bilgi gizliliğini ihlal etmeye eğilimli oldukları sonucuna varılabilir. Gizlilik ihlalleri literatürde birçok farklı yaklaşımla tartışılmaktadır. Cranor (2004) kişisel tercih paylaşma konusunda olası gizlilik tehlikelerini istenmeyen pazarlama, fiyat ayrımcılığı ve hesaba yetkisiz erişim gibi çeşitli kategorilerde sınıflandırmaktadır. Bu ihlalleri daha geniş bir grupta gözden geçirebilmek için, Friedman vd. (2015) gizlilik ihlallerini iki geniş kategoride sınıflandırır. Bunlar (i) özel kullanıcı verilerine doğrudan erişerek ve (ii) mevcut bilgiden yeni bilgilerin türetilmesiyle mahremiyet ihlalidir.

3.4.1. Kullanıcı verisine doğrudan erişim

Kullanıcı tercihi verilerine doğrudan erişim ciddi gizlilik ihlallerine neden olabilir. Bu tehditler, istenmeyen veri toplama, üçüncü şahıslarla veri paylaşımı ve çalışanlar tarafından izinsiz şekilde verilere erişim olarak üç temel alt başlıkta gruplandırılmıştır (Friedman vd., 2015; Cranor, 2004). Geleneksel *OF* sistemlerinde, kullanıcının tercih değerlerine bağlı olarak bir kullanıcı hakkında yeni bilgi çıkarımları yapılabilir. Ayrıca, e-ticaret servisleri, hizmet kalitesini iyileştirmek için sistem üzerinde gerçekleştirilen işlemler hakkında bağlamsal bilgileri takip etme eğilimindedir. Bu tür istenmeyen veri toplama işlemleri, kullanıcıların gizlilik endişelerini artırmaktadır. Çünkü içerik izleme ve kullanıcının tercih değerleri ile elde edilen verileri birleştirerek gizli tutulması gereken bilgileri elde etmek mümkündür. E-ticaret şirketlerinin, kullanıcı verilerini finansal kazanımlar amacıyla üçüncü şahıslarla paylaştığı ve hatta bunları sattığı bilinen bir gerçektir (Friedman vd., 2015). Bu şirketler, sistem kalitesini iyileştirmek için veri

madenciliği, izinsiz giriş tespiti veya istatistiksel raporlama gerçekleştirmek için kullanıcı verilerini periyodik olarak uzmanlarla veya üçüncü taraf hizmetleriyle paylaşmaktadır. Ayrıca Netflix, rekabete yönelik olarak araştırma toplulukları ile kullanıcı derecelendirme verilerini paylaşmaktadır. Anonim olmalarına rağmen, bu veriler gerçek kullanıcılarla eşleştirilebilir. Son olarak, hizmet sağlayıcılar veya e-ticaret şirketleri finansal zorluklarla karşılaştıklarında, değerli bir varlık olan kullanıcı tercih verilerini satarak gelir elde etmeye çalışabilirler. Öneri sistemleri, kullanıcıların bilgilerini çalışanların istenmeyen erişiminden korumak için birçok önlem almasına rağmen, bu bilgilerin finansal getirisi düşünüldüğünde alınan önlemler yetersiz kalabilir.

3.4.2. Kullanıcı verisine dolaylı erişim

OF sistemlerinde gizlilik ihlalleri sadece kullanıcı verilerine direkt olarak erişim sağlamakla ortaya çıkmaz, aynı zamanda mevcut bilgilerin işlenmesiyle yeni bilgilerin türetilmesi de bu ihlalleri meydana getirebilir. *OF* sistemleri tarafından toplanan veriler profil kişiselleştirme sürecinin bir sonucu olarak bireyin rızası olmadan onun yaşı, cinsiyeti ve etnik kökeni gibi demografik bilgilerin ifşa edilmesinde de kullanılabilir. Bunun sonucunda elde edilen bilgiler kullanılarak bireylere rızası dışında reklam göstermek ya da ürünlerde kişiye özel fiyat ayrımcılığı yapmak için kullanılabilir (Weinsberg vd., 2012). Pratikte, bireyler bu tür bilgileri gizlilik endişeleri nedeniyle çevrimiçi profillerinde ifşa etmekten kaçınırlar. Ancak, mevcut araştırmalar, bir *OF* sistemindeki kullanıcı derecelendirme geçmişini ele geçirerek kullanıcıların özel bilgilerinin kolayca elde edilebileceğini göstermektedir (Jeckmans vd., 2013). Özel ve hassas kullanıcı bilgilerinin çıkarımlarını yapmak için kötü niyetli girişimler genellikle diğer kullanıcıların verileriyle çeşitli şekillerde ilişki arar. Weinsberg vd. (2012) kullanıcıların cinsiyetlerini yüksek doğruluk düzeyine sahip olarak belirlemek için MovieLens ve Flixster veri kümelerinde çeşitli sınıflandırıcılar kullanmaktadır. Ramakrishnan vd. (2001) eklektik beğenilere sahip bireylerin gizlilik tehditlerine kolayca maruz kalacağını göstermek için ortaya koyduğu yaklaşımda istatistiksel veri tabanı sorguları kullanmaktadır.

Kullanıcı derecelendirme geçmişlerine ek olarak, kullanıcılar tarafından ortak olarak derecelendirilen ürünlere ait listeler "pasif gizlilik saldırıları" olarak adlandırılan gizlilik ihlallerine ilişkin önemli riskler taşımaktadır. Gerçek hayatta, Amazon.com gibi e-ticaret servis sağlayıcıları, kullanıcılar için ilgili öğe listelerini herkese açık olarak

yayınlamaktadır. Kötü niyetli bir kullanıcı bu listeyi kullanarak, hedef kullanıcının işlem geçmişinde belirlenen hedef öğenin bulunup bulunmadığını öğrenebilir. Bu işlem için, kötü niyetli bir kullanıcının herhangi bir *OF* sistemine abone olması ya da sistem veri tabanına erişmesi gerekmez (Chen vd., 2014). Ancak hedef ürünlerin işlemlerinde ilgili ürün listesini kullanarak belirlenen bir öğenin bulunup bulunmadığını öğrenebilir. Belirli bir kullanıcı tarafından oylanan bir ürün listesini elde eden saldırgan, bireyin gelir düzeyi ve yaşam tarzı gibi daha kişisel bilgiler hakkında gerçekçi bir tahmin yapabilir.

OF^k sistemleri öneri üretme sürecinde yapısı gereği genel beğeni ölçütü ile birlikte birden fazla alt-ölçüt kullanır. Bu ölçütlerden kullanıcının genel beğeni ölçütüne verdiği oy değeri ile alt-ölçütler arasında belirli bir uyum olması beklenmektedir. Ölçüt derecelendirme değerleri arasında böyle bir uyumun olmadığı, olağan dışı yüksek ya da düşük derecelendirme değeri verildiği durumlarda bireyler oldukça ciddi mahremiyet riskleri ile karşı karşıya gelebilmektedir. Bu riskler özellikle çıkarım yolu ile elde edilebilecek yeni bilgilerin elde edilmesine olanak sağlamaktadır. Bu gibi durumlar hem geleneksel *OF* sistemlerinde mevcut olan tehditlerin ortaya çıkma olasılığını arttırırken hem de yeni tehditler için risk teşkil etmektedir. Örneğin siyasi içerikli bir film, tek-ölçütlü *OF* sistemde yüksek bir puanla değerlendirildiğinde elde edilebilecek bilgiler kullanıcının bu filmi izlediği ve beğendiği bilgisidir. Ancak bu filmi hangi özelliğinden dolayı yüksek puanla derecelendirildiği bilgisi elde edilemez. Aynı film örneğini OF^k sistemi üzerinde yeniden ele alıp, sistemin kullanıcı beğeni değerlerini elde etmek için oyunculuk, görseller, yönetmenlik ve senaryo alt-ölçütleri ile birlikte genel bir tercih değeri toplayan çoklu-ölçütlü film öneri sistemine sunduğu derecelendirmeleri inceleyelim. OF^k sistemleri düşünüldüğünde bu film, senaryo alanında yüksek bir puan ile diğer alanlarda ise nispeten düşük puanlarla derecelendirilirse, bireyin izlediği diğer filmler de göz önünde bulundurularak politik görüşü hakkında bir çıkarım yapılması söz konusu olabilmektedir. Bu yaklaşıma benzer bir yaklaşımla, farklı *OF* sistemlerinde de alt-ölçütler arasındaki korelasyon ve uyum analiz edilerek bir kitabın konusu veya bir restoranın lokasyonu gibi ürüne özgü özellikler ve alt-ölçütler aracılığı ile bireye ait gizli bilgilerin çıkarılması mümkün hale gelebilir.

OF sistemlerine kıyasla, tek bir kişiye özgü oy verme ve ürün satın alma eğilimine sahip profillerdeki kullanıcılar OF^k sistemlerinde daha büyük risk teşkil etmektedirler. Buna bir örnek vermek gerekirse, bir kullanıcı bir kitap satış sitesinden İtalyan yemeklerine ait bir kitap ile birlikte ağ güvenliği kitabı satın aldığı bir senaryoda, bu satın

alma profiline sahip, sistemde az sayıda kullanıcı olabileceği için sistem veri setine sahip olan üçüncü şahıslar basit istatistiksel sorgular yardımı ile bu kullanıcının kimliğini ifşa edebilirler. Bu örnekteki satın alma profili kullanılarak maruz kalınacak gizlilik ihlalleri geleneksel *OF* sistemlerinde de mevcuttur. Ancak aynı örneği çoklu-ölçüt bakış açısı ile yeniden ele aldığımızda kitap satış sitesinin kullandığı alt-ölçütler aracılığı ile eklektik beğeniye sahip kullanıcının kimlik bilgisi yanı sıra alt-ölçütler içerisinde yapılacak çıkarımlarla birlikte bireyin kitabı sevme sebebinden yola çıkarak onun etnik kökeni, mesleği gibi daha özel bilgiler de ifşa edilebilmektedir.

Bu tür bilgileri elde etmek için OF^k sistemlerinde, bir kullanıcı tarafından verilen oylar arasından olağan dışı oy verme eğilimlerini kullanmak, doğru tahminler üretmek için olanak sağlayacaktır. Kullanıcıların cinsiyetlerini tanımlamak ya da tahmin etmek için popüler bir şarkıcı, film yıldızı ya da özellikle erkekler için futbol içerikli kaynaklar kullanılabilir. Örneğin, popüler bir futbolcu ile ilgili belgeselde oyunculuk alt-ölçütü dışındaki ölçütleri beğenmeyen bir kullanıcının cinsiyeti hakkında çıkarım yapmak mümkün hale gelebilir. Benzer bir senaryoda popüler bir şarkıcının hayatına ait bir belgeselde kullanıcının cinsiyeti hakkında çıkarım yapmanın yanı sıra yaş aralığı hakkında da tahmin yapmak mümkün hale gelmektedir. Özetle, genel ölçüt ve alt-ölçütler arasındaki ilişkiye dayanan daha düşük ya da yüksek puanlamalar kullanıcı profili hakkında yeni çıkarımlar yapmaya olanak sağlarken aynı zamanda da kullanıcı mahremiyetini geleneksel *OF* sistemlerine göre daha fazla ihlal etmektedir.

Bir başka örnekte restoran değerlendiren bir kullanıcının tek-ölçütle beğenilerini derecelendirdiğini varsayalım; oyların çoğunu nispeten pahalı ve lüks restoranlara verdiği düşünüldüğünde, kullanıcının bu restoranlara gitmek için yeterince varlıklı olduğunu varsaymak makul olacaktır. Ayrıca, kullanıcının gelir seviyesi, tercih edilen restoranların ortalama fiyat aralığına göre kabaca tahmin edilebilir ve kullanıcı kolayca fiyat ayrımcılığına maruz kalabilir. Bununla birlikte, çoklu-ölçütlü tercihler kullanıldığı varsayılırsa, bu kullanıcının maliyet ölçütüne diğer alt-ölçütlere nazaran daha düşük oylar verdiği gözlemlenirse, bu kullanıcının aslında tahmin edilenin aksine daha düşük gelir seviyesinde olduğu anlaşılabilir. Çoklu-ölçütlü derecelendirmelere dayalı çıkarımlarda bu türden bir çelişki, tercihlerin ayrıntıları hakkında daha fazla bilgi sağlayacağından, kullanıcıların daha ciddi gizlilik risklerine maruz kalabileceği anlaşılmaktadır. Sonuç olarak, kullanıcının normalde ifşa etmek istemediği gizli bilgiler tehlikeye girer. Ayrıca,

bu metodolojiyi uygulayarak hizmet sağlayıcı, sistemdeki tüm kullanıcıları gelir düzeylerine göre sınıflandırabilir ve hedefli pazarlamayı gruplara uygulayabilir.

Bir başka örnekte, geleneksel tek-ölçüte dayalı *OF* sistemi kullanan bir otel rezervasyon sitesi ile çoklu-ölçütlü versiyonunu kıyaslayalım. Geleneksel *OF* sistemi kullanan bir çevrimiçi hizmette kullanıcılar gittiği otellerle ilgili görüşlerini sadece tek bir genel ölçüt üzerinden değerlendirebilir. Genel beğeni değeri kullanılarak, kullanıcının sadece gittiği otel ve bu oteli ne kadar sevdiği bilgisi elde edilebilir. Özetle elde edilebilecek bilgi oldukça sınırlıdır. Aynı örneği çoklu-ölçüt bakış açısı ile yeniden ele aldığımızda ise kullanıcı açısından çoklu-ölçüt kullanımının getirdiği yeni mahremiyet riskleri gözlemlenebilir. Örneğin yaygın olarak kullanılan birçok otel rezervasyon ve gezi deneyim sitesinde birden çok alt-ölçüt kullanılmaktadır. Bu alt-ölçütlerden bazıları gezgin tipi başlığı altında toplanan; aile, çift, arkadaş, yalnız ve iş gibi alt ölçütlerdir. Sistemin daha doğru öneri üretme potansiyelini arttıran bu alt ölçütler kullanıcının içinde yaşadığı toplum ve aile yapısına göre farklılık gösterse de genel anlamda direkt olarak ifşa edildiği durumlarda bireyin gizliliğini göz ardı eden mahremiyet ihlalleridir.

Aynı otel örneğinden yola çıkıp farklı bir bakış ile inceleyecek olursak, yorum yapılan ya da oylanan diğer alt ölçütler yeni riskler ortaya çıkarabilmektedir. Konuyu sadece bireyin mahrem kalması gereken kimlik bilgilerini ya da gizli bilgilerini ifşa etmek olarak düşünmemek gerekmektedir. Ortaya çıkan bu mahremiyet ihlalinin doğuracağı yeni problemler de göz ardı edilmemelidir. Örneğin aile odası alt ölçütünü derecelendiren bir kullanıcının bu derecelendirmesinden dolayı çocuk sahibi olduğu bilgisi elde edilebilmektedir. Bu bilgi kullanılarak kullanıcı daha sonra yapacağı rezervasyonlarda aile tipi odalar ya da çocuk dostu otellerle ilgili fiyat ayrımcılığına maruz kalabilir. Görüldüğü üzere tek-ölçüt kullanan bir sistemde kullanıcı-ürün matrisinde sadece gidilen otel ve o otele verilen beğeni değeri bulunmakta iken çoklu-ölçüt kullanımı ile ortaya çıkan birden çok alt-ölçüt, kullanıcı mahremiyetini farklı yollarla ihlal edebilmektedir ve kullanılan her bir alt-ölçütün kendisine has mahremiyet riskleri olabilir.

Bilindiği üzere birçok çevrimiçi hizmet özellikle de ürün satış siteleri sahip olduğu kullanıcı verisini analiz ettirmek ya da gelir elde etmek için ihtiyaç duydukları takdirde üçüncü şirketler ile paylaşabilir ya da satabilirler. *OF* sistemi kullanan bu tip çevrimiçi hizmetler ellerindeki kullanıcı ürün matrislerinin yetersiz olduğu ve kullanıcılarına doğru öneriler üretemediği senaryolarda da bu tip işbirliklerini yapmaktadırlar. Bu gibi durumlarda bazen kullanıcı-ürün matrislerinde ortak kullanıcıların bulunduğu durumlar

söz konusu olurken bazen de ortak ürün listeleri bulunabilmektedir. Ortak kullanıcıların bulunduğu durumlar söz konusu olduğu durumlarda, OF^k sistemlerinde alt-ölçütler kullanılarak yapılacak kimlik ifşaları bireyler için oldukça tehlikeli sorunlara neden olabilmektedir. Otel örneği üzerinden devam edilecek olursa çocuk sahibi olduğu bilinen bir kullanıcı verisi üçüncü şirketler ile veri paylaşımı neticesinde farklı bir çevrimiçi hizmette de ifşa edilirse bu site üzerinden yapacağı çocuk ürünlerinde fiyat ayrımcılığına maruz kalabilmekte ve çocuk ürünleri hakkında e-posta bombardımanıyla karşılaşabilmektedir. Özetle, bir veri kümesindeki olası alt-ölçütlerin anlamsal ilişkileri göz önünde bulundurulduğunda, çoklu-ölçütlü bir profil hakkında yapılabilecek derecelendirme modeli, elde edilmek istenen çıkarımları daha ilginç hale getirebilmekte ve çoklu-ölçütlü tercih toplama, tek-ölçütlü koleksiyonlarda yer alan gizlilik risklerini artırmaktadır.

3.5. Sonuçlar

Bireysel mahremiyetle ilgili kaygılar, öneri hizmetlerinin kullanımı yaygınlaştıkça daha da artmaktadır. Gizlilik endişelerine sahip kullanıcılar, bu risklerinden kaçınmak için gerçek beğenileri yerine sahte tercihler sunmayı tercih edebilir ya da bu hizmetleri kullanmaktan tamamıyla vazgeçmektedir. Bazı araştırmalar kullanıcılar hakkındaki mahremiyet ihlallerinin, kişisel bilgilere doğrudan erişim sağlamadan çıkarım yoluyla da elde edilebileceğini göstermektedir. Bu bağlamda, bireylerin çoklu-ölçütlü tercih verileri aracılığı ile yaşam tarzı özellikleri, alışveriş alışkanlıkları, finansal durumu, yaşı, cinsiyeti, aile bilgisi ve etnik kökeni gibi kişisel tanımlayıcıları kötü niyetli girişimlerle elde edilebilmektedir.

Kullanıcı tercihlerinin bir ürünün/hizmetin birden çok yönüne göre toplanması durumunda bu tür gizlilik risklerinin ortaya çıkması olasılığı daha da artmaktadır. Tek-ölçütlü tercih toplama nedeniyle mevcut risklerin daha da artmasının yanı sıra, çoklu-ölçütlü sistemler, alt ölçütlerin birleştirici niteliği nedeniyle ek gizlilik riskleri ortaya çıkarmaktadır. Çoklu-ölçütlü derecelendirmeler, bir kullanıcının belirli bir ürünü veya hizmeti neden beğendiğini veya beğenmediğini belirleyerek öneri hizmet kalitesini iyileştirmeye yardımcı olsa da, aynı zamanda bireylerin gizliliğini tehlikeye atmak konusunda da risklidir. Bu tez, bu olası risklerin bazılarını tanımlamakta ve gelecek vaat eden çoklu-ölçütlü öneri hizmetlerine gizlilik koruma mekanizmalarının geliştirilmesi için dikkat çekmeyi amaçlamaktadır. Sonuç olarak, tek-ölçütlü sistemler için mevcut

gizliliği koruyan yöntemlerin etkinliklerini çoklu-ölçütlü platformlara genişletmek ve tercih özelliklerinin birleştirici doğasında yer alan ek gizlilik risklerini önlemek için yeni yaklaşımlar geliştirilmelidir.

4. GİZLİLİĞİ KORUYAN ÇOKLU-ÖLÇÜTLÜ ORTAK FİLTRELEME

OF^k yaklaşımlarında kullanılan çoklu-ölçütlü kullanıcı derecelendirmeleri, bir öğenin kullanıcı tarafından beğenilme seviyesini göstermekle birlikte bu öğenin hangi özelliğinden dolayı belirlenen seviyede beğenildiği bilgisine de sahiptir. Kullanıcıların, bir ürün/hizmet hakkındaki görüşlerini birden fazla ölçüt aracılığıyla ifade edebilmesi *OF*^{k} sistemlerinin daha kişiselleştirilmiş kullanıcı profillerine ulaşmasına yardımcı olmaktadır. Kullanıcıların alt-ölçütlere verdiği önemi keşfetmek, kullanıcıların beğeni profilleri arasındaki benzerlikleri daha kesin bir şekilde belirlemeyi ve dolayısıyla bireyler için daha kişiselleştirilmiş, aynı zamanda daha doğru tahminler elde etmesini sağlamaktadır (Jannach vd., 2012). Bununla birlikte çoklu-ölçütlü kullanıcı verisi toplayan *OF*^{k} sistemleri, kullanıcılarını bir ürünün veya hizmetin iyi tanımlanmış alt özellikleri için tercihlerini sunmaya yönlendirmektedir. Bunun sonucunda da geleneksel *OF* sistemleri ile kıyaslandığında bireyler daha ciddi gizlilik tehditleri ile karşı karşıya getirilmektedir.

Çoklu-ölçütlü kullanıcı derecelendirmeleri Bölüm 3'te de ifade edildiği gibi beraberinde daha fazla gizlilik riskleri getirmektedir. Alt-ölçütlerin sahip olduğu kendine özgü bilgiler kullanarak; kullanıcıya ait demografik bilgiler, yaşam tarzı özellikleri, alışveriş alışkanlıkları, finansal durum, yaş-cinsiyet gibi kullanıcıya ait gizli bilgilerin tahmin edilebilirliğini arttırmaktadır. Çoklu-ölçütlü verilerin kullanımından kaynaklanan daha ciddi gizlilik tehditlerine rağmen, bunları hafifletecek çözümler nispeten sınırlıdır (Yargıç ve Bilge, 2017). Bunun önemli bir sonucu olarak, kullanıcıların *OF*^{k} sistemlerine karşı önyargı oluşturmasına neden olmaktadır. Önyargılı kullanıcılar, *OF*^{k} sistemini kullanırken bireysel gizlilik koruma mekanizmalarını geliştirip bazen kasten yanlış bilgi beyan ederek sistemi manipüle etmekte, bazen de *OF*^{k} sistemi kullanmayı tamamen reddetmektedir (Ackerman vd., 1999). Ancak *OF*^{k} sistemleri, doğru tavsiyeler üretmek için gerçek ve özgün tercih verilerine ihtiyaç duymaktadır. *OF*^{k} sistemleri, kullanıcılarının gerçek beğeni değerlerini elde etmek ve bu verileri kullanarak doğru öneriler üretebilmek için gizlilik koruma yaklaşımlarına ihtiyaç duymaktadır.

GKOF sistemleri, tek-ölçütlü derecelendirme kullanımıyla ortaya çıkabilecek gizlilik tehditlerini ortadan kaldırmaya odaklanmaktadır ve çoklu-ölçütlü tercih verileri alanındaki gizlilik riskleri göz ardı edilmektedir. Bu bölümde, geleneksel tek-ölçütlü *OF* sistemlerinde verimli bir şekilde kullanılan gizlilik koruma yöntemleri, çoklu-ölçütlü derecelendirme verilerine adapte edilmiştir. *OF*^{k} sistemleri için *RK* ve *RD* temelli Gizliliği

Koruyan Çoklu-Ölçütlü Ortak Filtreleme (*GKOF*^{k}) yaklaşımları sunulmaktadır. Böylece, *OF*^{k} sistemleri için referans gizlilik ve doğruluk seviyeleri elde edilmiştir. Önerilen gizlilik koruma şemalarının oy ve kullanıcı dağılımı bakımından farklı karakteristik özelliklere sahip veri setleri üzerinde tahmin doğruluğu ve veri gizliliği düzeylerine etkilerini incelemek için *YM20*, *YM10* ve *YM5* veri setlerinde önerilen *GKOF*^{k} yaklaşımı test edilmiştir.

4.1. *RK* ve *RD* Yöntemleri ile *GKOF*^{k}

OF^k sistemlerinde *kullanıcı* × *ürün* matrisi genel derecelendirme değeri r_o ile birlikte k tane alt-ölçütten meydana gelmektedir. Geleneksel *GKOF* sistemlerinde uygulanan *RK* ve *RD* yöntemleri genel derecelendirme vektöründen oluşan iki boyutlu *kullanıcı* × *ürün* derecelendirmeleri için maskeleme işlemini gerçekleştirmektedir. *GKOF*^{k} sistemlerine geleneksel maskeleme işleminin uyarlanabilmesi için her bir ölçütün ayrı ayrı maskelenmesi gerekmektedir. Bu işlem gerçekleştirilirken, her bir kullanıcı derecelendirme ölçütü istemci tarafında oluşturulan maskeleme verisi ile gizlenmektedir. Burada ölçütler, kullanıcının ihtiyaç duyduğu gizlilik seviyesine göre belirlediği gizlilik ihtiyacına göre üretilmiş rastgele sayı vektörleri ile maskelenmektedir. Bu işlem sonucunda, çoklu-ölçütlü veri seti için oluşturulan maskeleme fonksiyonu $P = (G_0 + R_0, G_1 + R_1, \dots\ G_k + R_k)$ elde edilmektedir.

Bölüm 2.2'de açıklanan *GKOF* sistemlerinde kullanılan yaklaşıma benzer bir şekilde, *GKOF*^{k} yaklaşımında da veri maskeleme vektörü, servis sağlayıcı tarafından belirlenen ve en yüksek gizlilik seviyesini ifade eden σ_{max} ve β_{max} katsayısı kullanılarak kullanıcının ihtiyaç duyduğu gizlilik seviyesine göre üretilmektedir. Kullanıcı gizlilik seviyesini belirlemek için kullanılan σ katsayısı (0, σ_{max}] ve β katsayısı (0, β_{max}] değer aralığında kullanıcı tarafından belirlenmektedir. Çoklu-ölçütlü veri setleri üzerinde maskeleme vektörünün oluşturulması ve eklenmesine ait süreç Prosedür 4.1'de verilmiştir. *GKOF*^{k} sistemi için oluşturulan maskeleme prosedüründe geleneksel yönteme ek olarak, oluşturulan veri maskeleme vektörü her bir alt-ölçüte ayrı ayrı eklenir ve tüm ölçütler için maskelenmiş z-skor vektörleri ($z'_{uk} = z_{uk} + r_{uk}$) oluşturulur.

Prosedür 4.1. *GKOF^k sistemlerinde veri maskeleme prosedürü*

Require: Kullanıcı × ölçüt × ürün vektörü (G_k), σ_{max}, β_{max}

***z*-skoru değerinin hesaplanması** ($\rightarrow \boldsymbol{Z_k}$)

1 : **for all criteria in** G $(i \leftarrow 1\ to\ k)$ **do**

2 : $\overline{G_i} \leftarrow MEAN(G_i)$

3 : $\sigma_{G_i} \leftarrow STD(G_i)$

4 : **end for**

5 : **for all criteria in** G $(i \leftarrow 1\ to\ k)$ **do**

6 : **for all items in** G_i $(i \leftarrow 1\ to\ m)$ **do**

7 : $Z_{ij} = (G_{ij} - \bar{G}_i)/\ \sigma_{G_i}$

8 : **end for**

9 : **end for**

Gizlilik parametrelerinin belirlenmesi

10 : $\beta \leftarrow RND\ (0, \beta_{max}]$;

11 : $\sigma \leftarrow RND\ (0, \sigma_{max}]$

12 : $\alpha \leftarrow \sqrt{3}\sigma$;

13 : $e \leftarrow |E|$ ► # oy verilmemiş ürün

14 : $g \leftarrow |G|$ ► # gerçek kullanıcı oyları

15 : $F \leftarrow e \times \beta\%$ ► # doldurulacak hücre sayısı

Dağılımın belirlenmesi ve rastgele sayı türetilmesi

16 : $dist \leftarrow RANDOM\ (uniform,\ normal)$

17 : $R \leftarrow dist(g + F;\ \mu = 0, \sigma|\ \alpha)$

***z*-skoru değerlerinin maskelenmesi** ($\rightarrow \boldsymbol{Z'}$)

18 : **for all criteria in** G $(i \leftarrow 1\ to\ k)$ **do**

19 : **for all items in** G_i $(i \leftarrow 1\ to\ m)$ **do**

20 : $Z'_{ij} = (Z_{ij} + R_{ij})$

21 : **end for**

22 : **end for**

23 : return Z'

4.2. Gizlilik Analizi

Bu bölümde *RK* ve *RD* prosedürleri ile oluşturulan rastgele sayı vektörünün kullanıcı derecelendirmelerine sağladığı gizlilik seviyesini ölçekleyebilmek için kullanılacak metrikler açıklanmaktadır.

4.2.1. *RD* yönteminin gizlilik analizi

RD prosedüründe kullanıcının derecelendirmiş olduğu ürün listesini gizlemek adına gerçek kullanıcı vektörünün üzerine, gerçekte kullanıcı tarafından derecelendirilmemiş, orijinal derecelendirilmiş ürün sayısına göre $\%\,\beta$ oranında sahte derecelendirmeler eklenmektedir. Kullanıcı tarafından belirlenen β katsayısının derecelendirilmiş ürün listesini maskelemek adına sağladığı gizlilik miktarını ölçeklendirmek için Bilge ve Polat (2013) tarafından önerilen Denklem 4.1 kullanılmaktadır. Bu denklemde, kullanıcıların gerçekte sahip olduğu derecelendirmeler ve eklenen maskeleme verisindeki derecelendirmelere göre kullanıcı profilindeki maskelenmiş veri modellenmekte ve elde edilen modelin entropisi (Shannon, 2001) nihai gizlilik seviyesi olarak sunulmaktadır.

Bu denklemde, bir kullanıcının gerçek derecelendirme sayısı *ds*, derecelendirilmemiş boş hücrelerinin sayısı ise *e* olarak ifade edilmektedir. Buna göre $G = \{g_1, g_2, \dots, g_{ds}\}$ gerçek derecelendirmelerin olasılık dağılımını, $F = \{f_1, f_2, \dots, f_{e\times\%\beta}\}$ ise sahte derecelendirmelerin dağılımlarını temsil etmektedir. Ayrıca, *#G* ve *#F* sırasıyla *G* ve *F kümelerindeki* öğelerin sayısını göstermektedir.

$$U = \frac{\#G + \#F}{ds + (e \times \beta_{max})} \tag{4.1}$$

β_{max} katsayısı ile elde edilen belirsizlik miktarını ölçeklendirme için Denklem 4.1 ile elde edilen *U* değerinin Shannon entropisi, *H(U),* Denklem 4.2 ile derecelendirilmektedir. Bilgi teorisinde, Shannon entropisi (*H*) belirli bir veri kaynağındaki belirsizliğin miktarı olarak tanımlanmaktadır. Bir *kullanıcı* × *ürün* matrisinde, olası oy değeri $X = \{x_1, x_2, \dots, x_t\}$ ve bu oy değerlerinin gözlemlenme oranları $P = p_1, p_2, \dots, p_t$ olarak tanımladığı varsayılırsa, bu sistemin entropisi *H(X)* Denklem 4.2 ile elde edilmektedir.

$$H(X) = -\sum_{i=1}^{t} p_i log_2(p_i) \tag{4.2}$$

4.2.2. *RK* yönteminin gizlilik analizi

RK tabanlı maskeleme yönteminin kullanıcı mahremiyeti üzerindeki etkisini analiz etmek için Agrawal ve Aggarwal (2001) diferansiyel entropi tabanlı bir gizlilik ölçümü metriği önermiştir. Bir *kullanıcı* × *ürün* matrisinde, rastgele olası oy değerinin $X = \{x_1, x_2, \dots, x_t\}$ ve bu oy değerlerinin gözlemlenme oranlarının $P = p_1, p_2, \dots, p_t$ olarak tanımlandığını varsayalım. Rastgele değişken *X*'in diferansiyel entropisi $h(X)$ Denklem 4.3 ile elde edilmektedir.

$$h(X) = -\int_{\Omega_X} f_x(x) log_2 f_x(x) dx \tag{4.3}$$

Burada *X*'in tanım kümesi Ω_X olarak ifade edilirken, *X*'in sahip olduğu belirsizlik miktarı $h(X)$ olarak tanımlanmaktadır. Sürekli bir olasılık dağılımının gizlilik seviyesini ölçeklendirmek için Agrawal ve Aggarwal (2001) Denklem 4.4'ü kullanmaktadır.

$$\prod(X) = 2^{h(X)} \tag{4.4}$$

Bir rastgele değişken *X*'in gizlilik seviyesi $2^{H(X)}$ ile hesaplanabileceği göz önünde bulundurularak maskelenmiş kullanıcı verisi *P*'nin gizlilik seviyesini elde etmek için bağımsız rastgele değişkenlerden *V* orijinal kullanıcı oy vektörünü temsil ederken, *R* maskeleme işleminde kullanılacak rastgele sayı vektörünü temsil etmektedir. Maskelenmiş veri vektörünü elde etmek için $P = V + R$ ile maskelenmiş *P* vektörü elde edilir. *R*'nin ortalama koşullu mahremiyeti $\prod(V|P) = 2^{H(V|P)}$ olarak tanımlanır. Burada $2^{H(V|P)}$, verilen *P* için *V*'nin koşullu diferansiyel entropisini sembolize eder. Böylece, *P*'nin ifşa edilmesinden sonra *V*'nin gizlilik seviyesi, $\prod(V|P) = \prod(V) \times (1 - \Pr(V|P))$ ile elde edilir (Agrawal ve Aggarwal 2001). Burada; *V*'nin *P* üzerindeki koşullu gizlilik kaybı $\Pr(V|P)$ olarak gösterilir ve Denklem 4.5 ile elde edilir.

$$\Pr(V|P) = 1 - 2^{H(V|P - H(P))} \tag{4.5}$$

Ancak, bu yöntem tek-ölçüt tabanlı veri kümeleri için kullanılmaktadır. Bu yöntemi OF^k tabanlı sistemlere uyarlamak için, her bir ölçütün gizlilik seviyeleri ayrı ayrı ölçülmektedir. Her bir ölçüte ait elde edilen gizlilik seviyelerinden, en düşük gizlilik seviyesine sahip alt-ölçüt s istemin sahip olduğu nihai gizlilik değeri olarak sunulmaktadır.

4.3. Maskelenmiş Veri ile Öneri Üretme

GKOF^k sistemlerinin temel amacı, gerçek kullanıcı derecelendirmelerini maskelerken aynı zamanda da maskelenmiş derecelendirme vektörleri kullanarak doğruluk seviyesi yüksek öneriler üretilmesini sağlamaktır. Yapılan maskeleme işlemine rağmen, *GKOF*^{k} sistemi veri tabanında yeterli kullanıcı tercih verisi mevcutsa, maskelenmiş veri kullanılarak sistem tarafından kabul edilen ve önceden öngörülen doğruluk kayıpları ile halen öneri üretmek mümkündür (Bilge vd., 2013). Oluşturulacak öneriler *z*-skoru normalleştirme sürecine tabi tutulmuş ve maskelenmiş *kullanıcı* × *ürün* matrisi üzerinden üretileceği için kullanıcı komşulukları Denklem 2.1 revize edilerek elde edilmektedir (Polat ve Du,2005a; Bilge ve Polat, 2013). Maskelenmiş *z*-skoru vektörleri üzerinden *a* kullanıcısının *u* ile benzerlikleri z'_{ai} ve z'_{ui} kovaryansından oluşan modifiye edilmiş *PK* katsayısı Denklem 4.6'da verilmiştir.

$$PK_{au} \approx PCC'_{au} = \sum_{i}^{m} z'_{ai} \times z'_{ui} \tag{4.6}$$

Maskelenmiş *z*-skoru vektörleri üzerinden *a* kullanıcısının *u* ile benzerlikleri elde edildikten sonra *q* ürünü için öneri üretme işlemi Denklem 4.7 aracılığı ile yapılmaktadır (Polat ve Du,2005a, Bilge ve Polat, 2012).

$$P_{aq} \approx P'_{aq} = \bar{r}_a + \sigma_a \frac{\sum_{u=1}^{k} PK'_{au} \times z'_{uq}}{\sum_{u=1}^{k} PK'_{au}} \tag{4.7}$$

Böylece, gizliliği koruyan *OF* sistemi gizlenmiş kullanıcı vektörü üzerinde yüksek doğruluk seviyelerinde tahminler oluşturabilir.

4.4. Deneysel Yaklaşımlar ve Elde Edilen Sonuçlar

Bu bölümde öncelikli olarak kullanılan deneysel metodolojiden bahsedilmiştir, sonrasında ise *GKOF*^{k} sistemleri için önerilen gizlilik koruma yaklaşımları σ_{max} ve β_{max} gizlilik kontrol parametreleri üzerinden kapsamlı bir şekilde analiz edilmiştir. Elde edilen gizlilik seviyeleri ve öneri doğrulukları göz önünde bulundurularak *YM5*, *YM10* ve *YM20* veri setleri için ideal σ_{max} ve β_{max} parametreleri belirlenmiştir. Son olarak, belirlenen ideal gizlilik seviyeleri göz önünde bulundurularak elde edilen maskelenmiş kullanıcı derecelendirme vektörlerinin *GKOF*^{k} sisteminin öneri üretme doğruluğu üzerindeki

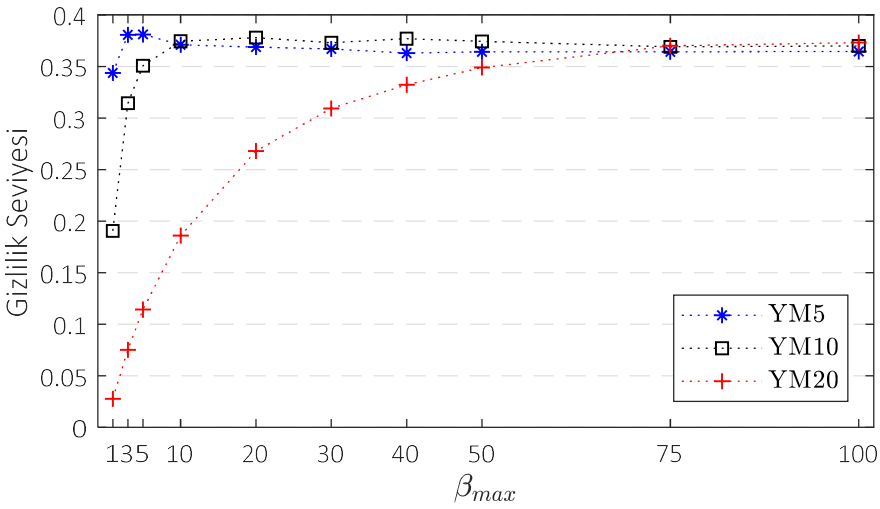
etkileri incelenmiş ve üretilen önerilerin istatistiksel olarak anlamlılık düzeyleri test edilmiştir.

Maskeleme verisinin öneri üretme doğruluğuna getirdiği negatif etkiyi tüm kullanıcı tercihleri üzerinde kapsamlı bir şekilde analiz edebilmek için birini dışarıda bırak çapraz doğrulama metodolojisi kullanılarak öneri üretme işlemi gerçekleştirilmektedir. Çapraz doğrulama metodolojisi ile her kullanıcı test için aktif kullanıcı olarak ele alınır ve kalan kullanıcılar eğitim verilerini oluşturur (Jannach vd., 2012; Bilge ve Yargıç, 2017). Böylece sistemde kayıtlı her bir kullanıcının her bir ürünü için ayrı ayrı öneri üretme süreci gerçekleştirilmiştir. Önerilen yöntemlerin etkilerini test etmek, kullanıcıların bireysel olarak belirlediği gizlilik seviyesi ve oluşturulan maskeleme verisinin diziliminden kaynaklanabilecek rastlantısallığın öneri doğruluğunda ortaya çıkarabileceği sapmaları hafifletmek için her deney seti 10 kez tekrarlanmıştır. Tüm deney setlerinden elde edilen sonuçların ortalamaları nihai gizlilik ve doğruluk değerleri olarak sunulmuştur.

Gross ve Acquisti (2005)'e göre, bireylerin mahremiyete yönelik tutumları birbirinden farklı olabilir. Bu nedenle, önerilen gizlilik koruma mekanizmalarının etkisini araştırmak için, deneyler çeşitli σ_{max} ve β_{max} gizlilik kontrol parametreleri üzerinden gerçekleştirilmektedir. Bireylerin gizlilik tercihlerini taklit etmek için, β_{max} parametresi %5'lik artan değer aralıkları ile %100 seviyesine kadar test edilmiştir. *RD* yönteminde, değişken β_{max} parametrelerinin gizlilik ve doğruluk üzerindeki etkisini test etmek için σ_{max} parametresinin 3'te sabit olduğu varsayılmıştır, σ_{max} için 3 değerinin seçilme nedeni bu değerin makul gizlilik seviyesine sahip olmasıdır. Bireylerin gizlilik seviyelerini belirlerken gerçek kullanıcı davranışlarını taklit etmek için gizlilik parametreleri (0,1] değer arasında rastgele bir sayı ile çarpılmıştır. *RK* yöntemini test etmek için yapılan deneysel çalışmalarda ise σ_{max} parametrelerinin sistem gizliliği üzerindeki etkisini test ederken β_{max} parametresinin 0'da sabit olduğu varsayılmıştır ve σ_{max} parametresi $[0{,}5;\ 5]$ aralığında $0{,}5$ aralıklarla artan değerlerle test edilmiştir. *RK* ve *RD* yöntemlerinde, oluşturulan maskeleme verisi $\mathcal{N}$ ve $\mathcal{U}$ dağılımlar ile oluşturulan rassal sayı vektörleri ile ayrı ayrı test edilmiştir. Deneysel çalışmaların detayları ve yapılan testlerin gizlilik ve doğruluk üzerine etkileri bir sonraki bölümde ayrıntılı olarak incelenmektedir.

4.4.1. *RD* yönteminin gizliliğe etkisi

RD yöntemi ile elde edilen gizlilik seviyeleri Pr(β_{max}), gerçek kullanıcı verilerinden oluşan Yahoo!Movies veri setinin üç alt kümesi olan *YM20*, *YM10* ve *YM5* veri setleri üzerinde değerlendirilmiştir. Bu veri setlerinde sırasıyla *YM5*, *YM10* ve *YM20* için toplamda 2565, 1164 ve 247 adet ürün tanımlıdır ve ortalama her bir kullanıcının oy verdiği ürün sayısı yaklaşık olarak 15, 25 ve 40 adettir. Bu koşullar altında, Denklem 4.1 kullanılarak modellenen kullanıcı profillerine ait Pr(β) değerleri [1,100] aralığında test edilmiş ve elde edilen gizlilik seviyeleri Şekil 4.1'de gösterilmektedir.



Şekil 4.1. *Değişken β_{max} değerlerinin gizlilik üzerine etkisi*

Şekil 4.1'de gösterilen mahremiyet seviyeleri, β_{max} katsayısının seviyesine göre değişkenlik göstermektedir. Ayrıca veri setinin seyreklik seviyesi ile elde edilen gizlilik seviyesi arasında güçlü bir ilişki de gözlemlenmektedir. Bölüm 2.3'te de ifade edildiği gibi Yahoo!Movies veri setinin alt kümelerinden oluşan *YM5*, *YM10* ve *YM20* veri setlerinin seyreklik oranları sırasıyla %99,44; %97,69 ve %83,65'tir. *Kullanıcı* × *ürün* matrisinin büyüklüğüne oranla yüksek seyreklik oranına sahip olan *YM5* ve *YM10* veri setleri için, %5-%10 gibi nispeten küçük oranlarda β_{max} değerlerinin kullanılması, bireylerin mahremiyetini sağlamak için yeterlidir. *YM5* ve *YM10* veri kümeleri için daha büyük bir β_{max} değeri kullanmamanın, kullanıcı gizliliğine olumlu yönde bir katkısı bulunmamaktadır. Bununla birlikte, kullanıcı mahremiyetini *YM20* gibi seyrekliği düşük bir veri setinde sağlamak için, β_{max} değeri %45-%50 gibi yüksek bir oranda kullanmak gerekmektedir. Bu sonuçlara dayanarak *YM20* gibi daha yoğun kullanıcı derecelendirmelerine sahip veri setlerinde, daha yüksek seviyelerde β_{max} değerinin

kullanılması kullanıcı mahremiyetini olumlu yönde etkilemektedir. *YM5* ve *YM10* gibi daha seyrek veri setlerinde ise, orijinal ve sahte derecelendirmeler arasındaki belirsizlik artışından dolayı nispeten düşük değerlerde β_{max} kullanımıyla daha yüksek seviyelerde mahremiyet elde etmek mümkün hale gelmektedir. *YM5* ve *YM10* gibi yüksek seyreklik seviyelerine sahip veri setlerinde ortalama olarak kullanıcının oy verdiği ürün sayısı düşük olduğu için β_{max} katsayısı düşük seviyelerde tutulsa da gizlilik seviyesine önemli katkılar sağlamaktadır. Ayrıca, *YM5* ve *YM10* veri setlerinde yüksek β_{max} seviyelerinde gizliliğin artması beklenirken bu seviyenin sabit kaldığı gözlemlenmektedir. Bunun nedeni şöyle açıklanabilir, β_{max} değerini yüksek seviyelerde tutarak kullanıcının gerçekte derecelendirdiği ürün sayısına oranla yüksek miktarda sahte oy değeri eklemek, gerçek ve sahte derecelendirmelerin belirsizlik düzeyini gereğinden fazla saptırdığı için mahremiyet seviyesine etkisini azaltmaktadır.

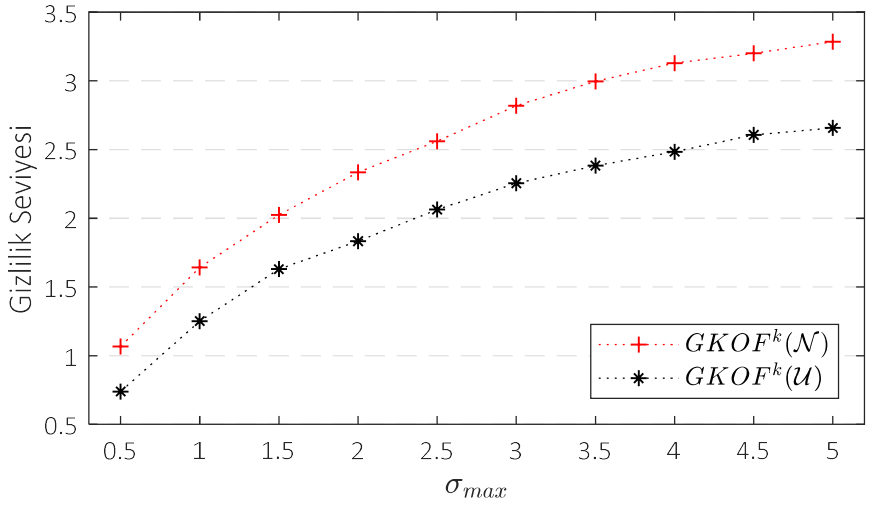
Özetle; en yüksek mahremiyet seviyesine ulaşmak için ideal β_{max} değerinin kullanılan veri setinin seyreklik oranı ile orantılı olarak belirlenmesi gerektiği sonucuna varılabilir. *YM20* gibi daha yoğun veri kümelerinde, gerçek derecelendirmelerin yeterli miktarda gizlenmesi için yüksek seviyelerde β_{max} değeri gereklidir. Bununla birlikte, β_{max} değerini gerçek oy değerinden fazla sayıda sahte oy değeri üretmeye zorlayan seviyelere kadar yükseltmek, gerçek ve sahte derecelendirmelerin belirsizlik düzeyini gereğinden fazla saptıracağı için beklenen gizlilik düzeyine katkıda bulunmaz. Bu nedenle, en yüksek gizlilik düzeyine ulaşmak için ideal β_{max} değerinin kullanıcının profilindeki gerçek derecelendirme sayısına uygun olması gerektiği sonucuna varılmaktadır.

4.4.2. *RK* yönteminin gizliliğe etkisi

$\mathcal{N}$ ve $\mathcal{U}$ dağılıma göre elde edilen rastgele gürültü vektörünün değişken σ_{max} seviyelerinde elde ettiği gizlilik seviyeleri $\prod(V|P)$ Şekil (4.2, 4.3 ve 4.4)'te sırasıyla *YM20*, *YM10* ve *YM5* veri setleri için gösterilmiştir.

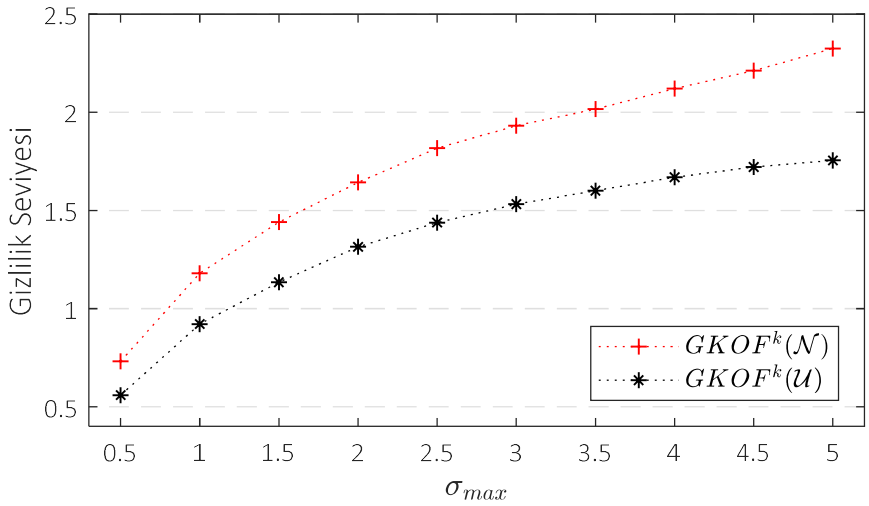
YM20 veri seti için $\mathcal{N}$ ve $\mathcal{U}$ dağılımlara göre elde edilen rassal sayıların $\sigma_{\max} = [0{,}5; 5]$ değer aralığı içerisindeki gizlilik sevileri Şekil 4.2'te verilmektedir. Beklendiği üzere artan σ değeri ile elde edilen gizlilik seviyesi arasında doğrusal bir ilişki gözlemlenmektedir. Artan σ_{max} değeri ile gerçek kullanıcı verilerine eklenecek olan maskeleme verilerinin rassallığı artacağından gizlilik seviyesinin de artması beklenen bir sonuçtur. $\mathcal{N}$ ve $\mathcal{U}$ dağılımlara göre elde edilen gizlilik seviyeleri karşılaştırıldığında $\mathcal{N}$

dağılımla üretilen rassal sayıların gizliliğe katkısı $\mathcal{U}$ dağılıma göre bütün σ_{max} değerleri göz önünde bulundurulduğuna % 27,44 oranında gizlilik seviyesinde iyileştirmeye neden olmaktadır.



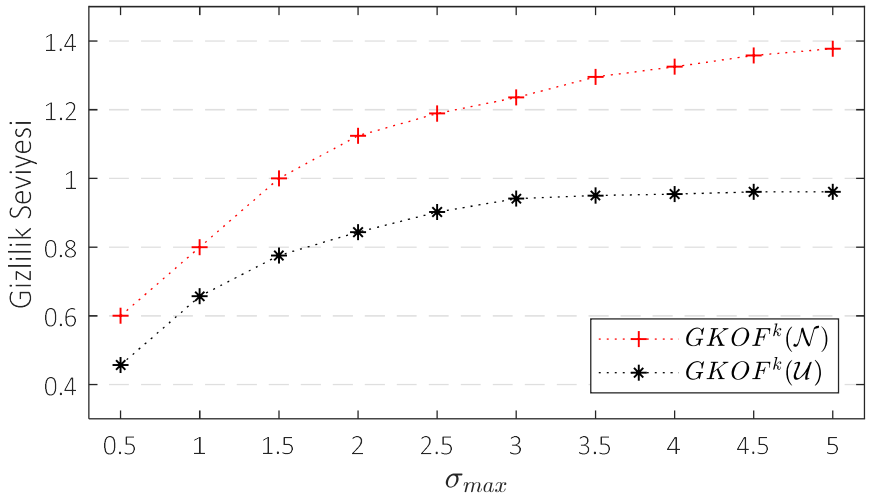
Şekil 4.2. *GKOF^k($\mathcal{N}$) ve GKOF^k ($\mathcal{U}$) için YM20 veri setinde değişken σ_{max} değerinin gizliliğe etkisi*

Aynı deneysel metodoloji ile *YM10* için yapılan deneysel çalışmada $\mathcal{N}$ dağılım ile üretilen rassal sayılar $\mathcal{U}$ dağılım ile elde edilen rassal sayılara göre % 27,74 oranında daha yüksek mahremiyet seviyelerine ulaşmaktadır. Kullanıcı ürün oy yoğunluğu bakımından *YM20*'ye göre daha seyrek olan *YM10* veri setinde yapılan deneylerde σ_{max} değerinin gizliliğe olan etkisi *YM20* veri setinde elde edilen sonuçlara paralellik gösterse de elde edilen nihai gizlilik seviyeleri daha düşük seviyelerde kalmaktadır. Bunun nedeni veri setinin seyreklik seviyesi bir başka deyişle kullanıcı başına ortalama derecelendirme sayısının *YM20* veri setine göre daha az olmasıdır. *YM20* veri setinde bir kullanıcı ortalama olarak 40 ürünü derecelendirirken, *YM10* veri setinde kullanıcı başına ortalama derecelendirilen ürün sayısı yaklaşık olarak 25 üründen oluşmaktadır. Azalan derecelendirme sayısı oluşturulan maskeleme vektörünün boyutunu da düşürmektedir. Bu nedenle, daha az elemana sahip olan maskeleme vektörü daha düşük seviyede gizlilik içermektedir.



Şekil 4.3. *GKOF^k ($\mathcal{N}$) ve GKOF^k ($\mathcal{U}$) için YM10 veri setinde değişken σ_{max} değerinin gizliliğe etkisi*

Son olarak *YM5* veri seti üzerinde karşılaştırılan σ_{max} değeri $\mathcal{N}$ ve $\mathcal{U}$ dağılım ile elde edilen maskeleme verilerinin gizliliğe sağladığı katkı göz önünde bulundurulduğunda *YM10* ve *YM20* ile elde edilen sonuçlara paralellik gösterip, $\mathcal{N}$ dağılım ile elde edilen mahremiyet seviyesi $\mathcal{U}$ dağılım ile elde edilen gizlilik seviyesine göre % 33,82 oranında daha yüksek gizlilik içermektedir. Ancak elde edilen gizlilik seviyeleri göz önünde bulundurulduğunda en düşük gizlilik seviyesi *YM5* veri seti üzerinde elde edilmiştir. Bunun nedeni, *YM10* veri setinin özelliklerinde de bahsedildiği gibi kişi başına düşen ortalama ürün derecelendirme sayısı ve buna paralel olarak oluşturulan daha düşük miktardaki maskeleme vektörüdür.



Şekil 4.4. *GKOF^k ($\mathcal{N}$) ve GKOF^k ($\mathcal{U}$) için YM5 veri setinde değişken σ_{max} değerinin gizliliğe etkisi*

Özetle; kullanıcı gizliliği ve σ_{max} arasında doğrusal bir ilişki bulunmaktadır ve σ_{max} değerindeki artış gizlilik seviyesini de olumlu yönde etkilemektedir. Artan σ_{max} değeri ile orijinal veri seti üzerine eklenen rassal sayı vektörünün değer aralığı ve bununla ilişkili olarak gizlilik seviyesi artacaktır. Ancak, kullanıcı için maskeleme işleminde kullanılacak ideal σ değeri sadece elde edilen gizlilik seviyesi üzerinden karar verilebilecek bir parametre değildir. Bu nedenle, kullanıcı için en uygun σ değeri ve servis sağlayıcı için en uygun σ_{max} değerini belirlemek için gizlilik ve öneri doğruluğu arasında bir denge kurulmalıdır.

4.4.3. Öneri üretme doğruluğu

RD ve *RK* yöntemlerinin değişken σ_{max} ve β_{max} parametreleri ile öneri doğruluğu üzerine etkisini test etmek için *YM20*, *YM10* ve *YM5* veri setleri üzerinde *k*-en yakın komşuluk tabanlı OF^k algoritmaları temel alınarak elde edilen öneri doğrulukları kıyaslanmaktadır. Öneri üretme sürecinde, birini dışarıda bırakarak çapraz doğrulama deney metodolojisi kullanılarak her bir kullanıcının oy verdiği her bir ürünü için öneri üretme işlemi gerçekleştirilmiştir. Bu işlem gerçekleştirilirken, öneri üretilecek her bir ürün kullanıcı derecelendirme matrisinden silinip, elde edilen yeni *kullanıcı* × *ürün* matrisine yeniden *RK* ve *RD* işlemleri uygulanmaktadır. Öneri üretilecek derecelendirme değerinin çıkarıldığı V' vektörü üzerine Prosedür 4.1'e göre yeniden oluşturulan R' vektörünün eklenmesiyle $P' = V' + R'$ vektörü elde edilmektedir. Bunun sonucunda, öneri üretilecek ürünün kullanıcı derecelendirme vektörü üzerinde ortalama ve standart sapma değerlerinde ortaya çıkaracağı sapmalar ortadan kaldırılarak daha gerçekçi öneri üretilmesi sağlanmaktadır.

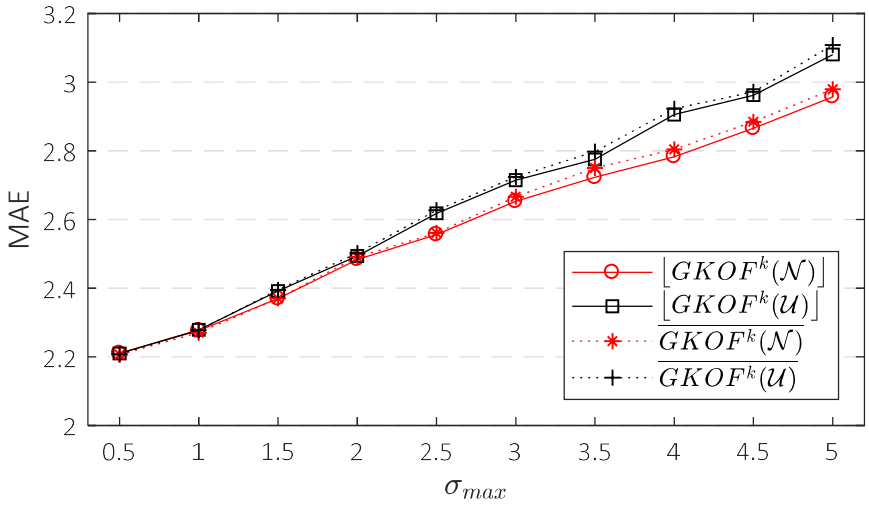
Öneri üretme sürecinde Bölüm 2.1.2'de tanımlanan $\bar{\cdot}$ ve $\lfloor\cdot\rfloor$ benzerlik elde etme yöntemlerine göre ürün önerileri üretilmektedir. $\mathcal{N}$ ve $\mathcal{U}$ dağılımla elde edilen R vektörlerinin öneri üretme doğruluğu üzerine etkisi *RK* ve *RD* prosedürlerinde ayrı ayrı değerlendirilmiştir. Ayrıca σ_{max} ve β_{max} katsayılarının öneri doğruluğu üzerine etkilerini analiz etmek için yapılan deneylerde, kullanıcı davranışlarını taklit etmek için sistem tarafından belirlenen σ_{max} ve β_{max} parametreleri (0,1] değer aralığında rastgele üretilen bir sayı ile çarpılmaktadır. Böylece gerçek sistem kullanıcılarının bireysel ihtiyaçlarına göre belirlediği gizlilik düzeyleri taklit edilmektedir. Bu işlem sonucunda ortaya çıkabilecek rastlantısallığı ortadan kaldırmak için her bir deney seti 10 kez tekrar edilmiş ve elde edilen sonuçların ortalaması nihai doğruluk seviyeleri olarak sunulmuştur. Elde

edilen deneysel sonuçların başarısını ölçeklendirmek için, gerçek oy değerleri ve bu değerler için üretilen tahminler arasındaki ortalama mutlak farkları ölçen istatistiksel doğruluk ölçütü olarak *MAE* kullanılmaktadır.

4.4.3.1. RK yönteminin öneri doğruluğuna etkisi

RK yönteminde değişken σ_{max} parametresinin öneri üretme doğruluğuna etkisini değerlendirmek için bu parametre $[0{,}5; 5]$ değer aralığı içerisinde artan 0,5'lik değerler ile test edilmiştir. Veri maskeleme işlemi için oluşturulan rastgele sayı vektörü $\mathcal{N}$ ve $\mathcal{U}$ dağıma göre üretilmiştir. *RK* yönteminde β_{max} katsayısına sıfır sabit değeri atanmıştır. Böylece oy verilmemiş ürünlere herhangi bir maskeleme verisi eklenmeden sadece gerçek oy değerlerinin maskelenmesi ile oluşan öneri doğruluk kayıpları test edilmiştir.

YM5 veri seti üzerinde, *GKOF*^{k} yaklaşımına göre elde edilen rastgele sayı vektörünün gerçek kullanıcı oy vektörüne eklenmesi ile oluşan yeni derecelendirme vektöründe maskeleme işleminin öneri doğruluğu açısından karşılaştırılması Şekil 4.5'te gösterilmektedir.

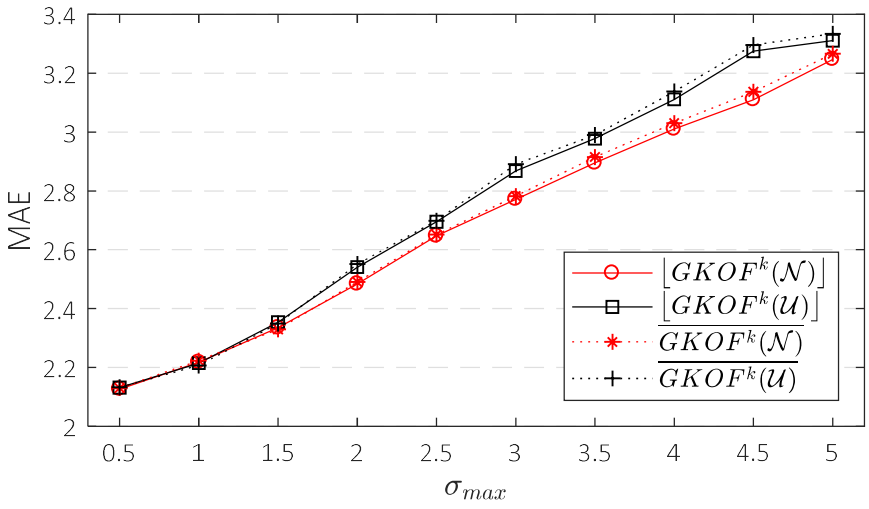


Şekil 4.5. *YM5 veri setinde σ_{max} katsayısının öneri üretme doğruluğuna etkisi*

Üretilen maskeleme verisinin dağılımları göz önünde bulundurulduğunda; $\mathcal{N}$ dağılıma göre maskelenen kullanıcı vektörü, tahmin üretme protokollerinden bağımsız olarak her bir değişken σ_{max} katsayısında $\mathcal{U}$ dağılımına göre daha düşük mutlak hata ile öneri üretilmesini sağlamaktadır. *GKOF*^{k} sistemlerinde benimsenen temel prensip veri gizliliğini sağlarken aynı zamanda da öneri üretme doğruluğundan mümkün olan en düşük seviyede kayıp vermektir. Şekil 4.4'te gösterilen gizlilik seviyeleri hatırlandığında;

σ_{max} katsayısının [0,5; 3] aralığında elde ettiği gizlilik seviyesi artış hızının 3'ten büyük değerler için yavaşladığı özellikle $\mathcal{U}$ dağılımında neredeyse sabitlendiği gözlemlenmektedir. Bu nedenle gizlilik ve öneri doğruluğunu dengelemek amacıyla *YM5* veri seti için ideal σ_{max} katsayısı 3 olarak belirlenmektedir. Böylece mahremiyet seviyesi mümkün olan en yüksek seviyede tutulurken öneri doğruluğundan makul düzeyde kayıp yaşanmaktadır.

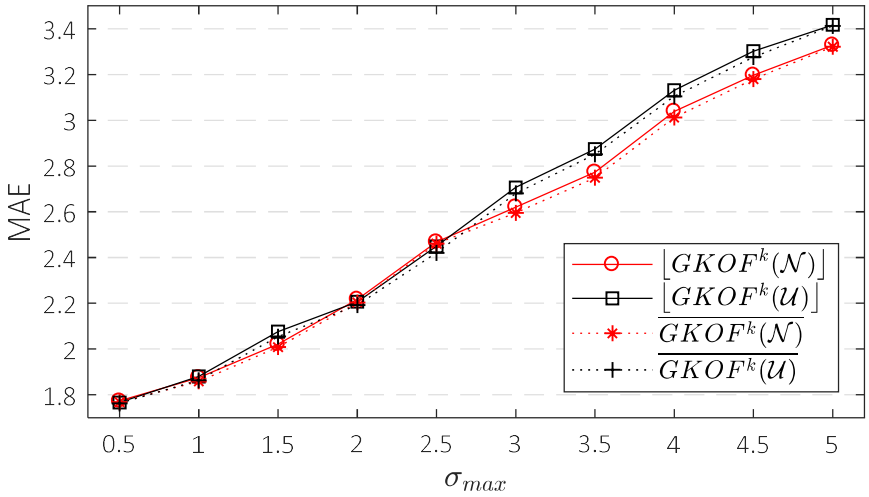
Kullanılan benzerlik tabanlı yaklaşımlar ($\lfloor GKOF^k \rfloor$, $\overline{GKOF^k}$) ve maskeleme vektörlerinin dağılımları göz önünde bulundurulduğunda; elde edilen *MAE* değerleri kabaca birbirine benzer eğilimler sergilemektedir. Ancak Şekil 4.4'te gösterilen gizlilik seviyeleri dikkate alındığında σ_{max} parametresi 3'e eşitken elde edilen gizlilik seviyelerinde $\mathcal{N}$ dağılım ile $\mathcal{U}$ dağılıma göre %23,38 oranında yüksek düzeyde veri gizliliği sağladığı görülmektedir. Dağılımların üretebildiği gizlilik seviyesi, elde edilen öneri doğruluğunda da değişkenlik göstererek en iyi mutlak hataların elde edildiği $\lfloor GKOF^k \rfloor$ yöntemi referans alınarak $\lfloor GKOF^k(\mathcal{N}) \rfloor$ yöntemi ile $\lfloor GKOF^k(\mathcal{U}) \rfloor$ yöntemine göre %3,47 oranında daha doğru öneriler üretilmesini sağlamaktadır. Özetle, *YM5* veri seti için en doğru öneriler $\lfloor GKOF^k(\mathcal{N}) \rfloor$ yaklaşımı ile elde edilmektedir.



Şekil 4.6. *YM10 veri setinde σ_{max} katsayısının öneri üretme doğruluğuna etkisi*

YM5 veri setine benzer şekilde *YM10* ve *YM20* veri setleri ile yapılan deneylerde, $\mathcal{N}$ dağılıma göre maskelenen veriler [0,5; 5] değer aralığı içerisinde değişken σ_{max} katsayılarında $\mathcal{U}$ dağılımına göre daha doğru öneriler üretilmesini sağlamaktadır. *YM10* ve *YM20* için σ_{max} katsayısının veri gizliliğine etkisini göz önünde bulundurduğumuzda,

Şekil **4.2** ve Şekil 4.3'te gösterildiği gibi ideal mahremiyet seviyesiyle Şekil 4.6 ve Şekil 4.7'de elde edilen öneri üretme doğruluğuna göre ideal σ_{max} katsayısı 3 olarak belirlenmiştir.



Şekil 4.7. *YM20 veri setinde σ_{max} katsayısının öneri üretme doğruluğuna etkisi*

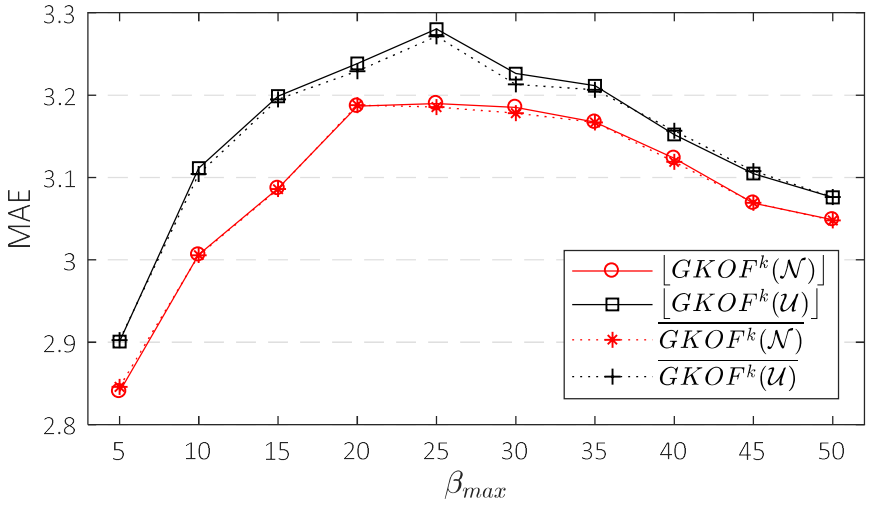
4.4.3.2. RD Yönteminin Öneri Doğruluğuna Etkisi

RD yöntemi ile oluşturulan *R* vektörünün öneri üretme doğruluğuna etkisini analiz etmek için doldurma vektörünün boyunu belirlemek için kullanılan β_{max} katsayı, [5,50] değer aralığı içerisinde test edilmiştir. *RD* yönteminde oluşturulan rastgele sayılar $\mathcal{N}$ ve $\mathcal{U}$ ile ayrı ayrı test edilmiştir. *RK* prosedürüne göre bütün veri setleri için ideal σ_{max} katsayısı 3 olarak belirlendiğinden, *RD* yönteminde bu değer kullanılmıştır. Elde edilen maskelenmiş kullanıcı oyları ile öneri üretme işlemi $\bar{\cdot}$ ve $\lfloor \cdot \rfloor$ benzerlik yaklaşımlarının her ikisine göre ayrı ayrı test edilmiştir. Son olarak, yapılan deneysel çalışma Yahoo!Movies veri setinin *YM5*, *YM10* ve *YM20* veri setlerinde ayrı ayrı değerlendirilmiştir.

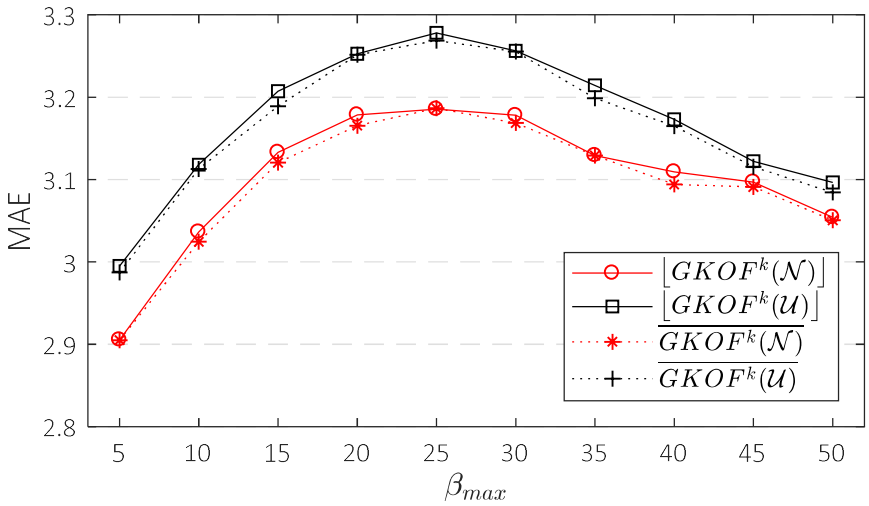
YM5 veri seti üzerinde, $GKOF^k$ yöntemi ile elde edilen rastgele sayıların kullanıcı tarafından belirlenen β katsayısına göre gerçek kullanıcı oy vektörüne eklenmesi ile oluşan yeni kullanıcı-ürün vektörünün öneri doğruluğu açısından karşılaştırılması Şekil 4.8'de gösterilmektedir.

Öneri üretme doğrulukları maskeleme verisi dağılımlarına göre incelendiğinde; $\mathcal{N}$ dağılıma göre maskelenen veriler, kullanıcı benzerliklerini belirleme protokolünden bağımsız olarak her bir değişken β_{max} katsayında $\mathcal{U}$ dağılımına göre daha doğru öneriler üretilmesini sağlamaktadır. Kullanıcı benzerliklerini elde etmek için kullanılan $\bar{\cdot}$ ve $\lfloor \cdot \rfloor$ göre elde edilen *MAE* değerleri incelendiğinde her iki dağılımda da $\lfloor GKOF^k \rfloor$ ve $\overline{GKOF^k}$

protokolleri ile üretilen öneri doğrulukları birbirine oldukça yakın değerlerdedir. Gizliliği koruyan *OF* sistemlerinde temel prensip kullanıcı verilerini doğru öneriler üretilebilecek kadar maskelemek, bir başka ifade ile gizlilik ve öneri doğruluğu arasında mutlak bir denge kurulmalıdır. Şekil 4.1'de gösterilen gizlilik seviyeleri hatırlandığında, *YM5* veri seti için ideal β_{max} katsayısı %5 olarak belirlenmektedir. Böylece mahremiyet seviyesi mümkün olan en yüksek seviyede tutulurken öneri doğruluğundan ideal düzeyde kayıp yaşanmaktadır.



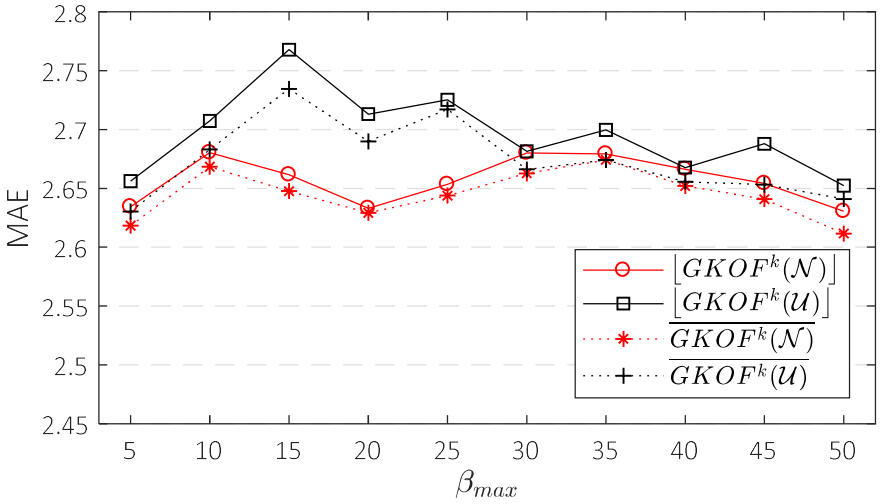
Şekil 4.8. *YM5 veri setinde β_{max} katsayısının öneri üretme doğruluğuna etkisi*



Şekil 4.9. *YM10 veri setinde β_{max} katsayısının öneri üretme doğruluğuna etkisi*

YM5 veri setine benzer şekilde *YM10* veri seti ile yapılan deneylerde, $\mathcal{N}$ dağılıma göre maskelenen veriler her bir değişken β_{max} katsayında $\mathcal{U}$ dağılımına göre daha doğru

öneriler üretilmesini sağlamaktadır. *YM10* için β_{max} katsayısının veri gizliliğine etkisini göz önünde bulundurduğumuzda, Şekil 4.1'de gösterilen gizlilik seviyesi ve elde edilen öneri üretme doğruluğu göz önünde bulundurulduğunda ideal β_{max} katsayısı %10 olarak belirlenmiştir.



Şekil 4.10. *YM20 veri setinde β_{max} katsayısının öneri üretme doğruluğuna etkisi*

$\mathcal{N}$ ve $\mathcal{U}$ dağılımlara göre elde edilen maskeleme verisinin, $\lfloor GKOF^k \rfloor$ ve $\overline{GKOF^k}$ benzerlik belirleme stratejilerine göre *YM20* veri seti ile elde edilen *MAE* sonuçları genel anlamda *YM10* ve *YM5* veri setleri ile elde edilen sonuçlara benzerlik göstermektedir. Ancak *YM20* veri setinde değişken β_{max} değerlerinin gizlilik üzerine etkisi incelendiğinde %50 seviyesine kadar veri gizliliğinin yeterince sağlanamadığı Şekil 4.1'de gösterilmiştir. Elde edilen *MAE* değerlerinde de %50 seviyelerinde en doğru önerilerin üretildiği gözlemlenmektedir. Bunun nedeni, *YM20* veri setinin olağan dışı seyreklik seviyesidir. Ortalama olarak bir kullanıcının 40 adet ürünü derecelendirdiği ve toplamda 247 adet üründen oluşan bir veri setinde β_{max} katsayısının %50 gibi yüksek değerlerde olması kullanıcı benzerliklerini etkilemekte ve normalde birbirinin komşuluğuna girmeyen kullanıcıların benzer derecelendirme profilleri oluşturmasına neden olmaktadır. Bunun sonucunda da elde edilen *MAE* değerlerinde azalma meydana gelmektedir.

4.5. Sonuçlar

Veri maskeleme prosedüründe $\mathcal{N}$ dağılım ile üretilen maskeleme verisinin kullanıcı gizliliğine katkısı $\mathcal{U}$ dağılıma göre bütün veri setleri üzerinde daha iyi sonuçlar vermektedir. Aynı zamanda, Şekil 4.5, Şekil 4.6 ve Şekil 4.7'de sunulan doğruluk

seviyeleri göz önünde bulundurulduğunda öneri üretme doğruluğu bakımından da $\mathcal{N}$ dağılım ile elde edilen maskeleme verisi ile karıştırılan kullanıcı verisi ile üretilen öneriler $\mathcal{U}$ dağılıma göre daha doğru sonuçlar vermektedir.

Önerilen $GKOF^k$ yaklaşımları, çoklu-ölçütlü kullanıcı derecelendirmeleri için yüksek seviyede gizlilik düzeyi sağlamada yararlı olsa da, maskelenmemiş ham kullanıcı verileriyle OF^k yaklaşımına göre üretilen öneri doğrulukları gizlilik koruma ortamındaki tahmini doğruluk kayıplarını gözlemlemek için kıyaslanmalıdır. Bu amaçla, maskelenmemiş gerçek kullanıcı derecelendirmeleri ile $\bar{\cdot}$ ve $\lfloor\cdot\rfloor$ benzerlik yaklaşımları kullanılarak öneri üretme işlemi gerçekleştirilerek gizlilik ve doğruluk ile ilgili çakışan hedeflere bağlı kayıplar gözlemlenmiştir. Bu karşılaştırma yapılırken *YM20* veri seti ideal gizlilik parametreleri $\sigma_{max} = 3$ ve $\beta_{max} = 50$, *YM10* veri seti için $\sigma_{max} = 3$ ve $\beta_{max} = 5$ ve *YM5* veri seti için $\sigma_{max} = 3$ ve $\beta_{max} = 5$ olarak kabul edilmiştir. Maskelenmemiş ham veri seti ile elde edilen öneri doğrulukları ve ideal maskeleme parametreleri ile maskelenen kullanıcı derecelendirmelerinden elde edilen öneri doğrulukları Tablo 4.1'de gösterilmektedir.

Tablo 4.1. *İdeal σ_{max} ve β_{max} seviyeleriyle elde edilen doğruluk seviyelerinin OF^k ile kıyaslanması*

Veri Seti	*Dağılım*	$\overline{OF^k}$	$\overline{GKOF^k}$	$\lfloor OF^k \rfloor$	$\lfloor GKOF^k \rfloor$
YM20	$\mathcal{U}$	**1,745**	2,641	**1,739**	2,652
	$\mathcal{N}$		**2,612**		2,631
YM10	$\mathcal{U}$	**2,104**	2,987	**2,099**	2,995
	$\mathcal{N}$		2,905		**2,905**
YM5	$\mathcal{U}$	**2,184**	2,903	**2,180**	2,901
	$\mathcal{N}$		2,846		**2,840**

Elde edilen en doğru sonuçlar göz önünde bulundurulduğunda *YM5*, *YM10* ve *YM20* veri kümeleri için doğruluk kayıpları % 23,24; % 27,75 ve % 33,19'dur. OF^k sistemleri için elde edilen kullanıcı gizliliği düzeylerinin karşılığında kaçınılmaz doğruluk kayıplarının ortaya çıkacağı kabul edilen bir gerçektir. Ancak, elde edilen doğruluk kayıpları makul ölçüde ve kabul edilebilir sınırlarda olmadığı görülmektedir. Geleneksel *RK* ve *RD* yaklaşımlarının direkt olarak çoklu-ölçütlü veri setlerine uyarlanması yüksek seviyede gizlilik sağlarken öneri üretme kalitesini negatif yönde etkilemektedir. Bu nedenle, OF^k sistemlerinde veri gizliliğini sağlarken aynı zamanda da daha doğru öneriler üretebilecek yeni yaklaşımlara ihtiyaç duyulmaktadır.

5. ENTROPİ TABANLI GİZLİLİĞİ KORUYAN ÇOKLU-ÖLÇÜTLÜ ORTAK FİLTRELEME

Çoklu-ölçütlü tercih verileri kullanımı, bir öğenin kullanıcı tarafından neden tercih edildiğini anlama ve kullanıcı profillerini daha ayrıntılı analiz etme fırsatı sunsa da, bireyleri daha ciddi mahremiyet tehditleri ile karşı karşıya getirmektedir. Geleneksel yaklaşımların, çoklu-ölçütlü tercih derecelendirmelerine uyarlandığı $GKOF^k$ sistemleri kullanıcı mahremiyet problemini hafifletecek çözümler sunmakla birlikte kullanılan veri maskeleme protokolleri, kullanıcı gizliliğini sağlarken aynı zamanda da sistemin öneri kalitesinde telafi edilemeyecek kadar yüksek seviyelerde bozulmalara neden olmaktadır. Bu nedenle, elde edilen gizlilik seviyeleri ile tahmin doğruluğu arasında bir denge kurmak için, entropi tabanlı gizliliği koruyan çoklu-ölçütlü ortak filtreleme yaklaşımı geliştirilmiştir.

Çoklu-ölçütlü veri alanında, tüm ölçütler kullanıcıları için aynı düzeyde önem seviyesine sahip değildir. Ölçütler içerisinde kullanıcısı için daha çok önem taşıyan dolayısıyla genel beğeni profilini daha iyi yansıtan ölçütler OF^k sistemlerinin öneri kalitesini arttırmak adına daha yüksek seviyede önem arz etmektedir. Ayrıca, kullanıcıların ürün derecelendirme eğilimlerinde, alt-ölçütler ve genel beğeni derecesi arasında bir bağımlılık söz konusudur. Bu nedenle, her bir alt-ölçüte ait tercih vektörünün genel beğenme üzerinde farklı etkileri olduğu iddia edilebilir, bu da farklı önem derecesine sahip alt-ölçütlerin farklı gizlilik seviyelerinde maskelenmesini motive eder. Böylece, elde edilen gizlilik seviyeleri ve öneri üretme doğruluğu arasında bir denge kurulabilmektedir.

5.1. Amaç ve Kapsam

OF^k sistemleri çoklu-ölçüt kullanımı ile ortaya çıkan yeni gizlilik tehditlerini hafifletmek için her bir ölçütü, kullanıcının ölçütlere verdiği önem derecesine ve derecelendirme profiline göre ayrı ayrı değerlendirebilen değişken koruma mekanizmalarına ihtiyaç duyulmaktadır. Buna ek olarak, gerçek kullanıcı derecelendirmelerinin rastgele sayılardan oluşan maskeleme vektörleri ile gizlenmesinden kaynaklanan doğruluk kayıplarını azaltabilecek yeni gizlilik koruma protokollerine ihtiyaç duyulmaktadır.

Hedeflenen amaç doğrultusunda, bu bölümde maskeleme işlemi sırasında ortaya çıkacak öneri doğruluk kayıplarını hafifletmeyi hedefleyen, kullanıcı ve kullanıcının

ölçütleri değerlendirme alışkanlıklarına göre maskeleme işleminde kullanılacak gizlilik parametrelerini belirleyen entropi tabanlı yeni bir gizlilik-koruma şeması önerilmektedir. Önerilen protokol, her bir kullanıcı ve kullanıcı derecelendirmelerinden oluşan her bir ölçüt için gizlilik koruma parametrelerinin dinamik olarak kontrol edilmesini sağlayan ve bunu sağlarken kullanıcının belirlediği gizlilik seviyesinden ödün vermeyen entropi tabanlı veri maskeleme yaklaşımıdır. Önerilen protokol temel olarak iki maskeleme stratejisinden oluşmaktadır. Bunlar;

(i) gelişmiş doğruluk düzeyine ulaşmak için daha az bilgi içeren bir başka ifade ile entropi değeri yüksek olan ölçütü daha düşük gizlilik seviyesine sahip rastgele sayı vektörü ile maskelemek,

(ii) kullanıcı gizliliğini daha fazla korumak için daha az bilgi içeren yani entropi değeri yüksek olan ölçütü daha yüksek gizlik seviyesine sahip rastgele sayı vektörü ile maskelemektir.

Önerilen gizlilik koruma şemalarının hem kullanıcı gizliliğine getirdiği farklılıkları hem de üretilen tahmin doğruluğuna olan etkilerini göstermek için, Yahoo!Movies çoklu-ölçütlü tercih veri kümesinin üç alt kümesi olan *YM5*, *YM10* ve *YM20* veri setleri deneysel olarak değerlendirilmektedir. Elde edilen sonuçlara göre, önerilen entropi tabanlı gizlilik koruma şeması, geleneksel gizlilik koruma senaryosu olan $GKOF^k$ yaklaşımının sağladığı özdeş bir gizlilik seviyesini muhafaza ederken önemli ölçüde daha doğru tahminler üretebilmektedir. Buna ek olarak, yeni entropi tabanlı gizlilik koruma şemasının, öneri doğruluğunu ciddi bir şekilde ihlal etmeden kullanıcı mahremiyetini koruyabildiği de ortaya konmaktadır.

5.2. Entropi Tabanlı Veri Karıştırma

Bilgi teorisinde entropi, belirli bir veri kaynağındaki belirsizliğin miktarı olarak tanımlanabilir (Shannon, 1948). Bir örneklemin yüksek entropi seviyesine sahip olması, o örneklemi oluşturan olaylar kümesinin görülme sıklığındaki rastlantısallıkla ilişkilidir. Bu tanımı çoklu-ölçüt alanında ele aldığımızda, bir kullanıcının önceden tecrübe ettiği ürünleri hep aynı tercih değerleri ile derecelendirmek yerine, tercihlerinde farklı derecelendirmelere yer vermesi o sistemin entropi değerinin artmasına neden olmaktadır. Kullanıcı derecelendirmelerindeki bu farklılık, gerçek kullanıcı derecelendirmelerinin öngörülebilirliğini azaltacaktır. Böylece, çoklu-ölçütlü veri kümesinde elde edilen gizlilik düzeyini beraberinde artıracaktır. Bahsi geçen bu önermeyi bir örnekle açıklamak

gerekirse; farklı oy dağılımlarına ve oy verme eğilimlerine sahip A ve B kullanıcılarına ait derecelendirme vektörlerinde kullanılan oy değerlerinin olasılık dağılımları aşağıdaki gibi olsun.

$$f(a) = \begin{cases} 0{,}5; & 0 \le a \le 1 \\ 0{,}5; & 4 \le a \le 5 \\ 0; & aksi\ halde \end{cases} \qquad f(b) = \begin{cases} 0{,}25; & 0 \le b \le 1 \\ 0{,}25; & 2 \le b \le 3 \\ 0{,}5; & 4 \le b \le 5 \\ 0; & aksi\ halde \end{cases}$$

Bu kullanıcılara ait derecelendirmelerin olasılık dağılımı ile gizlilik seviyesi arasındaki korelasyonu göstermek için Agrawal ve Aggarwal (2001) tarafından sunulan, bir rastgele değişkenin diferansiyel entropisine dayalı gizlilik ölçeklendirme yöntemi kullanılmıştır. Burada, rastgele bir olay olarak ifade edilen *X*'in diferansiyel entropisi Denklem 5.1'de verilmiştir.

$$h(X) = -\int_{\Omega X} f_x(x) log_2 f_x(x) dx \qquad \textbf{(5.1)}$$

Burada Ω_x X'in tanım kümesi olarak gösterilirken, $h(X)$ X olayının belirsizlik ölçüsü olarak tanımlanmaktadır. Sürekli dağılıma sahip bir olayda, gözlemlenen olasılık dağılımının gizlilik seviyesini ölçmek için Denklem 5.2 kullanılmaktadır (Agrawal ve Aggarwal, 2001).

$$\prod(X) = 2^{h(X)} \qquad \textbf{(5.2)}$$

Verilen örnekte A ve B kullanıcılarına ait derecelendirme vektörlerinin diferansiyel entropisi aşağıdaki gibi hesaplanmaktadır.

$$h(X) = -\int_{\Omega X} f_x(x) log_2 f_x(x) dx$$

$$h(A) = -\int_0^1 0{,}5\ log_2\ 0{,}5\ d_a - \int_4^5 0{,}5\ log_2\ 0{,}5\ d_a = 1$$

$$h(B) = -\int_0^1 0{,}25\ log_2\ 0{,}25\ d_b - \int_2^3 0{,}25\ log_2\ 0{,}25\ d_b - \int_4^5 0{,}5\ log_2\ 0{,}5\ d_b = 3/2$$

Diferansiyel entropisi elde edilen *A* ve *B* kullanıcı vektörlerine ait gizlilik seviyeleri;

$$\prod(A) = 2^1 = 2$$

$$\prod(B) = 2^{3/2} = 2{,}83$$

olarak ölçeklenmiş olur.

Elde edilen gizlilik seviyeleri göz önünde bulundurulduğuna, yüksek entropi değerine sahip olan B kullanıcısına ait derecelendirme vektörünün daha yüksek düzeyde mahremiyete sahip olduğu görülmektedir. Özetle; derecelendirme vektöründen elde edilen diferansiyel entropi değeri, sistemin gizlilik seviyesi ile doğrudan ilişkilidir.

$GKOF^k$ ve geleneksel *GKOF* sistemlerinde veri gizleme işleminde kullanıcı gizliliği, kullanıcının bireysel ihtiyaçlarına göre tercih ettiği σ parametresine göre gerçekleştirilmektedir. Ancak, her bir ölçüt farklı entropi değerlerine sahip olabileceğinden, her ölçüt değişken dağılım aralığına göre maskelenmelidir. Çünkü geleneksel yaklaşımda tanımlandığı gibi her bir ölçüt aynı σ değerine göre gizlendiğinde, bazı ölçütler ihtiyaç duyulduğu seviyeden daha yüksek seviyede maskelenmiş olacak ve sonucunda öneri doğruluğundan kayıplar meydana gelecektir. σ parametresinin belirlenme sürecinde göz önünde bulundurulması gereken en önemli koşul, doğruluk ve gizlilik arasında en uygun dengeyi kurmaktır. Bunu elde etmek için göz önünde bulundurulması gereken durumlardan ilki; daha yüksek entropi değerine sahip bir ölçütün büyük σ katsayısı ile maskelenmesi gizliliği artırsa da tahmin üretme kalitesini olumsuz yönde etkileyebilir. Aksine, ikinci durumda daha düşük entropi değerine sahip alt-ölçütün daha az maskelemesi, veri kümesinin gizliliği düzeyinde önemsiz bir kayıpla doğruluğu daha yüksek öneriler üretmesini sağlayabilir.

Bu durumların ortaya çıkabileceği kullanıcı gizliliği ve öneri doğruluğu senaryolarını değerlendirmek için yapılan çalışmada iki farklı maskeleme yöntemi önerilmiştir.

- S_σ olarak adlandırılan yaklaşımda, en büyük entropi değerine sahip ölçüt, en düşük σ değeriyle elde edilen rastgele sayı vektörü tarafından maskelenmektedir. Yüksek entropi değerine sahip bir ölçüt, örneklemi oluşturan derecelendirme kümesinin görülme sıklığı göz önünde bulundurulduğunda olarak oldukça rastlantısaldır. Bu nedenle, mahremiyetini korumak için yüksek σ değerleri ile maskelenmesi gerekmeyip, sahip olduğu entropi değerine ters orantılı olarak daha küçük bir σ değeriyle maskelenmektedir. Bunun sonucunda, derecelendirmelerin görülme olasılığından dolayı zaten yüksek seviyede entropi değerine sahip olan bir ölçüt yapısı gereği yüksek seviyede gizlilik sergilemekte ve büyük σ değerleri ile maskelenmeyerek sistemin öneri üretme doğruluğu tehlikeye atılmamaktadır.

- S^{σ} olarak adlandırılan yaklaşımda, en büyük entropi değerine sahip ve en az bilgi içeren ölçüt, en yüksek düzeydeki σ değeriyle maskelenmektedir. İlk yaklaşımın tersine, bu yöntem daha yüksek bir σ katsayısı ile daha yüksek entropi değerine sahip ölçütü maskelemektedir. Böylece sistem, mahremiyet düzeyini daha yüksek dereceye çıkarabilmektedir.

Sunulan entropi tabanlı rastgele veri karıştırma prosedüründe S_{σ} ve S^{σ} yaklaşımlarının ortak ana adımları aşağıdaki maddeler altında özetlenebilir.

i. Her bir kullanıcıya ait ölçütlerin entropisi ayrı ayrı hesaplanır.
ii. Elde edilen entropi değerlerinin ortalaması 1'e eşit olacak şekilde normalleştirilir (S_N).
iii. Ölçütlere ait normalleştirilmiş entropi değerleri S_{σ} veya S^{σ} yaklaşımlarına göre sıralanır, böylece σ_k katsayıları elde edilir.
iv. Elde edilen σ_k katsayıları, her bir ölçüt için σ katsayısı olarak kullanılır ve rastgele sayılar elde edilen nihai σ değerlerine göre üretilir.

Önerilen yöntemlerin çalışma prensiplerini göstermek üzere Tablo 5.1'de kullanıcı derecelendirmelerine ait $r_1, \dots, r_5$ olarak isimlendirilen beş ölçütün entropi katsayıları ve $S_{\sigma} - S^{\sigma}$ yaklaşımlarına göre elde edilen σ_k katsayılarına ait bir örnek verilmiştir.

Tablo 5.1. *S_{σ} ve S^{σ} için örnek σ katsayıları*

	k_1	k_2	k_3	k_4	k_5
S	1,5	1	3	2,5	2
S_N	0,75	0,5	1,5	1,25	1
S_{σ}	1,25	1,5	0,5	0,75	1
S^{σ}	0,75	0,5	1,5	1,25	1

Tablo 5.1'de verilen örnekte görüldüğü üzere, öncelikle her bir ölçüt için kullanıcı derecelendirme vektörünün entropi değerleri elde edilmiştir. Elde edilen entropi değerleri kullanılarak, kullanıcının belirlediği σ değerine göre bütün ölçütlere ait σ katsayılarının ortalamaları 1'e eşit olacak şekilde normalleştirilir. Böylece kullanıcı tarafından belirlenen gizlilik seviyesinden çok fazla ödün vermeden, ölçütlere sahip oldukları bilgi miktarına göre değişken σ katsayıları uygulanarak kullanıcı mahremiyeti korunmuş olur. Dolayısıyla, bir ölçüt daha yüksek bir σ değeriyle gizlenirken, diğerleri mahremiyet ve doğruluk arasında bir denge kurmak için daha küçük σ değerine göre maskelenir. Kullanıcı tarafından $\sigma = \sigma_{max}$ olarak belirlendiği durumlarda, en yüksek seviyede

gizlenecek olan ölçüt sistemin maksimum gizlilik seviyesinin üzerinde maskelenmektedir. Entropi tabanlı gizliliği koruma yaklaşımları Prosedür 5.1'de gösterilmektedir. S_σ ve S^σ yaklaşımları için oluşturulan maskeleme prosedüründe *GKOF*^{k} yaklaşımına ek olarak, ölçütlere ait entropi katsayılarının hesaplanıp her bir ölçüt için sahip oldukları entropi derecesine göre σ katsayılarının belirlendiği fonksiyon eklenmiştir.

Prosedür 5.1. *S_σ ve S^σ yaklaşımlarında veri maskeleme prosedürü*

Require: Kullanıcı × ölçüt × ürün vektörü (G_k), σ_{max}, β_{max}

z-skoru değerinin hesaplanması ($\rightarrow \boldsymbol{Z_k}$)

1 : **for all criteria in** G $(i \leftarrow 1\ to\ k)$ **do**

2 : $\overline{G_i} \leftarrow MEAN(G_i)$

3 : $\sigma_{G_i} \leftarrow STD(G_i)$

4 : **end for**

5 : **for all criteria in** G $(i \leftarrow 1\ to\ k)$ **do**

6 : **for all items in** G_i $(i \leftarrow 1\ to\ m)$ **do**

7 : $Z_{ij} = (G_{ij} - \bar{G}_i)/\ \sigma_{G_i}$

8 : **end for**

9 : **end for**

10 : **for all criteria in** G $(i \leftarrow 1\ to\ k)$ **do**

11 : $Ent_{G_i} = \sum_{j=1}^{r} p_j log\ (p_j)$

12 : **end for**

Maskeleme yaklaşımının ve entropi katsayılarının belirlenmesi

13 : $EntCoef_{G_k} = sort(Ent_G, EntMax|EntMin)$

Gizlilik parametrelerinin belirlenmesi

14 : $\beta \leftarrow RND\ (0, \beta_{max}]$

15 : $\sigma_{1\rightarrow k} \leftarrow RND\ (0, \sigma_{max}] \times EntCoef_{1\rightarrow k}$

16 : $\alpha_{1\rightarrow k} \leftarrow \sqrt{3}\sigma_{1\rightarrow k}$

17 : $e \leftarrow |E|$ ► # oy verilmemiş ürün

18 : $g \leftarrow |G|$ ► # gerçek kullanıcı oyları

19 : $F \leftarrow e \times \beta\%$ ► # doldurulacak hücre sayısı

Dağılımın belirlenmesi ve rastgele sayıların türetilmesi

20 : $dist \leftarrow RANDOM\ (uniform,\ normal)$

21 : **for all criteria in** $G\ (i \leftarrow 1\ to\ k)$ **do**

22 : $R_i \leftarrow dist(g + F;\ \mu = 0, \sigma_i |\ \alpha_i)$

23 : **end for**

***z*-skoru değerlerinin maskelenmesi** $(\rightarrow \boldsymbol{Z'})$

24 : **for all criteria in** $G\ (i \leftarrow 1\ to\ k)$ **do**

25 : **for all items in** $G_i\ (i \leftarrow 1\ to\ m)$ **do**

26 : $Z'_{ij} = (Z_{ij} + R_{ij})$

27 : **end for**

28 : **end for**

29 : **return** $\boldsymbol{Z'}$

5.3. Entropi Tabanlı Yaklaşımların Gizlilik ve Doğruluk Analizi

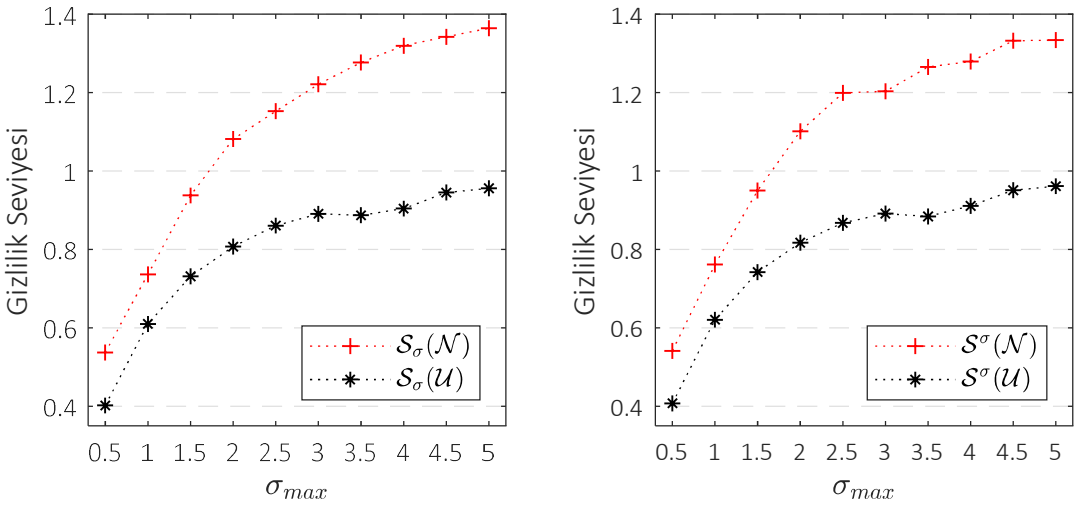
Bu bölümde öncelikle, önerilen entropi tabanlı S_σ ve S^σ yaklaşımlarının kullanıcı derecelendirme vektörlerine sağladığı gizlilik seviyeleri ve maskeleme işleminin öneri üretme doğruluğu üzerine etkisi değişken σ_{max} parametreleri kullanılarak analiz edilmiştir. Sonrasında, S_σ ve S^σ yaklaşımlarının geleneksel *GKOF*^{k} yaklaşımı ile gizlilik ve öneri doğruluğu açısından karşılaştırılması yapılmıştır. Son olarak, S_σ ve S^σ yaklaşımları ile maskelenen kullanıcı derecelendirmeleri ile üretilen önerilerin istatistiksel anlamlılıkları analiz edilmiştir.

Deneysel değerlendirmelerde öneri üretme süreci *GKOF*^{k} yaklaşımında da kullanılan birini dışarıda bırakarak çapraz doğrulama metodu kullanılarak oluşturulmuştur. Derecelendirme vektörünü maskelemek için oluşturulan R_i vektörlerinde, kullanıcı davranışlarını taklit etmek için [0,1] değer aralığı içerisinde rastgele bir sayı ile çarpılan σ katsayısının ve üretilen maskeleme vektörünün diziliminde ortaya çıkan rastlantısallıktan doğan farklılıkları ortadan kaldırmak için her deney seti aynı koşullar altında 10 kez tekrarlanmış. Sunulan nihai gizlilik ve doğruluk değerleri yinelenen tekrarlı deneylerde elde edilen sonuçların ortalamalarından oluşmaktadır. Ayrıca, yapılan deneysel çalışmalarda σ parametresinin kullanıcı gizliliği ve doğruluğu üzerindeki etkisini test ederken β_{max} parametresinin 0 değerinde olduğu varsayılmıştır ve σ_{max} parametresi $[0{,}5;\ 5]$ aralığında 0,5 aralıklarla artan değerlerle test edilmiştir. *RK* prosedürüyle oluşturulan R vektörü, $\mathcal{N}$ ve $\mathcal{U}$ dağılıma göre ayrı ayrı test edilmiştir.

Deneysel çalışmaların detayları ve yapılan testlerin gizlilik ve doğruluk üzerine etkileri sonraki bölümlerde ayrıntılı olarak incelenmektedir.

5.3.1. S_σ ve S^σ yaklaşımlarının kullanıcı gizliliğine etkisi

$\mathcal{N}$ ve $\mathcal{U}$ dağılımlara göre elde edilen rastgele gürültü vektörünün değişken σ_{max} seviyelerinde elde ettiği gizlilik seviyeleri Bölüm 4.2.2'de ifade edildiği gibi $\prod(V|P)$ fonksiyonu ile ölçeklenmektedir. Elde edilen gizlilik seviyeleri *YM5*, *YM10* ve *YM20* veri setleri için analiz edilmiştir. İlk olarak; *YM5* veri seti için S_σ ve S^σ yaklaşımları ile $\mathcal{N}$ ve $\mathcal{U}$ dağılımlara göre elde edilen maskeleme vektörünün [0,5;5] değer aralığı içerisinde elde ettiği gizlilik sevileri Şekil 5.1'de gösterilmektedir. Değişken σ_{max} değeri için $\mathcal{N}$ ve $\mathcal{U}$ dağılım ile elde edilen maskeleme verilerinin gizliliğe sağladığı katkısı incelendiğinde, $\mathcal{N}$ dağılım ile üretilen maskeleme verisi S_σ ve S^σ yaklaşımlarının her ikisinde de $\mathcal{U}$ dağılıma göre daha iyi sonuçlar elde edildiği görülmektedir. S_σ protokolünde $\mathcal{N}$ dağılıma sahip maskeleme verisi $\mathcal{U}$ dağılıma göre kullanılan bütün σ_{max} katsayıları için ortalama olarak %36,21 oranında gizliliği arttırmaktadır. Bu artış S^σ protokolünde %35,40 seviyesinde gözlemlenmektedir.

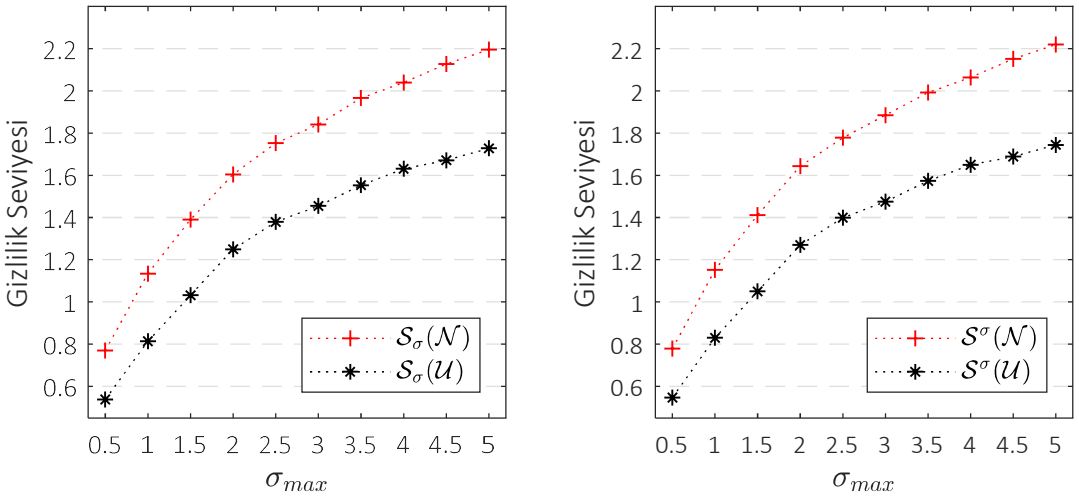


(a) (b)

Şekil 5.1. *S_σ ve S^σ için YM5 veri setinde değişken σ_{max} değerinin gizliliğe etkisi*

S_σ ve S^σ protokolleri kullanılarak oluşturulan maskeleme verisiyle *YM10* veri seti için elde edilen gizlilik seviyeleri Şekil 5.2'de gösterilmektedir. Elde edilen sonuçlara göre; $\mathcal{N}$ dağılımla üretilen rassal sayıların gizliliğe katkısı $\mathcal{U}$ dağılıma göre oldukça yüksektir. Gözlemlenen gizlilik seviyeleri dağılım bazında kıyaslandığında; S_σ

protokolünde $\mathcal{N}$ dağılıma sahip maskeleme verisi $\mathcal{U}$ dağılıma göre ortalama olarak %30,52 oranında gizliliği arttırırken, S^{σ} protokolünde %30,62 oranında bir artış söz konusudur.



(a) (b)

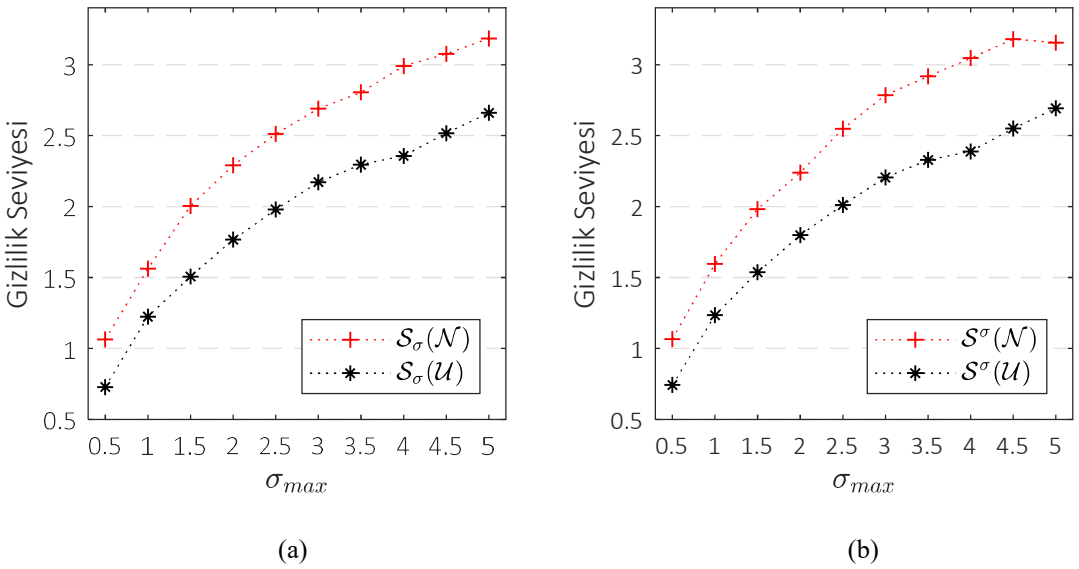
Şekil 5.2. *S_{σ} ve S^{σ} için YM10 veri setinde değişken σ_{max} değerinin gizliliğe etkisi*

Son olarak *YM20* veri seti için S_{σ} ve S^{σ} yöntemleri ile $\mathcal{N}$ ve $\mathcal{U}$ dağılımlara göre elde edilen rassal sayıların [0,5;5] değer aralığı içerisindeki gizlilik seviyeleri Şekil 5.3(a,b)'de verilmektedir. *GKOF*^{k} yönteminde de gözlemlendiği üzere artan σ_{max} değeri ile elde edilen gizlilik seviyesi arasındaki doğru orantı S_{σ} ve S^{σ} yöntemleri ile elde edilen gizlilik seviyelerinde de gözlemlenmektedir. Bununla birlikte $\mathcal{N}$ ve $\mathcal{U}$ dağılımlara göre elde edilen gizlilik seviyeleri karşılaştırıldığında, $\mathcal{N}$ dağılımla üretilen maskeleme verisinin gizliliğe katkısı $\mathcal{U}$ dağılıma göre her iki yöntemde de daha fazladır. Değişken σ_{max} değerlerine göre dağılımların gizliliğe etkisi bütün σ_{max} değerleri göz önünde bulundurularak kıyaslandığında; S_{σ} protokolünde $\mathcal{N}$ dağılım ile elde edilen maskeleme verisinin gizliliği katkısı % 27,79 olarak hesaplanırken, S^{σ} yönteminde % 27,40 olarak hesaplanmıştır.

YM5, *YM10* ve *YM20* veri setlerinde dağılımların kullanıcı gizliliğine katkısı göz önünde bulundurulduğunda $\mathcal{N}$ dağılım ile üretilen maskeleme verisi her koşulda $\mathcal{U}$dağılım ile üretilen maskeleme verisine göre daha yüksek seviyede gizlilik sağlamaktadır. Bunun nedeni, $\mathcal{N}$ dağılım ile üretilen rastgele sayı vektörünün gerçek kullanıcı oy verme eğilimini daha iyi taklit edebilmesidir.

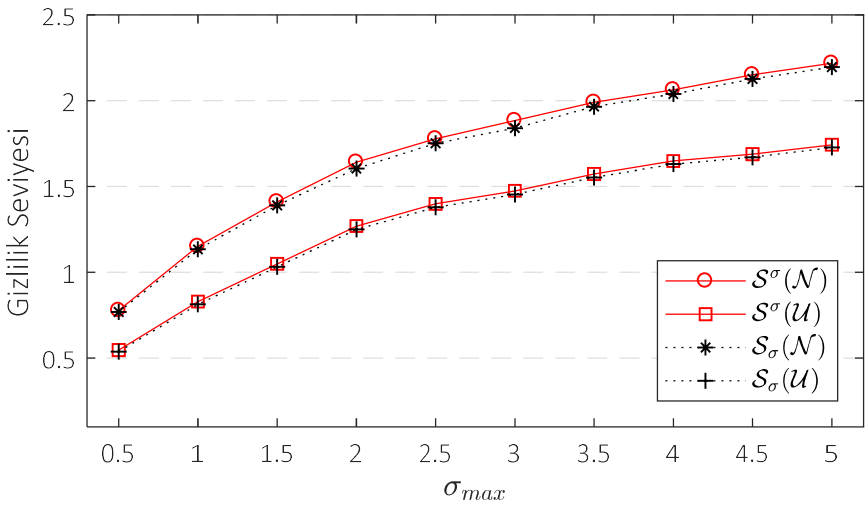
Elde edilen gizlilik seviyeleri her bir veri setinde benzer eğilimler sergilemektedir. Bu nedenle S_{σ} ve S^{σ} yaklaşımları ile elde edilen gizlilik seviyeleri Şekil 5.4'te yalnızca *YM10* veri seti için gösterilmektedir. Elde edilen sonuçlarda S^{σ} yaklaşımında $\mathcal{N}$ dağılım

ile oluşturulan maskeleme verisi ile elde edilen gizlilik seviyeleri S_σ yaklaşımı ile elde edilene göre %1,49 oranında daha yüksek seviyede gizlilik sağlamaktadır. Bunun nedeni, S^σ yaklaşımda en yüksek entropi değerine sahip ölçütün en büyük σ katsayısı ile oluşturulan maskeleme verisi ile gizlenmesiyle ortaya çıkan gizlilik seviyesi artışıdır.



(a) (b)

Şekil 5.3. *S_σ ve S^σiçin YM20 veri setinde değişken σ_{max} değerinin gizliliğe etkisi*



Şekil 5.4. *YM10 veri seti için S_σ ve S^σ yaklaşımları ile elde edilen gizlilik seviyeleri*

5.3.2. S_σ ve S^σ yaklaşımlarının öneri üretme doğruluğuna katkısı

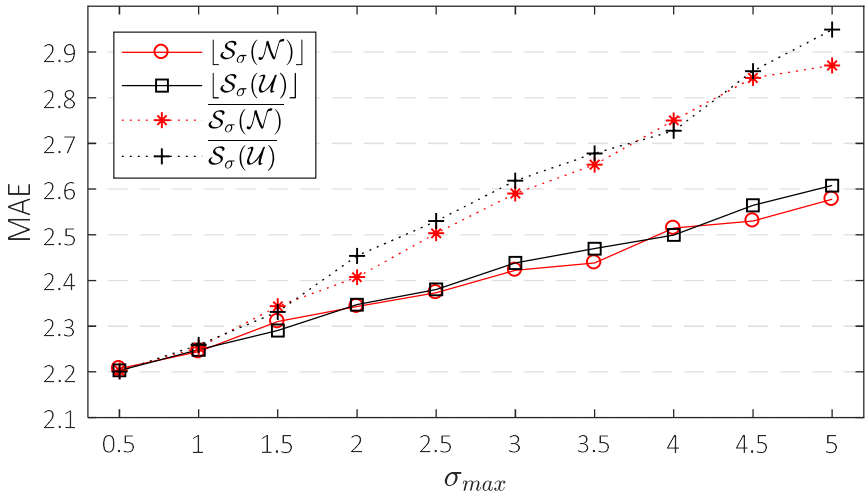
Gizliliği korumak adına uygulanan S_σ ve S^σ protokollerinde kullanılan değişken σ_{max} parametrelerinin öneri doğruluğu üzerine etkisini test etmek için farklı seyreklik seviyelerine sahip *YM20*, *YM10* ve *YM5* veri setleri üzerinde geleneksel *k*-en yakın

komşuluk tabanlı OF^k algoritmaları temel alınarak elde edilen öneri doğrulukları kıyaslanmaktadır. S_σ ve S^σ yaklaşımlar ile gizlenen kullanıcı derecelendirmeleriyle öneri üretmek için Bölüm 4.3'te açıklanan öneri üretme yaklaşımı kullanılmaktadır. Öneri üretme sürecinde, birini dışarıda bırakarak çapraz doğrulama yöntemi kullanılarak Bölüm 2.1.2'de tanımlanan $\bar{\cdot}$ ve $\lfloor\cdot\rfloor$ benzerlik elde etme metodolojilerine göre ürün önerileri üretilmektedir.

Prosedür 5.1'e göre yapılan *RK* işlemi ile elde edilen *R* vektörü $\mathcal{N}$ ve $\mathcal{U}$ dağılımları için ayrı ayrı değerlendirilmiştir. Değişken $\sigma_{\max}$ parametresinin öneri üretme doğruluğuna etkisini değerlendirmek için bu parametre $[0{,}5; 5]$ değer aralığı içerisinde artan 0,5'lik değerler ile test edilmiştir. *RK* prosedürü uygulanırken kullanıcının derecelendirilmemiş ürün listesini gizlemek için oluşturulan maskeleme vektörünün büyüklüğünü belirleyen β_{max} katsayısına sıfır sabit değeri atanmıştır. Bunun nedeni; derecelendirilmemiş ürünlere herhangi bir gizleme işlemine tabi tutulmayarak sadece gerçek kullanıcının derecelendirmelerinin maskelenmesiyle oluşan öneri doğruluk kayıplarının test edilmesidir. $GKOF^k$ işlemine benzer bir şekilde, kullanıcı tarafından belirlendiği varsayılan σ parametresinin rassallığını taklit edebilmek için sistem tarafından belirlenen σ_{max} katsayısı (0,1] değer aralığında rastgele üretilen bir sayı ile çarpılmaktadır. Böylece gerçek kullanıcı davranışları taklit edilmektedir. Ayrıca öneri üretilecek q nesnesi için derecelendirme değeri *kullanıcı* × *ürün* matrisinden silinmektedir. Elde edilen derecelendirme vektörü V' kullanılarak yeni P' vektörünü elde etmek için; V' vektörü üzerine Prosedür 5.1'e göre yeniden derlenen R' vektörü oluşturulmaktadır. Sonunda nihai P' vektörü; $P' = V' + R'$ işlemi ile elde edilmektedir. Yeni V' vektörüne göre maskeleme vektörünün tekrar oluşturulma nedeni, q nesnesine ait derecelendirme değerinin, *kullanıcı* × *ürün* matrisinde ortalama ve standart sapma değerlerine etkisini ortadan kaldırmaktır. Ayrıca R' vektörünü oluşturan rastgele sayıların diziliminden kaynaklanabilecek doğruluk sapmalarını azaltmak için her bir deney seti 10 kez tekrar edilmiş ve sonuç olarak bu testlerin ortalama değerleri sunulmuştur. Son olarak, üretilen önerilerin doğruluk derecesini değerlendirmek için, gerçek oy değerleri ve bu değerler için üretilen tahminler arasındaki ortalama mutlak farkları ölçen istatistiksel doğruluk ölçütü olarak *MAE* kullanılmaktadır.

5.3.2.1. S_σ öneri doğruluğu

S_σ yaklaşımı ile elde edilen maskelenmiş kullanıcı oyları ile öneri üretme işlemi $\lfloor S_\sigma \rfloor$ ve $\overline{S_\sigma}$ benzerlik yaklaşımlarının her ikisine göre ayrı ayrı test edilmiştir. *YM5* veri seti için değişken σ_{max} katsayılarına göre gerçek kullanıcı derecelendirmelerinin maskelenmesi ile oluşan yeni *kullanıcı* × *ürün* vektörünün öneri doğruluğu açısından karşılaştırılması Şekil 5.5'te gösterilmektedir. Elde edilen öneri doğrulukları kullanılan dağılımlara göre incelendiğinde; $\mathcal{N}$ dağılıma göre oluşturulan maskeleme verisi ile gizlenen ölçütler, tahmin üretme stratejilerinden bağımsız olarak her bir değişken σ_{max} katsayında $\mathcal{U}$ dağılımına göre daha doğru öneriler üretilmesini sağlamaktadır. Elde edilen öneri doğrulukları tahmin üretme stratejilerine göre incelendiğinde; $\lfloor S_\sigma \rfloor$ kullanılarak oluşturulan komşuluklar $\overline{S_\sigma}$ yöntemine göre daha doğru öneriler üretilmesine olanak sağlamaktadır.

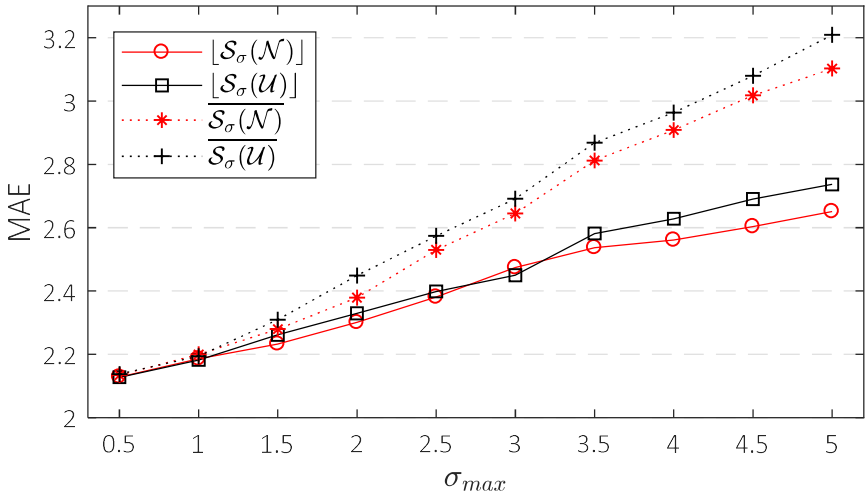


Şekil 5.5. *S_σ doğruluk seviyeleri YM5*

Veri maskeleme işleminde kabul edilen temel prensip; veri gizliliği ve öneri üretme doğruluğu arasında mutlak bir denge kurmaktır. Şekil 5.1(a)'da gösterilen gizlilik seviyeleri hatırlandığında; $\mathcal{N}$ ve $\mathcal{U}$ dağılım için σ katsayısı için en uygun değer 3 olarak kabul edilmektedir. Bunun nedeni, σ katsayısının sağlamış olduğu gizlilik miktarının 3 ve 3'ten büyük değerler için artış miktarının azalmasıdır. Dolayısıyla, gizlilik ve öneri doğruluğunu dengelemek amacıyla *YM5* veri seti için ideal σ_{max} katsayısı [2,5;3] değer aralığında belirlenmektedir. Böylece mahremiyet seviyesi mümkün olan en yüksek seviyede tutulurken öneri doğruluğundan ideal düzeyde kayıp yaşanmaktadır. Dağılımların üretebildiği gizlilik seviyesi, elde edilen öneri doğruluğunda da değişkenlik

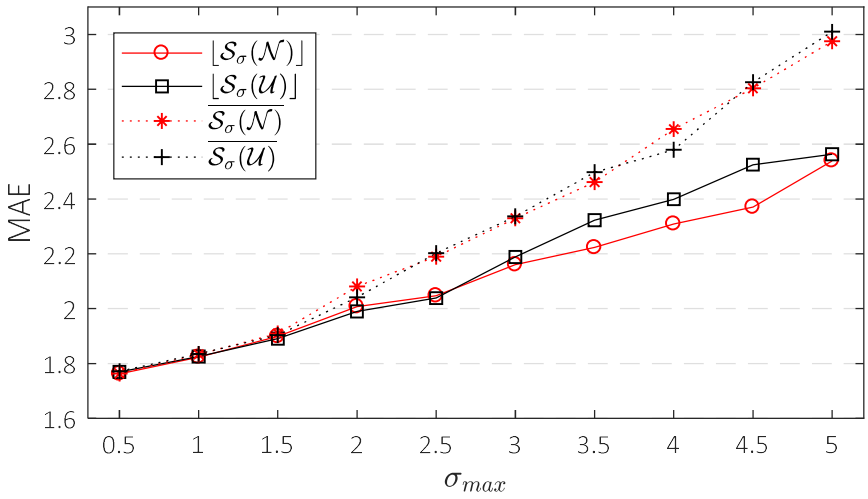
göstererek en iyi mutlak hataların elde edildiği $\lfloor S_\sigma \rfloor$ yöntemi referans alınarak $\lfloor S_\sigma\ (\mathcal{N}) \rfloor$ yöntemi ile $\lfloor S_\sigma\ (\mathcal{U}) \rfloor$ yöntemine göre ortalama olarak %2,08 oranında daha doğru öneriler üretilmesini sağlamaktadır. Bu nedenle *YM5* veri seti için ideal dağılım olarak $\mathcal{N}$ dağılımı seçilmiştir.

YM10 veri seti ile yapılan deneylerde, $\mathcal{N}$ dağılıma göre maskelenen veriler her bir değişken σ_{max} katsayında $\mathcal{U}$ dağılımına göre daha doğru öneriler üretilmesini sağlamaktadır. *YM10* için σ_{max} katsayısının veri gizliliğine etkisini göz önünde bulundurulduğunda, Şekil 5.6'da gösterilen mahremiyet seviyeleri ve elde edilen öneri üretme doğruluğu göz önünde bulundurulduğunda ideal σ_{max} katsayısı 3 olarak belirlenmiştir. Buna ek olarak en iyi mutlak hataların elde edildiği $\lfloor S_\sigma \rfloor$ yöntemi referans alınarak $\lfloor S_\sigma\ (\mathcal{N}) \rfloor$ yaklaşımı ile üretilen öneriler doğruluk bakımından $\lfloor S_\sigma\ (\mathcal{U}) \rfloor$ yöntemine göre %1,36 oranında daha doğru öneriler üretilmesini sağlamaktadır.



Şekil 5.6. *S_σ doğruluk seviyeleri YM10*

Son olarak, seyreklik oranı diğer veri setlerine oranla oldukça düşük düzeylerde olan *YM20* veri seti kullanılarak değişken σ_{max} katsayılarının öneri üretme doğruluğu açısından etkileri Şekil 5.7'de gösterilmektedir. Dağılım ve benzerlik hesaplama stratejileri göz önünde bulundurulduğunda, *YM20* veri seti *YM10* ve *YM5* ile aynı eğilime sahiptir. Elde edilen sonuçlara göre en doğru öneriler $\lfloor S_\sigma(\mathcal{N}) \rfloor$ yöntemi ile elde edilmiştir.



Şekil 5.7. *S_σ doğruluk seviyeleri YM20*

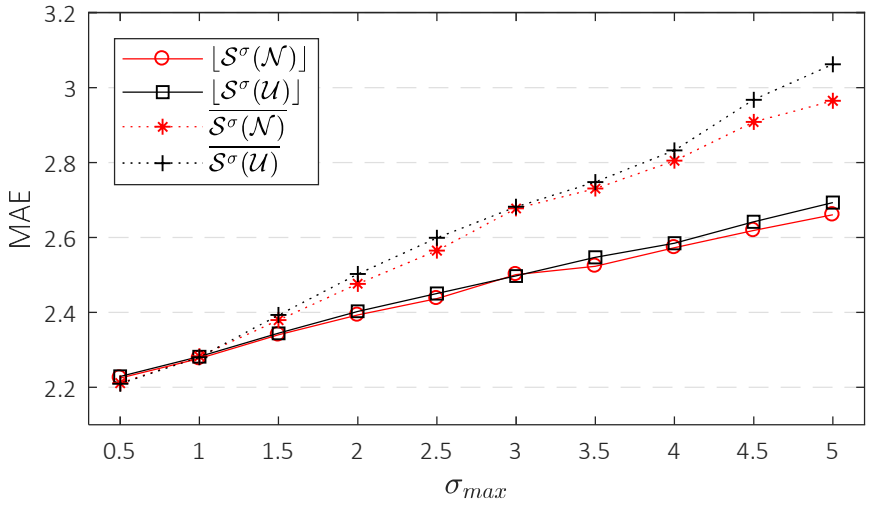
5.3.2.2. S^σ öneri doğruluğu

Prosedür 5.1'de açıklanan maksimum entropi değerine sahip ölçütü en büyük σ değeri ile elde edilen maskeleme vektörü ile gizlendiği S^σ prosedüründe her kullanıcının gizlilik düzeyi, *OF* sistemi tarafından belirlenen $(0,\sigma_{\max}]$ değer aralığındaki bir σ değerinin seçilerek söz konusu değer ile derecelendirme maskeleme vektörlerini oluşturulması ve üretilen sayıların gerçek oy değerlerini maskelemesi temeline dayanmaktadır. Bu yöntemde değişken $\sigma_{\max}$ parametresinin öneri üretme doğruluğuna etkisini değerlendirmek için bu parametre $[0{,}5; 5]$ değer aralığı içerisinde artan 0,5'lik değerler ile test edilmiştir. Elde edilen maskelenmiş kullanıcı oyları ile öneri üretme işlemi $\lfloor S^\sigma \rfloor$ ve $\overline{S^\sigma}$ benzerlik yaklaşımlarının her ikisine göre ayrı ayrı test edilmiştir.

YM5 veri seti için S^σ protokolünde kullanıcı tarafından belirlenen σ katsayısına göre gerçek kullanıcı oy vektörüne eklenmesi ile oluşan yeni derecelendirme vektörünün öneri doğruluğu açısından karşılaştırılması Şekil 5.8'de gösterilmektedir. Öneri üretme doğrulukları maskeleme verisini oluşturmak için kullanılan $\mathcal{N}$ ve $\mathcal{U}$ dağılımlara göre incelendiğinde; $\mathcal{N}$ dağılıma göre maskelenen veriler, her bir değişken σ_{max} katsayında $\mathcal{U}$ dağılımına göre daha doğru öneriler üretilmesini sağlamaktadır.

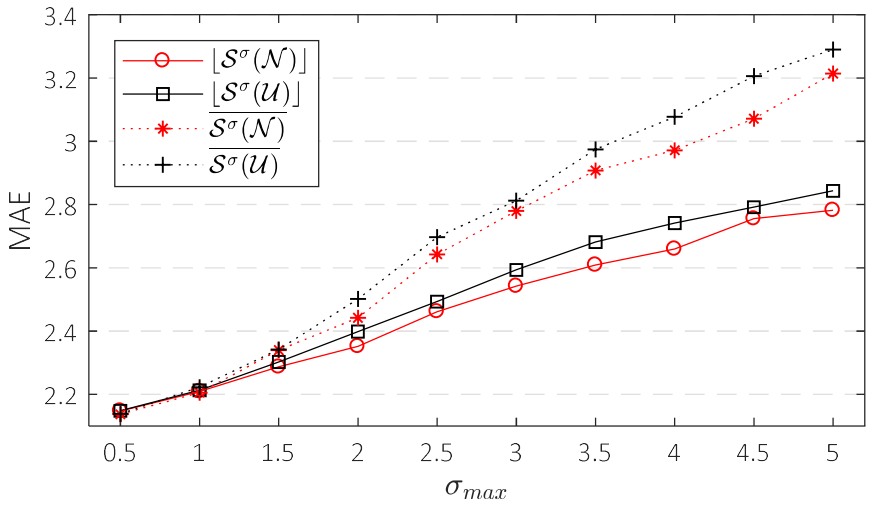
Şekil 5.1(b)'de gösterilen gizlilik seviyeleri hatırlandığında; $\mathcal{N}$ dağılım için σ değerinin $[0{,}5; 2{,}5]$ aralığında, $\mathcal{U}$ dağılımda da benzer şekilde $[0{,}5; 2{,}5]$ aralığında elde ettiği gizlilik seviyesi artış hızının 3'ten büyük değerler için yavaşladığını söylemek mümkündür. Bu nedenle gizlilik ve öneri doğruluğunu dengelemek amacıyla *YM5* veri seti için ideal σ_{max} katsayısını $\mathcal{N}$ ve $\mathcal{U}$ dağılımda 3 olarak belirlenmektedir. Böylece

mahremiyet seviyesi mümkün olan en yüksek seviyede tutulurken öneri doğruluğundan ideal düzeyde kayıp yaşanmaktadır.



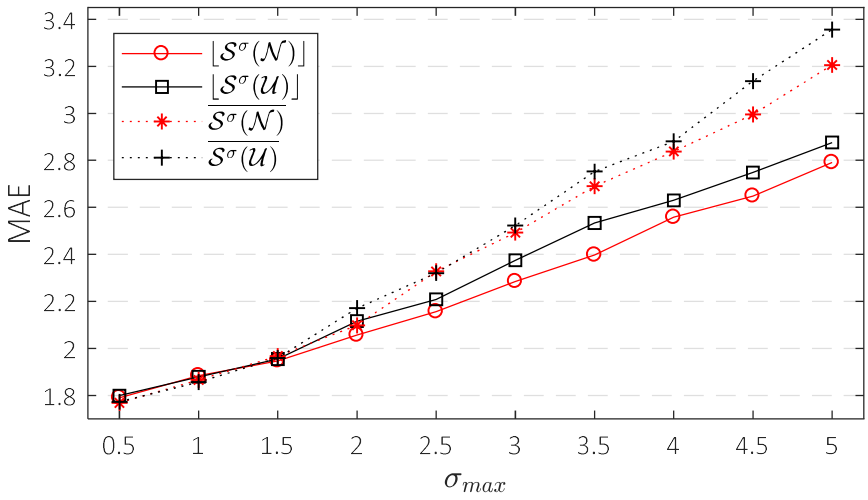
Şekil 5.8. *S^{σ} doğruluk seviyeleri YM5*

YM10 veri seti için elde edilen sonuçlar *YM5* ile elde edilen sonuçlara benzerlik göstermektedir. Özetle yapılan deneylerde; $\mathcal{N}$ dağılıma göre maskelenen veriler her bir değişken σ_{max} katsayısında $\mathcal{U}$ dağılımına göre daha doğru öneriler üretilmesini sağlamaktadır. *YM10* için σ_{max} katsayısının veri gizliliğine etkisi hatırlanıp Şekil 5.9'da gösterilen öneri üretme doğruluğu göz önünde bulundurularak ideal σ_{max} katsayısı 3 olarak belirlenmiştir.



Şekil 5.9. *S^{σ} doğruluk seviyeleri YM10*

Son olarak *YM20* veri seti kullanılarak değişken σ_{max} katsayılarının öneri üretme doğruluğu açısından etkileri Şekil 5.10'da gösterilmektedir. Dağılım ve benzerlik hesaplama stratejileri göz önünde bulundurulduğunda *YM20* veri seti *YM10* ve *YM5* ile aynı eğilime sahiptir. Elde edilen sonuçlara göre en doğru öneriler $\lfloor S^{\sigma}(\mathcal{N}) \rfloor$ yöntemi ile elde edilip gizlilik ve doğruluk arasında denge kurabilmek için σ_{max} katsayısı 3 olarak belirlenmiştir.



Şekil 5.10. *S^{σ} doğruluk seviyeleri YM20*

Gizlilik ve σ_{max} katsayısı arasında doğrusal bir ilişki bulunmakla birlikte σ_{max} değerindeki artış gizlilik seviyesini olumlu yönde arttırmaktadır. Artan σ_{max} değeri ile orijinal veri seti üzerine eklenen sayı vektörünün rassallığı artacağı için gizlilik seviyesi de artacaktır. Ancak, kullanıcı ya da servis sağlayıcı için maskeleme işleminde kullanılacak ideal σ değeri sadece elde edilen gizlilik seviyesi üzerinden karar verilebilecek bir parametre değildir. Bu nedenle, σ değerini belirlemek için gizlilik ve öneri doğruluğu arasında bir denge kurulmalıdır. Buna ek olarak S_{σ} ve S^{σ} yaklaşımları ile üretilen önerilerde $\mathcal{N}$ dağılım ile maskelenen kullanıcı derecelendirme vektörü $\mathcal{U}$ dağılım ile maskelenene göre *MAE* değerleri kıyaslandığında her iki yaklaşımda da daha üstün sonuçlar elde etmektedir. Bunun nedeni, $\mathcal{N}$ dağılımın kullanıcı oy verme eğilimini daha iyi taklit etmesidir.

5.3.3. İstatistiksel anlamlılık

Elde edilen sonuçların istatistiksel olarak anlamlı olup-olmadığını test edebilmek için *t*-test yöntemi kullanılmıştır. $GKOF^{k}$ ile elde edilen hata değerlerini entropi tabanlı

S_σ ve S^σ şemaları ile karşılaştırmak için istatistiksel anlamlılık *t*-testleri gerçekleştirilmiştir. Ortalama mutlak hata sonuçları, 10 kez tekrarlanan deneysel doğruluk değerlerinin ortalaması alınarak ele alınmıştır. $\sigma_{max} = 3$ ve $\beta_{max} = 5$ değerlerinde bir dizi *t*-testi yapılmış ve elde edilen sonuçlar Tablo 5.2'de gösterilmiştir. S_σ için tek kuyruklu hipotezlerin sonuçlarında bütün deneylerde ve tüm $\overline{S^\sigma}$ ve $\lfloor S_\sigma \rfloor$ senaryoları için %99 güven düzeyinde istatistiksel olarak anlamlı olduğu görünmektedir. S^σ yöntemi için elde edilen sonuçlar %95 güven düzeyinde istatistiksel olarak anlamlıdır.

Tablo 5.2. *Önerilen yaklaşımlar ile elde edilen öneri doğruluğu artışlarının istatistiksel anlamlılıkları*

Veri Seti	*Dağılım*	$\overline{S^\sigma}$	$\lfloor S^\sigma \rfloor$	$\overline{S_\sigma}$	$\lfloor S_\sigma \rfloor$
YM20	$\mathcal{U}$	**0,043****	0,003*	**0,000***	0,000*
	$\mathcal{N}$	0,002*	**0,000***	0,000*	0,000*
YM10	$\mathcal{U}$	**0,012****	0,000*	**0,000***	0,000*
	$\mathcal{N}$	0,029**	0,000*	0,000*	**0,000***
YM5	$\mathcal{U}$	**0,000***	0,000*	**0,000***	0,000*
	$\mathcal{N}$	0,003*	0,000*	0,000*	**0,000***

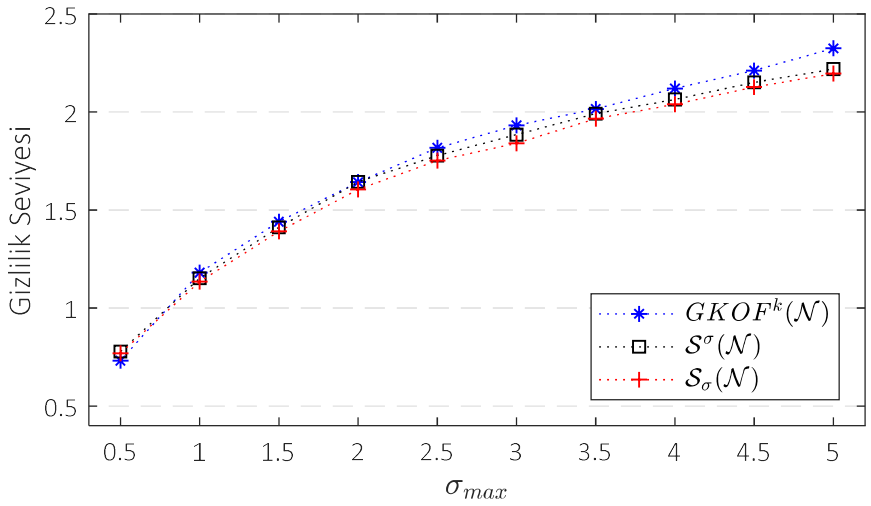
**%95 seviyesinde güvenilir, *%99 seviyesinde güvenilir

5.4. Sonuçlar

Bu bölümde geleneksel *GKOF* yöntemlerinin *OF*^{k} sistemlerine adapte edildiği *GKOF*^{k} yöntemi ile kullanıcının oy verme eğilimi ve alt-ölçütlere verdiği öneme göre dinamik bir şekilde maskeleme işlemi gerçekleştirebilen S_σ ve S^σ yöntemleri kullanılarak elde edilen gizlilik seviyeleri kıyaslanacaktır. Buna ek olarak ortaya konan yeni protokollerin öneri üretme doğruluğu üzerine ektisi maskelenmemiş ham kullanıcı verileri ile elde edilen öneri doğrulukları ile kıyaslanarak, üretilen önerilerin istatistiksel olarak anlamlılıkları tartışılacaktır.

5.4.1. Gizlilik seviyeleri

Ortaya konan *GKOF*^{k}, S_σ ve S^σ yöntemleri genel olarak incelendiğinde her üç protokolde de $\mathcal{N}$ dağılım elde edilen gizlilik seviyeleri açısından $\mathcal{U}$ dağılıma göre daha yüksek seviyelerde gizlilik elde etmiştir. Bunun nedeni, $\mathcal{N}$ dağılımın kullanıcı oy verme eğilimini daha uygun bir şekilde taklit etmesinin yanı sıra $\mathcal{N}$ dağılım ile üretilen *R* vektörünün daha çeşitli rassal sayılardan oluşmasıdır. Geliştirilen her üç yöntem ile elde edilen gizlilik seviyeleri bütün veri setlerinde benzer eğilimleri gösterdiği için kıyaslanan sonuçlar YM10 veri setinde incelenerek Şekil 5.11'de gösterilmektedir.



Şekil 5.11. *YM10 veri seti için $\mathcal{N}$dağılım ile elde edilen gizlilik seviyeleri*

Elde edilen gizlilik seviyeleri kıyaslandığında, önerilen bütün yöntemlerde gizlilik seviyeleri paralellik gösterse de *GKOF*^{k} protokolü S_σ protokolüne göre %3,01, S^σ protokolüne göre ise %1,47 seviyesinde mahremiyeti daha iyi korumaktadır. Bu farklılığın nedeni, *R* vektörünü oluşturan rassal sayılar her üç protokolde de aynı şekilde oluşturuluyor olsa da S_σ ve S^σ protokollerinde ölçütlerin sahip olduğu entropi katsayısına göre farklı σ değerleri kullanılmaktadır. Bu nedenle, elde edilen gizlilik seviyelerinde maskeleme vektörünü oluşan rassal sayıların dağılımından dolayı göz ardı edilebilecek gizlilik seviyesi farklılıkları ortaya çıkabilmektedir. *GKOF*^{k} protokolünde her bir alt-ölçüt aynı σ_{max} değeri ile elde edilen farklı *R* vektörleri ile maskelendiği için mahremiyet seviyesinin diğer protokollerden yüksek çıkması beklenen bir sonuçtur. Bunun dışında gözlemlenen bir diğer sonuç, Yahoo!Movies veri setinin farklı seyreklik seviyelerine sahip alt kümelerinde gözlemlenen mahremiyet seviyeleri farklılıklarıdır.

YM20, *YM10* ve *YM5* veri setleri için en yüksek gizlilik seviyesini sağlayan $\mathcal{N}$ dağılım ile elde edilen ideal gizlilik seviyeleri Tablo 5.3'te verilmektedir. Elde edilen değerler önerilen bütün protokoller göz önünde bulundurularak incelendiğinde; en yüksek gizlilik seviyesi *YM20* veri setinde elde edilmiş, *YM10* ve *YM5* veri setlerinde azalarak devam etmiştir. Bunun temel nedeni veri setlerinin farklı düzeylerdeki seyreklik oranlarıdır. *YM20* veri setinde, bir kullanıcı yaklaşık 40 ürüne oy vermektedir ve bu veri setinde 247 adet ürün bulunmaktadır. Yani, oluşturulan maskeleme vektörünün eleman sayısı, toplam ürün sayısının yaklaşık %16'sını kapsamaktadır. *YM10* veri setinde bu oran yaklaşık %2,3; *YM5* veri setinde ise %0,7'dir. Bunun anlamı, daha az sayıda ürün

derecelendirmesine sahip seyrek veri setlerinde gerçek derecelendirme vektörüne eklenen rastgele sayı miktarı da daha az olacağı için kullanıcı derecelendirmeleri daha düşük seviyede karıştırılmaktadır. Bir başka ifade ile oluşturulan maskeleme vektörünün eleman sayısı toplam ürün sayısına orantılandığında ortaya çıkan oran düşük seviyelerdeyse genel gizlilik seviyesi de düşük seviyelerde seyretmektedir. Bu nedenle artan seyreklik ve azalan oy-ürün oranı ile ilişkili olarak elde edilen gizlilik seviyeleri de değişkenlik göstermektedir.

Tablo 5.3. *YM veri setlerinde elde edilen ideal gizlilik seviyeleri*

$\sigma_{max} = 3$ $\beta_{max} = 0$	*YM20*	*YM10*	*YM5*
$GKOF^k(\mathcal{N})$	2,82	1,93	1,24
$S^{\sigma}(\mathcal{N})$	2,79	1,89	1,20
$S_{\sigma}(\mathcal{N})$	2,76	1,84	1,22

5.4.2. Öneri doğruluğu

Veri maskeleme prosedüründe $\mathcal{N}$ dağılım ile üretilen maskeleme verisinin kullanıcı gizliliğine katkısı $\mathcal{U}$ dağılıma göre bütün veri setleri üzerinde daha iyi sonuçlar vermektedir. Aynı zamanda, Şekil 5.5. S_{σ} doğruluk seviyeleri YM5 ve Şekil Şekil ***5.10.*** S^{σ} doğruluk seviyeleri YM20 arasında sunulan doğruluk seviyeleri göz önünde bulundurulduğunda öneri üretme doğruluğu bakımından da $\mathcal{N}$ dağılım ile elde edilen maskeleme verisi ile gizlenen kullanıcı verisi ile üretilen öneriler $\mathcal{U}$ dağılıma göre daha doğru sonuçlar vermektedir.

Önerilen $GKOF^k$ yaklaşımları, çoklu-ölçütlü kullanıcı derecelendirmeleri için yüksek seviyede gizlilik düzeyi sağlamada yararlı olsa da, maskelenmemiş ham kullanıcı verileriyle OF^k yaklaşımına göre üretilen öneri doğrulukları gizlilik koruma ortamındaki tahmini doğruluk kayıplarını gözlemlemek için kıyaslanmalıdır. Bu amaçla, maskelenmemiş gerçek kullanıcı derecelendirmeleri ile $\bar{\cdot}$ ve $\lfloor\cdot\rfloor$ benzerlik yaklaşımları kullanılarak öneri üretme işlemi gerçekleştirilerek gizlilik ve doğruluk ile ilgili çakışan hedeflere bağlı kayıplar gözlemlenmiştir. Bu karşılaştırma yapılırken *YM20*, *YM10* ve *YM5* veri setleri için ideal gizlilik parametreleri $\sigma_{max} = 3$ olarak belirlenmiştir.

Maskelenmemiş ham veri seti, $GKOF^k$, S_{σ} ve S^{σ} ile elde edilen öneri doğrulukları ve ideal maskeleme parametreleri ile maskelenen kullanıcı derecelendirmelerinden elde edilen öneri doğrulukları Tablo 5.4'te gösterilmektedir. Elde edilen sonuçlar $GKOF^k$

yaklaşıma göre incelendiğinde; *YM5*, *YM10* ve *YM20* veri kümeleri için doğruluk kayıpları % 23,24; % 27,75 ve % 33,19'dur. Gözlemlenen öneri hatalarını azaltmak için geliştirilen entropi tabanlı yaklaşımlardan en yüksek doğruluk seviyesinin elde edilmesini sağlayan $\lfloor S_\sigma \rfloor$ yaklaşımının $\lfloor OF^k \rfloor$ yaklaşımına göre elde edilen öneri doğrulukları ile karşılaştırıldığında *YM5*, *YM10* ve *YM20* veri kümeleri için doğruluk kayıpları sırasıyla %11,4; %15,1 ve %23,2'dir. Elde edilen kullanıcı öneri kayıpları incelendiğinde, entropi tabanlı $\lfloor S_\sigma \rfloor$ yaklaşımının geleneksel gizlilik koruma yöntemlerine göre oldukça üstün sonuçlar elde ettiği gözlemlenmektedir.

Tablo 5.4. *Maskelenmemiş veri ve ideal gizlilik parametreleriyle maskelenmiş veriden elde edilen doğruluk seviyeleri*

Veri Seti	*Dağılım*	$\overline{OF^k}$	$\overline{GKOF^k}$	$\overline{S^\sigma}$	$\overline{S_\sigma}$	$\lfloor OF^k \rfloor$	$\lfloor GKOF^k \rfloor$	$\lfloor S^\sigma \rfloor$	$\lfloor S_\sigma \rfloor$
YM20	$\mathcal{U}$	**1,745**	2,641	2,536	2,286	**1,739**	2,652	2,382	2,101
	$\mathcal{N}$		2,612	2,420	**2,235**		2,631	2,279	**2,067**
YM10	$\mathcal{U}$	**2,104**	2,987	2,974	2,816	**2,099**	2,995	2,663	2,528
	$\mathcal{N}$		2,905	2,813	**2,729**		2,905	2,539	**2,471**
YM5	$\mathcal{U}$	**2,184**	2,903	2,909	2,730	**2,180**	2,901	2,561	2,481
	$\mathcal{N}$		2,846	2,838	**2,664**		2,840	2,565	**2,459**

6. VERİ MASKELEME İŞLEMİNDE OLAĞAN DIŞI OYLARIN ETKİSİ ve ÖNERİ DOĞRULUĞUNUN İYİLEŞTİRİLMESİ

OF sistemleri, öneri üretme kalitesini olumsuz olarak etkileyen oldukça seyrek kullanıcı verileri ile baş etmek zorundadır (Bilge ve Yargıç, 2017). *OF* sistemlerinde öneri üretme sürecinin başarısı kullanılan yöntemler kadar, *OF* sisteminin üzerinde çalıştığı veri seti ile de ilişkilidir (Su, Khoshgoftaar, 2009). *OF* sistemleri kullanıcı açısından tatmin edici doğruluk seviyelerinde öneriler üretebilmek için, gerçek kullanıcı tercihlerine dayalı, gürültüden arındırılmış ve yeterli miktarda *kullanıcı* × *ürün* derecelendirmesine sahip tercih verilerine ihtiyaç duymaktadır. Aksi durumlarda, niteliksiz koleksiyonlara dayalı, doğru ve güvenilir olmayan tahminlerin üretilmesi, özellikle ticari öneri sistemleri için önemli bir sorun haline gelmektedir (Su, Khoshgoftaar, 2009). Pratikte, OF^k sistemleri, kullanıcıların farklı ürün ve hizmetlere ait tercih verilerinden oluşan büyük bir *kullanıcı* × *ürün* matrisi üzerinde öneri üretme işlemi gerçekleştirmektedir (Adomavicius ve Kwon, 2007). Bununla birlikte, ortalama bir sistem kullanıcısı, tüm ürün listesi göz önünde bulundurulduğunda son derece seyrek bir *kullanıcı* × *ürün* matrisine sahiptir ve sistemin sahip olduğu ürün miktarının çok küçük bir oranını derecelendirmektedir. İyi bilinen OF^k veri tabanları için ortalama seyreklik seviyesi %98'den yüksektir (Jannach, Karakaya ve Gedikli, 2012; Nilashi vd., 2015). Kullanıcıların tercihlerinden oluşan bu verinin kısıtlılığı, kullanıcılar arasındaki korelasyonları keşfetmeyi olumsuz olarak etkiler, hatta bazı durumlarda tamamen engeller. Çünkü geleneksel komşuluk tabanlı benzerlikler yalnızca kullanıcılar arasındaki ortak derecelendirilen ürün listesi üzerinde gerçekleştirilir (Zheng, Agnani ve Singh, 2017).

Kullanıcı tabanlı benzerlik yaklaşımı kullanan *OF* sistemleri öneri üretmek için aktif kullanıcı ile diğer kullanıcılar arasındaki benzerliklerden faydalanmaktadır. Derecelendirilen ürün listesi ve derecelendirme profillerinin yanı sıra aktif kullanıcı ve diğer kullanıcıların kendilerine özgü oy verme eğilimleri, üretilen önerilerin kalitesini etkileyen temel faktörlerdendir (Zheng, Agnani ve Singh, 2017). *OF* sistemlerinde kullanıcı profilleri genel olarak üç ana başlık altında sınıflandırılmaktadır (McCrae, Piatek ve Langley, 2004). Bunlar;

- Veri setinin genelini oluşturan ve nicelik bakımından en geniş kullanıcı sayısını oluşturan ilk sınıf altında toplanan kullanıcılar, oy verme profilleri

bakımından diğer kullanıcılar ile kolaylıkla komşuluk kuran ve yüksek seviyede derecelendirme korelasyonuna sahip olan kullanıcılardır.

- İkinci sınıf altında gruplanan kullanıcılar, ilk kullanıcı grubunun aksine genellikle çok az sayıda kullanıcı ile komşuluk kurabilen hatta bazı durumlarda hiçbir kullanıcı ile benzer profile sahip olmayan kullanıcılardır.
- Üçüncü sınıfta gruplanan Sıra Dışı Kullanıcılar (*SK*), diğer kullanıcılarla düşük korelasyonlara neden olan farklı ve alışılmamış beğeni profillerine sahip kullanıcı grubudur (McCrae, Piatek ve Langley, 2004). Bir başka deyişle, olağan dışı kullanıcı profilleri diğer kullanıcı profilleri ile ne tam anlamıyla uyumludur ne de uyumsuzdur (Ghazanfar, Prugel-Bennett, 2011; McCrae, Piatek ve Langley, 2004; Zheng, Agnani ve Singh, 2017). Buna ek olarak ikinci sınıf altında toplanan kullanıcı grubundan farklı olarak, diğer kullanıcıların komşuluklarına dâhil olup öneri kalitesini düşürmeye neden olan oy verme eğilimleri bulunmaktadır.

İlk iki grubu oluşturan kullanıcılar genel olarak *OF* sistemleri açısından problem teşkil etmeyen kullanıcı profilleridir. Hiçbir komşuluğu olmayan bir kullanıcı için kullanıcı tabanlı bir *OF* sisteminin öneri üretememesi kabul edilebilir bir senaryodur. Ancak, kullanıcılarla düşük korelasyonlara neden olan farklı ve alışılmamış beğeni profillerine sahip üçüncü tip kullanıcılar diğer kullanıcıların komşuluk kümelerine dâhil olarak hem kendisi hem de diğer kullanıcılar için alışılmamış ve öngörülemeyen hatalı öneriler üretilmesine neden olmaktadır.

6.1. *OF* Sistemlerinde Sıra Dışı Kullanıcı Problemi

OF sistemlerinin geliştirilmesine yönelik yapılan çalışmalarda olağan dışı oy verme eğiliminde olan kullanıcıların belirlenmesi ve bu kullanıcılar için üretilen önerilerin doğrulukları konusunda yapılan çalışmalar özellikle gizliliği koruyan çoklu-ölçütlü ortak filtreleme alanında bulunmamaktadır. Olağan dışı oy verme eğilimindeki kullanıcı profilleri genellikle diğer kullanıcılar ile pozitif korelasyona sahip olmayan, diğer kullanıcıların komşuluklarına dâhil olamayan ya da diğer kullanıcılar ile pozitif korelasyonu olup sıra dışı oy verme profilleri nedeniyle hem kendisi hem de diğer kullanıcılar için hatalı öneriler üretilmesine neden olan kullanıcı profillerinden oluşmaktadır (McCrae, Piatek ve Langley, 2004; Claypool vd., 1999). Sıra dışı olarak

nitelendirilen bu kullanıcılar, *OF* sisteminin ürettiği öneri kalitesini olumsuz etkilemektedir (Ghazanfar, Prugel-Bennett, 2014; Ghazanfar, Prugel-Bennett, 2011; Zheng, Agnani ve Singh, 2017; McCrae, Piatek ve Langley, 2004; Ruiz-Montiel ve Aldana-Montes 2009; Su ve Khoshgoftaar, 2009). Bu nedenle sıra dışı kullanıcı profillerinin *OF* sisteminde tanımlanması ve bireysel olarak ele alınması gerekmektedir (Zheng, Agnani ve Singh, 2017).

Geleneksel *OF* sistemlerinde sıra dışı kullanıcı sorununa işaret eden ve bu kullanıcılarının özelliklerini tanımlayan birçok araştırma bulunmakla birlikte, sıra dışı kullanıcı profillerini saptamaya yönelik çözümleri bulmak için yapılan çalışmalar nicelik olarak oldukça sınırlıdır (Claypool vd. 1999; McCrae, Piatek, ve Langley, 2004; Ruiz-Montiel ve Aldana-Montes 2009; Su ve Khoshgoftaar , 2009). Yapılan çalışmalara örnek olarak; Ghazanfar ve Prugel-Bennett (2011,2014) sıra dışı kullanıcıları saptamaya yönelik yaptığı çalışmada; değişken kullanıcı benzerlikleri için eşik değerleri belirleyerek bu değerler doğrultusunda sıra dışı kullanıcı kümeleme tekniği kullanmıştır. Böylece farklı kullanıcı benzerlik eşik değerleri belirleyerek sıra dışı kullanıcı profillerini diğer kullanıcı profillerinden ayırmaktadır. Ancak kümeleme tabanlı yaklaşımların en büyük dezavantajı, kümelenme sürecinde uygun değer küme sayısının bulunma zorluğu ve kümelenme sürecindeki yakınsamaların giderilmesi için yüksek hesaplama maliyetinin bulunmasıdır. Kümeleme tabanlı yöntemlerin dışında Gras, Brun ve Boyer (2016), kullanıcı derecelendirmelerinin dağılımına dayalı olarak oy dağılımlarındaki sapmaları kullanmaktadır. Ayrıca, öneri tahminlerinde ortaya çıkan sıra dışı hata katsayıları da *SK* belirlemek için kullanmaktadır. Ancak, öneri doğruluğunda ortaya çıkan hata değerleri yalnızca bir kullanıcının sıra dışı olup olmadığını değerlendirmek için yeterli değildir. Bunun nedeni, sıra dışı oy verme eğiliminin tek başına büyük öneri hatalarına yol açan tek neden olmamasıdır. Diğer bir deyişle, büyük tahmin hatalarıyla ilişkili bir kullanıcının sadece hata oranları referans alınarak sıra dışı olarak nitelendirilmesi her zaman için doğru değildir (Zheng, Agnani ve Singh, 2017). Bir diğer yaklaşımda, sıra dışı olarak tanımlanan kullanıcıların diğer birçok kullanıcıyla düşük korelasyona sahip olmaları ve çok az sayıda kullanıcının komşuluğuna girebilmelerinden yararlanılarak bu kullanıcıları belirleme işlemi gerçekleştirilmektedir (Claypool vd., 1999). Kullanıcı benzerliklerinden yararlanılarak sıra dışı kullanıcı profillerinin belirlendiği diğer bir yaklaşımda; kullanıcıların birbirleri ile olan benzerlik ilişkileri istatistiksel olarak analiz edilip, aykırı saplamalara neden olan kullanıcı profilleri sıra dışı olarak sınıflandırılmaktadır (Zheng,

Agnani ve Singh, 2017). Ancak, kullanıcı benzerliğine dayalı ikili korelasyonlar kullanılarak yapılan sınıflandırmalar bazı dezavantajlara sahiptir. Bunlardan ilki; eğer kullanıcılar arasında ortak derecelendirilmiş ürünler yoksa kullanıcı tabanlı benzerlik değerinin hesaplanması mümkün değildir. Diğer problem ise iki kullanıcı tarafından ortak olarak derecelendirilen sınırla sayıda öğe mevcut ise bu kullanıcıların öneri üretme işleminde referans alınması öneri doğruluğunda hataya neden olabilir.

6.2. Problem Tanımı ve Amaç

Sıra dışı kullanıcı ya da oy değerlerini bulmak adına yapılan çalışmalar, ham kullanıcı verileri, yani maskelenmemiş orijinal kullanıcı verileri üzerinde çalışmaktadır. Ancak, rastgele karıştırma ya da rastgele doldurma prosedürlerinde de olduğu gibi orijinal kullanıcı verisinin değiştirilerek gizlendiği bir veri setinde kullanıcı korelasyonu, kümeleme ya da kullanıcı oylarına ait dağılımlar kullanılarak sıra dışı kullanıcı ya da oy değerlerini belirlemek ve onlara özgü öneri üretme çözümleri sunmak mümkün değildir. Çünkü maskeleme işlemi sonucunda kullanıcının ürün derecelendirme vektöründe yapılan değişiklik sıra dışılığı belirlemeye engel olmaktadır.

GKOF yöntemlerinde, veri maskeleme işlemi istemci tarafında gerçekleştirilen bir süreçtir ve sunucu sadece maskelenmiş kullanıcı-ürün matrisine sahiptir. Bu esnada, istemci diğer kullanıcı oy vektörlerine erişim sağlayamadığı için maskelenmemiş ham kullanıcı verisi olarak kullanıcı komşuluklarını belirlemek ve buna göre sıra dışı kullanıcıları sınıflandırmak mümkün değildir. Bu nedenle, maskelenmiş veri setleri üzerinde geleneksel *OF* ve OF^k sistemlerinde yapılanın aksine kullanıcı tabanlı sıra dışılık belirleme stratejilerinin yerine aktif kullanıcıya ait oy vektörü üzerinde ürün tabanlı sıra dışılık belirleme yaklaşımları kullanılmalıdır. Ürün tabanlı sıra dışılık belirleme stratejisinde, aktif kullanıcının oy verme eğilimi analiz edilerek oy verme eğilimine göre standart dışı oy değerleri ile derecelendirdiği ürün listesi belirlenmelidir. Bu nedenle yapılan çalışmada, OF^k sistemlerinde aktif kullanıcıya ait sıra dışı derecelendirmelerin belirlenmesi ve bu oy değerlerine özgü bir veri maskeleme prosedürünün uygulanması amacıyla yeni bir strateji sunulmuştur.

6.3. Sıra Dışı Derecelendirmelerin Belirlenmesi

Geleneksel *OF* sistemleri tek bir oy değerine ait *kullanıcı* × *ürün* vektöründen meydana gelmektedir. Ancak OF^k sistemleri birden fazla alt-ölçütten meydana

gelmektedir. Bu nedenle, geleneksel yaklaşımlar çoklu-ölçütlü veri setlerine direkt olarak adapte edilemezler. OF^k sistemlerinde sıra dışı oy değerleri belirleme işleminde göz önünde bulundurulması gereken en temel problemler; hangi ölçüt ya da ölçütlerin kullanıcı için en yüksek bilgiye sahip olduğunu belirlemek ve bu ölçütleri referans ölçüt olarak kabul edip sıra dışı oy değerlerini belirlemektir. Bu nedenle, sıra dışı oy değerlerini belirlemeye yönelik yapılan çalışmaların yanı sıra ölçüt belirleme stratejileri de geliştirilmiştir.

Sıra dışı oy değerleri genel olarak, diğer oy değerlerinden çok fazla büyük ya da küçük oy değerleridir, bir başka deyişle aykırı kullanıcı oylarını ifade etmektedir. Aykırı değerleri belirlemek için literatürde birçok çalışma bulunmaktadır. Sıra dışı oy değerlerinin belirlenmesi için kullanıcı oy değerlerinin dağılımlarına dayalı dört temel strateji kullanılmaktadır. Bunlar; Rosner metodu (*gesd*) (Chelishchev,Popov, ve Sørby, 2018), *median* (Leys vd., 2013), *mean* (Leys vd., 2013) ve Tukey testi (*quartiles*) (Wang, Caja, ve Gómez, 2018) stratejileridir.

6.4. Ölçüt Belirleme Stratejileri

OF^k sistemlerinde kullanılan veri setleri yapısı gereği birden çok ölçütten meydana gelmektedir. Bu ölçütler içerisinde kullanıcı profilini en iyi şekilde temsil eden ölçüt ya da ölçütleri belirlemek kullanıcının oy verme eğilimini doğru şekilde analiz etmeyi sağlar. Bu ölçütler kullanılarak uygulanacak sıra dışılık belirleme stratejileri öneri doğruluğu bakımından iyileşmelere neden olmaktadır. Bu kapsamda, $GKOF^k$ sistemlerinde öneri doğruluğunu geliştirmek için dört farklı yaklaşım değerlendirilmiştir.
Bu yaklaşımlarda en önemli ölçüt ya da ölçütler olarak;

- genel beğeni ölçütü,
- en yüksek entropi değerine sahip ölçüt,
- bütün alt-ölçütler ve
- en yüksek entropi ile genel beğeni ölçütünün birleşim kümesi kullanılmaktadır.

Kullanılan sıra dışı oy değerleri ve ölçüt belirleme stratejileri tanımlanırken $GKOF^k$ sistemlerinde veri maskeleme işlemi referans alınarak veri maskeleme işlemi gerçekleştirilecektir. Ancak bu prosedürden farklı olarak; R oluşturulduktan sonra sıra dışı oy değerlerini ve referans ölçütü belirleyip R vektörünü yeniden sıralamaya yarayan

sıra dışılık belirleme fonksiyonu eklenmiştir. Sıra dışılık belirleme fonksiyonunu oluşturan temel basamaklar;

- sıra dışı oy değerleri için referans ölçütün belirlenmesi,
- sıra dışı oy değerlerini belirleme fonksiyonun seçilmesi,
- sıra dışı oy değerlerinin (*SO*) ve toplam sıra dışı oy değeri sayısının (*#SO*) belirlenmesi,
- maskeleme verisinin artan düzene göre sıralanması (*Sr*),
- sıralanmış maskeleme vektörünün ilk *#SO* elemanının sıra dışı oy değerlerini maskelemesi ve,
- Kalan maskeleme verisinin rastgele olarak yeniden sıralanıp sıra dışı olarak işaretlenmeyen oy değerlerini maskelemesidir.

Sıra dışı oy değerlerini belirlemek için kullanılan bütün stratejiler Prosedür 6.1 üzerinde, önerilen yönteme göre oluşturulmuş değişken sıra dışılık belirleme fonksiyonları (*SBF*) ile test edilmiştir.

Prosedür 6.1. *$GKOF^k$ sistemlerinde sıra dışı oy değerlerinin belirlenmesi ve R vektörünün yeniden oluşturulması*

Require: Kullanıcı × ölçüt × ürün vektörü (G_k), $\sigma_{max}, \beta_{max}$

Bütün ölçütler için z-skoru değerlerinin hesaplanması ($\rightarrow \boldsymbol{Z_k}$)

1 : **for** all criteria in G $(i \leftarrow 1\ to\ k)$ **do**

2 : $\bar{G}_i \leftarrow MEAN(G_i)$

3 : $\sigma_{G_i} \leftarrow STD(G_i)$

4 : **end for**

5 : **for** all criteria in G $(i \leftarrow 1\ to\ k)$ **do**

6 : **for** all items in G_i $(i \leftarrow 1\ to\ m)$ **do**

7 : $Z_{ij} = (G_{ij} - \bar{G}_i)/\ \sigma_{G_i}$

8 : **end for**

9 : **end for**

Gizlilik parametrelerinin belirlenmesi

10 : $\beta \leftarrow RND\ (0, \beta_{max})$

11 : $\sigma \leftarrow RND\ (0, \sigma_{max})$

12 : $\alpha \leftarrow \sqrt{3}\sigma$

13: $e \leftarrow |E|$ ► # oy verilmemiş ürün

14 : $g \leftarrow |G|$ ► # gerçek kullanıcı oyları

15 : $F \leftarrow e \times \beta\%$ ► # doldurulacak hücre sayısı

Dağılımın belirlemesi ve rastgele sayı türetilmesi

16 : $dist \leftarrow RANDOM\ (uniform|\ normal)$

17 : $R_{tmp} \leftarrow dist(g + F;\ \mu = 0, \sigma|\ \alpha)$

Sıra dışılık belirleme fonksiyonu (*SBF*)

18 : $R \leftarrow (SBF(TD|S_{max}|GD|GDS_{max}, R_{tmp})$

***z*-skoru değerlerinin maskelenmesi ($\rightarrow Z'$):**

19 : **for** all criteria in G $(i \leftarrow 1\ to\ k)$ **do**

20 : **for** all items in G_i $(i \leftarrow 1\ to\ m)$ **do**

21 : $Z'_{ij} = (Z_{ij} + R_{ij})$

22 : **end for**

23 : **end for**

24 : **return** Z'

6.4.1. Genel beğeni ölçütü (GB)

Genel beğeni değeri olarak isimlendirilen ölçüt (G_0), kullanıcının alt-ölçütlerden bağımsız olarak oy verdiği ürünü bütün özellikleriyle tek bir ölçüt altında değerlendirmesine olanak sağlayan ve çoklu-ölçütlü sistemlerde kullanılan bir derecelendirme ölçütüdür. Bu ölçüt diğer alt-ölçütlerden bağımsız olarak kullanıcının alt-ölçütlerde ifade edemediği özelliklere ait beğeni değerlerini ve ürün hakkındaki son kararını ifade ettiği için bütün ölçütler arasında en yüksek değere sahip ölçüttür. Ayrıca, öneri üretme sürecinde kullanıcıya sunulan nihai tahmin değeri olarak G_0 ölçütü için üretilen tahmin esas alınmaktadır. Bu nedenle; kullanıcın bu ölçütü derecelendirirken kullandığı sıra dışı oy değerlerini belirlemek ve maskeleme verisini bu ölçütten elde edilen sıra dışı oy değerlerine göre yeniden sıralamak öneri doğruluğunu arttırmak adına pozitif katkı sağlamaktadır.

GB stratejisinde sıra dışı oy değerlerini belirlemek ve gerçek kullanıcı verisini maskelemek için yapılan işlemler aşağıda sıralanmaktadır.

i. Kullanıcıya ait ham derecelendirme değerlerine sahip olan G_0 ölçütü içerisinde tercih edilen sıra dışı oy değeri belirleme tekniğine göre

($SO_{gesd}, SO_{mean}, SO_{median}, SO_{quartiles}$) sıra dışı oy vektörü ($SO$) oluşturulur.

ii. R vektörü negatif ve pozitif sayılardan meydana geldiği için $|R|$ küçükten büyüğe doğru sıralanıp indekslenir.

iii. Her bir ölçüt için; sıralanmış R vektörünün (Sr) ilk $\#SO$ elemanı sıra dışı oy değerlerini maskeleme için kullanılır.

iv. Kalan normal oy değerlerini gizlemek için; maskeleme vektöründe sıra dışı oy değerlerini gizlemek için kullanılmayan değerler rastgele bir biçimde yeniden sıralanır ve gerçek oy değerlerine eklenir.

Bu stratejide kullanılan sıra dışılık fonksiyonu ve veri maskeleme işlemi Prosedür 6.2'de gösterilmektedir.

Prosedür 6.2. *GB stratejisinde kullanılan sıra dışılık belirleme ve veri maskeleme prosedürü*

Sıra dışı oy belirleme stratejisinin belirlenmesi

1 : $alg \leftarrow (gesd, mean, median, quartiles)$

2 : $SO \leftarrow isoutlier(G_o, alg)$ ► Genel derecelendirme üzerinde sıra dışı oy değerlerini indeksle

3 : $Sr \leftarrow sort(|R|, descend)$ ► R vektörünü sırala

Sıra dışı oy değerlerine maskeleme veri indekslerinin atanması

4 : **for** all items in SO $(i \leftarrow 1\ to\ k)$ **do**

5 : $R(SO_i) = Sr_i$

6 : **end for**

Nihai maskeleme verisinin oluşturulması

7 : $Rson \leftarrow randomize\ index\ of\ (\forall Sr \in R\ \neg(Sr = R(SO)))$

6.4.2. En yüksek entropi (S_{max})

Bilgi teorisinde entropi (S) kavramı, bir veri kaynağındaki belirsizliğin miktarı olarak tanımlanmaktadır (Shannon, 1948). Bir veri setinin sahip olduğu yüksek entropi değeri, o veri setini oluşturan değerlerinin olasılık dağılımlarının rassallığı ile ilişkilidir. Veri setini oluşturan ölçütler arasında entropi değeri referans alınarak yapılacak bir sıralama, maskeleme işlemindeki önem sırasını da belirlemeye yardımcı olabilecektir. Bu nedenle entropi değeri, ölçütlerin önem sırasını belirlemek için kullanılmıştır.

Entropi değerinin yüksek olması kullanıcının derecelendirme değerlerinin rassal olarak dağılımı ile doğru orantılıdır. Gerçek oy değerlerindeki bu rassallık kullanıcı oy dağılımlarından elde edilebilecek bilgi miktarını belirlemektedir. Bu nedenle, sıra dışı oy değerlerini belirlerken en yüksek entropiye sahip ölçütü referans alma stratejisi kullanılmaktadır. Bu stratejide kullanılan sıra dışılık fonksiyonu ve veri maskeleme işlemi Prosedür 6.3'de gösterilmektedir.

Prosedür 6.3. S_{max} *stratejisinde kullanılan sıra dışılık belirleme ve veri maskeleme prosedürü*

Her ölçüt için *S* değerinin hesaplanması

1 : **for all criteria in** G $(i \leftarrow 1\ to\ k)$ **do**

2 : $Ent_G = \sum_{j=1}^{r} p_j log\ (p_j)$ ► Derecelendirmelerin görülme olasılığı

3 : **end for**

Sıra dışı oy belirleme stratejisinin belirlenmesi

4 : $alg \leftarrow\ (gesd, mean, median, quartiles)$

En yüksek *S* değerine sahip ölçüt (*t*) için sıra dışı oy değerlerinin indekslenmesi

5 : $SO \leftarrow isoutlier(G_t, alg)$

6 : $Sr \leftarrow sort(|R|, descend)$ ► R vektörünü sırala

7 : **for all items in** $\boldsymbol{SO}$ $(i \leftarrow 1\ to\ k)$ **do**

8 : $R(SO_i) = Sr_i$

9 : **end for**

Nihai maskeleme verisinin oluşturulması

10 : $Rson \leftarrow \boldsymbol{randomize\ index\ of}\ (\forall \boldsymbol{Sr}\ \in \boldsymbol{R}\ \neg(\boldsymbol{Sr} = \boldsymbol{R}(\boldsymbol{SO})))$

S_{max} stratejisinde sıra dışı oy değerlerini belirlemek ve gerçek kullanıcı verisini maskelemek için yapılan işlemler aşağıda anlatılmaktadır.

i. Genel derecelendirme ölçütü de dâhil olmak üzere, kullanıcının bütün derecelendirme ölçütleri için entropi değerleri ayrı ayrı hesaplanır.

ii. En yüksek entropi katsayısına sahip ölçütteki derecelendirme değerleri referans alınarak, kullanıcıya ait ham derecelendirme değerleri, sıra dışı oy değeri belirleme tekniğine göre ($SO_{gesd}, SO_{mean}, SO_{median}, SO_{quartiles}$) SO vektörü oluşturulmaktadır.

iii. R vektörü negatif ve pozitif sayılardan meydana geldiği için $|R|$ küçükten büyüğe doğru sıralanıp indekslenir.

iv. Her bir alt-ölçüt için; Sr'ın ilk $\#SO$ elemanı sıra dışı oy değerlerini maskeleme için kullanılır.

v. Kalan normal oy değerlerini gizlemek için; maskeleme vektöründe sıra dışı oy değerlerini gizlemek için kullanılmayan değerler rastgele bir biçimde yeniden sıralanır ve gerçek oy değerlerine eklenir.

6.4.3. Genel beğeni değeri ve en yüksek entropi (GBS_{max})

Bu stratejide en yüksek entropi değerine sahip alt-ölçüt ile birlikte genel derecelendirme ölçütünde bulunan sıra dışı oy değerlerinin birleşim kümesi, sıra dışı oy değeri olarak tanımlanmaktadır. Bu stratejide kullanılan sıra dışılık fonksiyonu ve veri maskeleme işlemi Prosedür 6.4'te gösterilmektedir.

Prosedür 6.4. *GBS_{max} stratejisinde kullanılan sıra dışılık belirleme ve veri maskeleme prosedürü*

Sıra dışı oy belirleme stratejisinin belirlenmesi

1 : $alg \leftarrow (gesd, mean, median, quartiles)$

Genel derecelendirme üzerinde sıra dışı oy değerlerinin indekslenmesi

2 : $SO_o \leftarrow isoutlier(G_o, alg)$

En yüksek S değerine sahip ölçütte sıra dışı oy değerlerinin indekslenmesi

3 : $SO_{S_{max}} \leftarrow isoutlier(G_{s_{max}}, alg)$

4 : $SO = SO_{S_{max}} \cup SO_o$

Sort R vector in descend or ascend order

5 : $Sr \leftarrow sort(|R|, descend)$ ► R vektörünü sırala

6 : **for all items in SO** $(i \leftarrow 1\ to\ k)$ **do**

7 : $\quad R(SO_i) = Sr_i$

8 : **end for**

Nihai maskeleme verisinin oluşturulması

9 : $Rson \leftarrow \boldsymbol{randomize\ index\ of}\ (\forall \boldsymbol{Sr} \in \boldsymbol{R}\ \neg(\boldsymbol{Sr} = \boldsymbol{R}(\boldsymbol{SO})))$

GBS_{max} stratejisinde sıra dışı oy değerlerini belirlemek ve gerçek kullanıcı verisini maskelemek için yapılan işlemler aşağıda maddelenmektedir.

i. *Kullanıcı* × *ürün* matrisindeki G_0 ölçütü dışında bütün alt-ölçütlerin entropi değerleri ayrı ayrı hesaplanır.

ii. Kullanıcıya ait ham derecelendirme değerleri en yüksek entropi katsayısına sahip alt-ölçüt üzerinde, tercih edilen sıra dışı oy değeri belirleme tekniğine göre (SO_{gesd}, SO_{mea} , $SO_{median}, SO_{quartiles}$) sıra dışı oy vektörü ($SO_e$) oluşturulur.

iii. G_0 ölçütüne göre *ii*. adımda kullanılan sıra dışı oy değeri belirleme işlemi yinelenir ve G_0 ölçütüne göre sıra dışı oy vektörü (SO_0) elde edilir.

iv. Nihai SO vektörü $SO_{S_{max}}$ ve SO_0 vektörlerin birleşim kümesi alınarak elde edilir ($SO = \left(SO_{S_{max}} \cup SO_0\right)$).

v. R vektörü negatif ve pozitif sayılardan meydana geldiği için $|R|$ küçükten büyüğe doğru sıralanıp indekslenir.

vi. Her bir alt-ölçüt için; Sr'ın ilk $\#SO$ elemanı sıra dışı oy değerlerini maskeleme için kullanılır.

vii. Kalan normal oy değerlerini gizlemek için; maskeleme vektöründe sıra dışı oy değerlerini gizlemek için kullanılmayan değerler rastgele bir biçimde yeniden sıralanır ve gerçek oy değerlerine eklenir.

6.4.4. Tüm derecelendirmeler (TD)

Her ölçüt için sıra dışı oy değerlerinin ayrı ayrı değerlendirildiği bu stratejide öncelikle aktif kullanıcı oyları üzerinde her bir ölçüt için olağan dışı oy değerleri indekslenmektedir. Bu indekslerin birleşim kümesi nihai sıra dışı oy değerleri olarak işaretlenmektedir. TD stratejisinde sıra dışı oy değerlerini belirlemek ve gerçek kullanıcı verisini maskelemek için yapılan işlemler aşağıda maddelenmektedir.

i. Kullanıcıya ait ham derecelendirme değerlerine sahip olan bütün alt-ölçütler için, belirlenen sıra dışı oy değeri belirleme tekniğine göre ($SO_{gesd}, SO_{mean}, SO_{median}, SO_{quartiles}$) sıra dışı oy vektörleri (SO_k) oluşturulmaktadır.

ii. Bütün ölçütler için ayrı ayrı belirlenen sıra dışı oy değerlerinin birleşim kümesi alınarak ölçütler için nihai SO elde edilir.

iii. R vektörü negatif ve pozitif sayılardan meydana geldiği için $|R|$ küçükten büyüğe doğru sıralanıp indekslenir.

iv. Her ölçüt için Sr'ın ilk *#SO* elemanı sıra dışı oy değerlerini maskeleme için kullanılır.

v. Kalan normal oy değerlerini gizlemek için; maskeleme vektöründe sıra dışı oy değerlerini gizlemek için kullanılmayan değerler rastgele bir biçimde yeniden sıralanır ve gerçek oy değerlerine eklenir.

Bu stratejide kullanılan sıra dışılık fonksiyonu ve veri maskeleme işlemi Prosedür 6.5'de gösterilmektedir.

Prosedür 6.5. *TD stratejisinde kullanılan sıra dışılık belirleme ve veri maskeleme prosedürü*

Sıra dışı oy belirleme stratejinin belirlenmesi

$alg \leftarrow (gesd, mean, median, quartiles)$

Bütün ölçütler üzerinde sıra dışı oy değerlerinin indekslenmesi

$SO_k \leftarrow isoutlier(G_k, alg)$

$SO = SO_1 \cup \cdots \cup SO_k$

$Sr \leftarrow sort(|R|, descend)$ ► R vektörünü sırala

for all items in $\boldsymbol{SO}$ $(i \leftarrow 1\ to\ k)$ **do**

$R(SO_i) = Sr_i$

end for

Nihai maskeleme verisinin oluşturulması

$Rson \leftarrow \boldsymbol{randomize\ index\ of}\ (\forall \boldsymbol{Sr} \in \boldsymbol{R}\ \neg(\boldsymbol{Sr} = \boldsymbol{R}(\boldsymbol{SO})))$

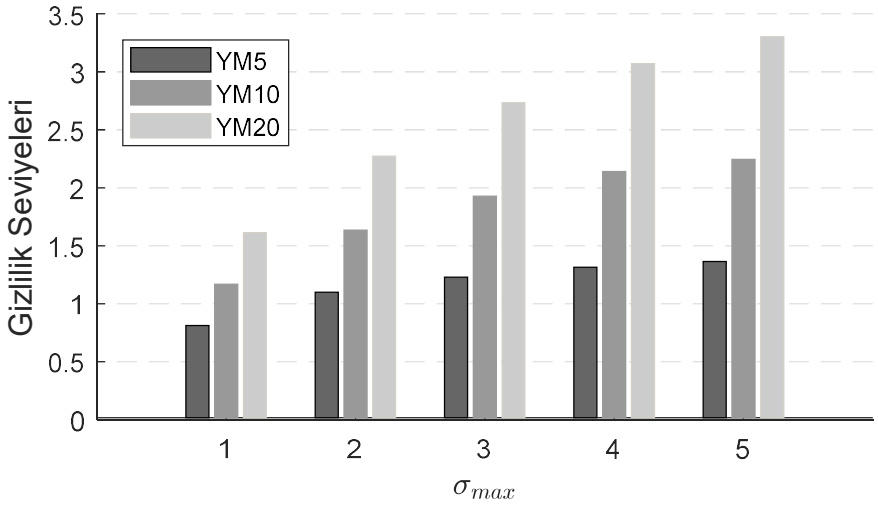
6.5. Deneysel Sonuçlar

Bu bölümde, önerilen sıra dışı oy değerleri belirleme stratejilerinin gizlilik ve öneri üretme doğruluğuna etkisi analiz edilmiştir. Bu amaçla sıra dışı oy değerlerinin belirlenmesinde kullanılan; sıra dışı oy belirleme ve sıra dışı ölçüt belirleme stratejileri *YM5*, *YM10* ve *YM20* veri setleri üzerinde ayrı ayrı değerlendirilip ideal stratejiler belirlenmiştir. Bununla birlikte, $GKOF^k$ sistemlerinde kullanıcı mahremiyet seviyesini belirlemek amacıyla kullanılan σ parametresi [1,5] değer aralığı içerisinde analiz edilmiş ve değişken σ parametresinin öneri üretme doğruluğu üzerindeki etkileri incelenerek üretilen önerilerin istatistiksel olarak anlamlılık düzeyleri test edilmiştir.

6.5.1. Gizlilik analizi

Prosedür 6.1 referans alınarak üretilen maskeleme verisi ile gizliliği sağlanan maskelenmiş kullanıcı verisi *P*'nin gizlilik seviyesi Bölüm 4.4.2'de tanımlanan yöntem kullanılarak ölçeklenmiştir. Gizlilik seviyesini ölçeklemek için kullanılan bu yöntem tek boyutlu bir kullanıcı ve maskeleme vektörü üzerinde gizlilik seviyesini değerlendirebilmektedir. Çoklu-ölçütlü veri setleri birden fazla ölçütten meydana gelmektedir. Bu nedenle, kullanılan deneysel veri setlerinde her bir ölçüt için gizlilik seviyesi değeri ayrı ayrı hesaplanmış ve elde edilen değerler içerisindeki en düşük değer nihai gizlilik seviyesi olarak deneysel sonuçlar bölümünde sunulmuştur.

Bölüm 4 ve 5'te veri gizliliği ve öneri üretme doğruluğu bakımından elde edilen sonuçlar göz önünde bulundurulduğunda; $\mathcal{U}$ dağılıma göre $\mathcal{N}$ dağılım ile üretilen maskeleme verisinin gizlilik ve doğruluk alanında daha iyi sonuçlar verdiği görülmektedir. Bu nedenle, sıra dışı oy değerleri belirleme stratejilerinin veri gizliliği bakımından değerlendirilmesi sürecinde *R* vektörü yalnızca $\mathcal{N}$ dağılım referans alınarak oluşturulmuştur. *RK* yönteminde kullanıcı gizliliği belirlemek için kullanılan σ değeri [1,5] aralığında test edilmiştir. $\mathcal{N}$ dağılımla elde edilen rastgele gürültü vektörünün değişken σ_{max} seviyelerinde elde ettiği gizlilik seviyeleri Şekil 6.1'de *YM20*, *YM10* ve *YM5* veri setleri için gösterilmiştir. Gizlilik seviyesi değerlendirilirken, *R* vektörünün rassallığı nedeniyle ortaya çıkabilecek sapmaları ortadan kaldırmak için yapılan her deney seti aynı parametreler kullanılarak 10 kez tekrarlanmış ve nihai değerler olarak bu değerlerin ortalaması sunulmuştur.



Şekil 6.1. *YM veri setlerinde değişken σ_{max} değerinin gizliliğe etkisi*

RK prosedürü uygulanarak maskelenen verilerden elde edilebilecek gizlilik seviyesi temel olarak iki farklı değişkenden etkilenmektedir. Bu değişkenlerden ilki maskelenecek verinin seyreklik oranı, ikincisi ise veri maskeleme işleminde kullanılan σ_{max} katsayısıdır. Yahoo!Movies veri setinin alt kümelerinden oluşan *YM5*, *YM10* ve *YM20* veri setlerinde ortalama olarak bir kullanıcının oy verdiği ürünlerin toplam ürün sayısına oranı sırasıyla; %0,7; %2,3 ve %16'dır. Oluşturulan maskeleme vektörü ile elde edilebilecek gizlilik seviyesi, veri setini oluşturan ürün sayısı ve bu ürünlerin kullanıcılar tarafından derecelendirilme oranları ile ilişkilidir. Bunun anlamı oluşturulan maskeleme vektörünün kullanıcı ve ürün sayısına orantılandığında veri setlerinin seyreklik seviyelerindeki farklılıklardan ötürü *R* vektörünün genel gizlilik seviyesine etkisi de azalmaktadır. Bu nedenle artan seyreklik ve azalan oy-ürün oranı ile ilişkili olarak elde edilen gizlilik seviyeleri de azalmaktadır. Buna ek olarak sistemin genel gizlilik seviyesi bütün kullanıcılardan elde edilen gizlilik seviyelerinin ortalama değeridir. Oluşturulan maskeleme verisinin kullanıcı verisine sağladığı gizlilik seviyesi hesaplanırken, orijinal kullanıcı verisindeki oy değerlerinin olasılıksal dağılımları kullanılmaktadır. *YM5* veri seti gibi bir kullanıcının sadece 5 adet ürüne oy verebildiği bir sistemde verilen bu oy değerlerinin olasılık dağılımları en rassal oy dağılımında bile en fazla 0,2 olabilecektir. Bu gibi kullanıcılar için elde edilen gizlilik seviyeleri oy dağılımlarındaki rassallık nedeniyle düşük olacağından genel mahremiyet seviyesini de düşürmektedir. Ayrıca; artan σ değeri ile elde edilen gizlilik seviyesi arasında doğrusal bir ilişki gözlemlenmektedir. Artan σ değeri ile orijinal kullanıcı verilerine eklenecek olan maskeleme verilerinin değer aralığı artacağından gizlilik seviyesinin de artması beklenen bir sonuçtur. Bu veriler doğrultusunda, *YM5* veri seti ile elde edilen gizlilik seviyesi *YM10* ve *YM20* veri setlerine göre daha düşük seviyelerdedir.

6.5.2. Doğruluk analizi

Maskeleme verisinin kullanıcı oy verme profiline göre yeniden sıralanmasının önerildiği sıra dışı oy belirleme stratejilerinin öneri üretme doğruluğu bakımından incelenmesi iki ana başlık altında yapılmaktadır. Bunlar; (i) önerilen sıra dışı oy belirleme stratejisinin öneri doğruluğuna etkisi, (ii) önerilen ölçüt belirleme stratejisinin öneri doğruluğuna etkisidir. Sıra dışı oy belirleme ve *R* vektörünün yeniden sıralanması ile ortaya çıkan öneri doğrulukları geleneksel öneri üretme prosedürlerinden $\lfloor\cdot\rfloor$ ve çoklu-değişkenli regresyon (*mvr*) yaklaşımlarına göre oluşturulmuştur. Bu yöntemlerin

değişken σ_{max} parametreleri üzerinde öneri doğruluğuna getirdiği farklılıklar *YM20*, *YM10* ve *YM5* veri setleri üzerinde geleneksel k-en yakın komşuluk strateji kullanılarak test edilmiştir. Her bir veri seti için öneri üretme işleminde komşuluğuna başvurulan kullanıcı sayısı 10 olarak belirlenmiştir. Öneri üretme sürecinde, birini dışarıda bırakarak çapraz doğrulama metodu kullanılarak her bir kullanıcının oy verdiği her bir ürün için ayrı ayrı öneri üretilmiştir.

Deneysel yöntem oluşturulurken, kullanıcıların değişken gizlilik seviyesi ihtiyaçlarını taklit edebilmek adına σ parametresi, sistem tarafından belirlenen [1,5] değer aralığındaki σ_{max} parametrelerinin (0,1] değer aralığındaki rastgele üretilen bir sayı ile çarpılması ile oluşturulmaktadır. Bu işlemler sırasında oluşturulan maskeleme vektörü elemanlarının rastgele diziliminden kaynaklanan sapmaları hafifletmek için her deney seti 10 kez tekrarlanmış ve sunulan nihai mutlak hata değerleri bu değerlerin ortalaması alınarak belirlenmiştir. Elde edilen öneri doğruluklarını ölçeklendirmek için, gerçek oy değerleri ve bu değerler için üretilen tahminler arasındaki ortalama mutlak farkları ölçen istatistiksel doğruluk ölçütü olarak *MAE* kullanılmaktadır.

6.5.2.1. Sıra dışı oy belirleme stratejilerinin öneri doğruluğuna etkisi

Kullanıcıların gerçek oy değerleri üzerinde sıra dışı olarak nitelendirilen oy değerlerini belirlemek için *gesd*, *median*, *mean* ve *quartiles* yöntemleri kullanılmaktadır. Bu yöntemlerin $GKOF^k$ sistemleri üzerinde öneri doğruluğuna etkileri *YM20*, *YM10* ve *YM5* veri setleri kullanılarak çoklu-değişkenli regresyon yaklaşımı ile elde edilen öneri doğrulukları üzerinden karşılaştırılmaktadır. Öneri üretme işleminde sıra dışı oy değeri belirleme stratejileri test edilirken; her ölçüt için sıra dışı oy değerlerini ayrı ayrı indeksleyen *TD* stratejisi kullanılmaktadır. Bu yöntemler ile sıra dışı oy değeri olarak işaretlenen kullanıcı oyları sayısı Tablo 6.1'de gösterilmektedir.

Tablo 6.1. *Sıra dışı oy olarak etiketlenen gerçek kullanıcı verileri sayısı*

	YM5	*YM10*	*YM20*
gesd	5,393	3,015	693
median	14,537	6,344	1,324
mean	859	652	194
quartiles	6,919	3,762	867

YM5 veri seti kullanılarak *TD* stratejisi üzerinde *gesd*, *median*, *mean* ve *quartiles* yöntemleri ile elde edilen sıra dışı oy değerlerine özgü maskeleme işlemi sonucunda ortaya çıkan *MAE* değerleri Tablo 6.2'de gösterilmektedir. Değişken σ_{max} değerleri üzerinde elde edilen en iyi öneri doğrulukları kıyaslandığında, $\sigma_{max} = 3$ değeri dışında bütün değer aralıklarında *mean* ve *quartiles* yöntemleri ile elde edilen sıra dışı oy değerlerine özgü maskeleme işlemi *YM5* veri seti için diğer yöntemlere göre daha doğru öneriler üretmektedir.

Tablo 6.2. *Sıra dışı oy belirleme stratejilerinin YM5 veri setinde genel öneri üretme doğruluğuna etkisi*

	gesd	*median*	*mean*	*quartiles*	$GKOF_{mvr}^{k}$
$\sigma_{max} = 1$	2,1513	2,1534	**2,1473**	2,1483	2,1654
$\sigma_{max} = 2$	2,2769	2,2763	**2,2700**	2,2703	2,3166
$\sigma_{max} = 3$	2,3687	2,3732	2,3695	**2,3649**	2,4088
$\sigma_{max} = 4$	2,5396	2,5388	**2,5332**	2,5391	2,5863
$\sigma_{max} = 5$	2,7000	2,7071	**2,6798**	2,6814	2,7600

Aynı maskeleme prosedürü ile *YM5* veri seti üzerinde sıra dışı oy olarak belirlenen kullanıcı oyları üzerinde elde edilen öneri doğrulukları Tablo 6.3'te gösterilmektedir. Elde edilen doğruluk seviyeleri incelendiğinde en doğru önerilerin üretildiği strateji *median* yöntemidir. Ancak Tablo 6.1'de verilen sıra dışı oy değerleri olarak etiketlenen oy değeri sayıları göz önünde bulundurulduğunda, *median* yöntemi ile etiketlenen oy sayısı diğer yöntemler ile elde edilen oy değerlerinden oldukça fazladır. Tablo 6.3'te gösterilen ve sadece sıra dışı oy değerleri için üretilen *MAE* değerleri analiz edilirken sıra dışı oy değeri olarak nitelendirilen oy değerlerinin sayısı ile hata katsayısı ilişkilidir. Sıra dışı oy değerleri olarak işaretlenen gerçek kullanıcı oyları kullanıcının genel oy verme eğilimiyle uyumsuz olduğu için, yalnızca sıra dışı oy değerleri için üretilen önerilerin yüksek hata katsayılarına sahip olması beklenen bir sonuçtur. Bir veri setinde, sıra dışı oy değerleri sayısı arttırıldıkça, bu değerler kullanıcının gerçek oy verme eğilimine yakınsamaya başlar. Bu nedenle *median* yöntemi ile işaretlenen 14,537 adet oy değeri ile üretilen önerilerin *MAE* değerinin diğer yöntemlere göre düşük olması beklenen bir sonuçtur. Sıra dışı oy değeri olarak işaretlenen oy sayısı arttıkça, kullanıcı profiline yakın oy değerleri de işleme dâhil edildiğinden *MAE* düşecektir. Belirlenen stratejide sıra dışı oy olarak işaretlenen oy değerlerinin gerçek oy değerlerine sayısal olarak oranı mutlak hata bakımından önemlidir. *MAE* değerleri göz önünde bulundurulduğunda *mean* ve *quartiles* yöntemleri en doğru önerilerin üretilmesine neden olmasına rağmen *YM5* veri

seti için *quartiles* yöntemi ideal sıra dışı oy belirleme stratejisi olarak seçilmiştir. *YM5* veri seti gibi az sayıda ürün derecelendirmesine sahip *kullanıcı* × *ürün* vektöründe, *quartiles* yönteminin kullanıcı derecelendirme profilinde ortaya çıkan sıra dışı değerleri belirlemede daha başarılı sonuçlar vermesidir. Örneğin kullanıcı derecelendirme vektörü $x = [1\ 2\ 1\ 2\ 1\ 5]$ olarak verilen bir *kullanıcı* × *ürün* vektöründe, '5' değeri kullanıcının genel oy verme eğiliminden farklılık göstermekte ve sıra dışı oy değeri olarak nitelendirilmektedir. Ancak verilen örnekte *mean* ve *quartiles* yöntemleri ile sıra dışı oy değerlerini belirlediğimizde *mean* yöntemi derecelendirme vektöründe herhangi bir sıra dışı oy değeri belirleyemezken *quartiles* yöntemi ile '5' değerini sıra dışı olarak belirlenmektedir. Bu örnekte, *mean* yönteminin sıra dışı oy değerlerini tespit edememesinin nedeni, derecelendirme vektörünün ortalamasından üç standart sapmadan daha büyük elemanları sıra dışı oy değeri olarak belirlemesidir.

Tablo 6.3. *Sıra dışı oy belirleme stratejilerinin YM5 veri setinde yalnızca sıra dışı oy değerleri için üretilen önerilerin doğruluğu*

	gesd	*median*	*mean*	*quartiles*
$\sigma_{max} = 1$	4,9254	3,3483	6,6004	4,2715
$\sigma_{max} = 2$	5,0133	3,4383	6,6856	4,3930
$\sigma_{max} = 3$	5,0692	3,5035	6,8384	4,4204
$\sigma_{max} = 4$	5,0498	3,6122	6,7991	4,4838
$\sigma_{max} = 5$	5,1084	3,6875	6,8632	4,5168

YM10 veri seti kullanılarak *TD* stratejisi üzerinde *gesd*, *median*, *mean* ve *quartiles* yöntemleri ile elde edilen sıra dışı oy değerlerine özgü maskeleme işlemi sonucunda ortaya çıkan *MAE* değerleri Tablo 6.4'te, yalnızca sıra dışı oy değerleri üzerinde elde edilen öneri doğrulukları Tablo 6.5'te gösterilmektedir. Değişken σ_{max} değerleri üzerinde elde edilen *MAE* değerleri kıyaslandığında, *quartiles* ve *mean* yöntemi ile elde edilen sıra dışı oy değerlerinin öneri kalitesini diğer yöntemlere göre daha çok arttırdığı gözlemlenmektedir.

Tablo 6.4. *Sıra dışı oy belirleme stratejilerinin YM10 veri setinde genel öneri üretme doğruluğuna etkisi*

	gesd	*median*	*mean*	*quartiles*	$GKOF^{k}_{mvr}$
$\sigma_{max} = 1$	2,0034	2,0057	2,0050	**1,9987**	2,0373
$\sigma_{max} = 2$	2,1282	2,1306	2,1286	**2,1221**	2,1946
$\sigma_{max} = 3$	2,3124	2,3258	**2,3111**	2,3173	2,4050
$\sigma_{max} = 4$	2,5084	2,5118	**2,5046**	2,5053	2,6236
$\sigma_{max} = 5$	2,6924	2,6966	2,6866	2,**6850**	2,7943

Yalnızca sıra dışı oy değerleri için üretilen önerilerin *MAE* katsayıları kıyaslandığında elde edilen sonuçlar *YM5* veri setinde elde edilen sonuçlara ve nedenlerine paralellik göstermektedir. *Mean* yöntemi ile daha doğru öneriler üretilmekle birlikte, *YM5* veri seti için de ifade edildiği gibi hata katsayıları sıra dışı oy değeri sayısı ile ilişkilidir. Bu nedenle, *YM10* veri seti için de ideal sıra dışı oy belirleme stratejisi olarak *quartiles* olarak belirlenmiştir.

Tablo 6.5. *Sıra dışı oy belirleme stratejilerinin YM10 veri setinde yalnızca sıra dışı oy değerleri için üretilen önerilerin doğruluğu*

	gesd	*median*	*mean*	*quartiles*
$\sigma_{max} = 1$	4,9359	3,5334	6,4258	4,2462
$\sigma_{max} = 2$	5,1016	3,6921	6,5355	4,3715
$\sigma_{max} = 3$	5,2037	3,8411	6,6481	4,5304
$\sigma_{max} = 4$	5,2539	3,9558	6,6455	4,6035
$\sigma_{max} = 5$	5,3172	4,0412	6,6915	4,7001

YM20 veri setinde *gesd*, *median*, *mean* ve *quartiles* yöntemleri ile elde edilen sıra dışı oy değerlerine özgü maskeleme işlemi sonucunda ortaya çıkan *MAE* değerleri Tablo 6.6'da gösterilmektedir. Değişken σ değerleri üzerinde elde edilen en iyi öneri doğrulukları kıyaslandığında, $\sigma = 1$ değeri dışında bütün değer aralıklarında *gesd* yöntemi ile elde edilen sıra dışı oy değerleri kullanılarak TD prosedürüne göre yeniden sıralanan maskeleme verisinin öneri doğruluğuna katkısı *YM20* veri seti için genel olarak diğer yöntemlere göre daha doğru öneriler üretmektedir.

Tablo 6.6. *Sıra dışı oy belirleme stratejilerinin YM20 veri setinde genel öneri üretme doğruluğuna etkisi*

	gesd	*median*	*mean*	*quartiles*	$GKOF_{mvr}^{k}$
$\sigma_{max} = 1$	1,7296	**1,7236**	1,7254	1,7266	1,7759
$\sigma_{max} = 2$	**1,8530**	1,8651	1,8579	1,8624	1,9730
$\sigma_{max} = 3$	**2,1237**	2,1320	2,1245	2,1402	2,3043
$\sigma_{max} = 4$	**2,3382**	2,3618	2,3550	2,3547	2,5070
$\sigma_{max} = 5$	**2,6358**	2,6494	2,6436	2,6500	2,8015

Yalnızca sıra dışı oy değerleri üzerinde elde edilen öneri doğrulukları Tablo 6.7'de gösterilmektedir. Elde edilen doğruluk seviyeleri incelendiğinde en doğru önerilerin üretildiği strateji *median* yöntemidir. Ancak Tablo 6.1'de verilen sıra dışı oy değerleri olarak etiketlenen kullanıcı oy değerleri göz önünde bulundurulduğunda, *YM20* veri seti

için ideal sıra dışı oy belirleme stratejisi olarak *gesd* olarak belirlenmiştir. Bunun nedeni, *YM20* veri setini oluşturan *kullanıcı* × *ürün* vektörünün *YM5* ve *YM10* veri setlerine göre daha fazla derecelendirilmiş üründen oluşması ve *gesd* yönteminin birbirini maskeleyen birden fazla sıra dışı derecelendirme olduğu durumlarda diğer yöntemlere göre daha iyi performans gösterebilmesidir.

Tablo 6.7. *Sıra dışı oy belirleme stratejilerinin YM20 veri setinde yalnızca sıra dışı oy değerleri için üretilen önerilerin doğruluğu*

	gesd	*median*	*mean*	*quartiles*
$\sigma_{max} = 1$	4,8378	3,3818	6,6609	4,0817
$\sigma_{max} = 2$	5,0591	3,5825	6,8909	4,2532
$\sigma_{max} = 3$	5,2566	3,7663	7,0184	4,4735
$\sigma_{max} = 4$	5,3446	3,9482	7,0284	4,6483
$\sigma_{max} = 5$	5,5602	4,10642	6,9581	4,7701

6.5.2.2. Alt-ölçüt belirleme stratejilerinin öneri doğruluğuna etkisi

Sıra dışı oy değerlerini belirlemek için önerilen stratejilerin hangi ölçüt üzerinde daha etkili sonuçlar verdiğini test etmek için Bölüm 6.4'de tanımlanan GD, S_{max}, GDS_{max} ve TD sıra dışı oy değerlerine özgü maskeleme verisi sıralama stratejileri kullanılmaktadır. Yapılan çalışmada Bölüm 6.5.2.1'de elde edilen sonuçlar doğrultusunda; *YM20* veri seti için *gesd*, *YM10* ve *YM5* veri setleri için *quartiles* stratejilerine göre sıra dışı oy değerleri belirlenmiş ve geleneksel $\lfloor\cdot\rceil$ benzerlik metodolojisi ve *mvr* ile elde edilen *MAE* değerleri karşılaştırılmaktadır.

Önerilen veri gizleme stratejileri ve sıra dışılık belirleme stratejileri kullanılmadan direkt olarak ham maskeleme verileri ($GKOF^k$) ile maskelenen derecelendirme değerleri ile elde edilen öneri doğrulukları regresyon tabanlı yaklaşım ve$\lfloor\cdot\rceil$ tabanlı yaklaşım için Tablo 6.8'de gösterilmektedir. [1,5] değer aralığında değişken σ_{max} parametreleri ile üretilen maskeleme verileri kullanılarak gizlenen kullanıcı oylarında; (i) *YM5* veri seti için $GKOF^k_{mvr}$ yöntemi aracılığıyla üretilen öneri doğrulukları $\lfloor GKOF^k \rceil$ yöntemine göre %7,82 oranında daha doğru sonuçlar üretmektedir, (ii) *YM10* veri setinde de $GKOF^k_{mvr}$ aracılığı ile elde edilen önerilerde $\lfloor GKOF^k \rceil$ yaklaşımına göre ortalama olarak %11,68 oranında daha doğru öneriler üretmektedir, (iii)benzer şekilde; *YM20* veri setinde de $GKOF^k_{mvr}$ aracılığıyla %12,99 oranında daha doğru öneriler üretebilmektedir. Özetle; $GKOF^k_{mvr}$ yöntemi maskelenmiş veri üzerinde daha başarılı öneriler üretmektedir. Bu

nedenle, önerilen ölçüt belirleme stratejileri $GKOF_{mvr}^{k}$ tabanlı yaklaşım ile elde edilen *MAE* değerleri kullanılarak sunulmuştur.

Tablo 6.8. $\lfloor GKOF^{k} \rfloor$ *ve* $GKOF_{mvr}^{k}$ *MAE karşılaştırmaları*

	YM5		*YM10*		*YM20*	
σ	$\lfloor GKOF^{k} \rfloor$	$GKOF_{mvr}^{k}$	$\lfloor GKOF^{k} \rfloor$	$GKOF_{mvr}^{k}$	$\lfloor GKOF^{k} \rfloor$	$GKOF_{mvr}^{k}$
1	2,2041	2,1639	2,1371	2,04274	1,8501	1,7759
2	2,4304	2,2805	2,4275	2,1952	2,1916	1,9729
3	2,6672	2,4277	2,7380	2,38944	2,6760	2,3043
4	2,8824	2,5921	3,0107	2,5793	2,9911	2,5070
5	3,0328	2,7195	3,2566	2,7783	3,3547	2,8015

YM5 veri setinde, önerilen ölçüt belirleme stratejilerinin öneri üretme doğruluğunu etkisi Tablo 6.9'da gösterilmektedir. Değişken σ_{max} parametreleri ile üretilen maskeleme verileri, önerilen alt ölçüt belirleme stratejileri ile yeniden sıralanıp veri maskeleme işlemine dâhil edildiğinde *OF* sisteminin öneri üretme kalitesinde artış elde edilmektedir. Bu yöntemler içerisinde genel olarak en başarılı sonuçların elde edilmesini sağlayan GD prosedürü σ_{max} parametresinin 1'den büyük olduğu bütün değerlerde diğer yöntemlere göre daha doğru öneriler üretmektedir. *GD* stratejisi, $\lfloor GKOF^{k} \rfloor$ yöntemine göre %10,52; $GKOF_{mvr}^{k}$ yöntemine göre %1,88 oranında öneri üretme kalitesinde artış sağlamaktadır. Kullanılan diğer prosedürlerde TD için %1,55; S_{max} için %1,62; GDS_{max} ve için %1,53 oranında öneri kalitesinde artış gözlenmektedir.

Tablo 6.9. *YM5 veri seti için ölçüt belirleme stratejilerinin MAE değerleri*

σ	TD	S_{max}	GD	GDS_{max}
1	2,1514	2,14922	2,1482	**2,1463**
2	2,2496	2,2544	**2,2472**	2,2551
3	2,3890	2,3864	**2,3810**	2,3964
4	2,5385	2,5364	**2,5309**	2,5376
5	2,6663	2,6630	**2,6510**	2,6650

YM10 veri setinde, önerilen ölçüt belirleme stratejilerinin öneri üretme doğruluğuna etkisi Tablo 6.10'da gösterilmektedir. Maskeleme verisi yeniden sıralama stratejilerinin her biri ham maskeleme verileri ile elde edilen önerilerden daha başarılı öneriler üretilmesini sağlamaktadır. Genel olarak en başarılı sonuçların elde edildiği GD stratejisi ile $\lfloor GKOF^{k} \rfloor$ prosedürüne göre %17,21; $GKOF_{mvr}^{k}$ prosedürüne göre %3,52 oranında öneri üretme kalitesinde artış gözlemlenmektedir. Kullanılan diğer prosedürlerde ise;

TD için %3,38; S_{max} için %3,46 ve GDS_{max} için %3,42 oranında öneri kalitesinde artış gözlenmektedir.

Tablo 6.10. *YM10 veri seti için ölçüt belirleme stratejilerinin MAE katsayıları*

σ	TD	S_{max}	GD	GDS_{max}
1	2,0064	2,0046	2,0040	**2,0005**
2	2,1227	2,1239	2,1233	**2,1215**
3	2,3049	2,3017	2,3006	**2,2989**
4	2,4852	2,4895	**2,4831**	2,4852
5	2,6747	2,6648	**2,6600**	2,6824

YM20 veri setinde, önerilen ölçüt belirleme stratejilerinin öneri üretme doğruluğuna etkisi

Tablo ***6.11***'de gösterilmektedir. Önerilen stratejilerin her biri ham maskeleme verileri ile elde edilen önerilerden daha başarılı öneriler üretilmesini sağlamaktadır. Elde edilen hata katsayılarında ortalama olarak en başarılı sonuçlar S_{max} ve GD prosedürü ile elde edilmektedir. Daha yüksek seviyelerdeki σ değerleri ile maskelemede daha başarılı sonuçlar elde edilmesini sağlayan yaklaşım GD stratejisi olduğu için *YM20* veri seti için hata değerleri bu prosedür ideal olarak kabul edilmiştir. GD stratejisi, $\lfloor GKOF^k \rfloor$ prosedürüne göre %22,01, $GKOF^k_{mvr}$ stratejisine göre %6,17 oranında öneri üretme kalitesinde artış elde edilmektedir. Kullanılan diğer stratejilerde ise TD %5,88; S_{max} için %5,99 ve S_{max} için %5,65 oranında öneri kalitesinde artış sağlamaktadır.

Tablo 6.11. *YM20 veri seti için ölçüt belirleme stratejilerinin MAE katsayıları*

σ	TD	S_{max}	GD	GDS_{max}
1	1,7359	**1,72765**	1,7354	1,7349
2	1,8460	**1,8459**	1,8563	1,8543
3	2,0948	2,1063	**2,0845**	2,0976
4	2,3777	2,3588	**2,3584**	2,3859
5	2,6521	2,6570	**2,6428**	2,6577

6.5.3. İstatistiksel anlamlılık

TD, S_{max}, GD ve GDS_{max} stratejileri ile elde edilen öneri doğruluklarının, $GKOF^k_{mvr}$ referans stratejisi ile üretilen öneri doğrulukları ile arasında istatistiksel olarak anlamlı bir farklılık olup-olmadığını test edebilmek için *t-testi* kullanılmıştır. Yapılan tek yönlü *t-testi* sonucunda elde edilen p değerleri Tablo 6.12'de gösterilmektedir. Ölçüt

belirleme stratejileri için yapılan tek yönlü hipotezleri sonuçlarında bütün stratejilerin %99 güven düzeyinde istatistiksel olarak anlamlı olduğu görünmektedir.

Tablo 6.12. *Önerilen yaklaşımlar ile elde edilen öneri doğruluğu artışlarının istatistiksel anlamlılıkları*

	TD	S_{max}	GD	GDS_{max}
YM5	0,0039*	0,0045*	0,0045*	0,0042*
YM10	0,0013*	0,0015*	0,0014*	0,0007*
YM20	0,0016*	0,0018*	0,0025*	0,0014*

* %99 güven seviyesi

7. SONUÇLAR

Öneri sistemleri, modern zamanlardaki aşırı bilgi yükü sorunuyla başa çıkmada son derece başarılı sonuçlar elde etmektedir. Ancak bireylerin mahremiyetine, kullanıcıların bu sistemlerin faydalarından istifade etmelerini engelleyebilecek seviyede, ciddi tehditler oluşturmaktadır. Öneri teknolojilerinde nispeten yeni bir yaklaşım olan OF^k sistemleri, kullanıcıların genel tercihlerinin yanı sıra ürünlere ait alt-ölçütlerle ilgili beğeni değerlerini de toplayarak bu tehditleri daha da artırmaktadır. Çoklu-ölçütlü veri setleri kullanıcılara ek gizlilik riskleri getirmesine rağmen, *ÖS* alanında gün geçtikçe daha da yaygın hale gelmektedir.

Bu tezde, çoklu-ölçütlü tercih verilerinin kullanımıyla ortaya çıkabilecek mahremiyet riskleri ve bu riskleri gidermeye yönelik bireylerin gizliliklerini tehlikeye atmadan yüksek doğruluk seviyesinde öneriler üretebilecek yaklaşımlar sunulmaktadır. Önerilen yaklaşımlar, sağladığı gizlilik düzeyleri ve öneri üretme doğruluğu açısından analiz edilmektedir. Gerçek kullanıcı verileri üzerinde gerçekleştirilen deneyler, önerilen gizlilik koruma yaklaşımlarının belirlenen amaçlara ulaştığını göstermektedir. Tezin hedeflenen amaçlar doğrultusunda sağladığı temel katkılar aşağıda özetlenmektedir.

- Çoklu-ölçütlü tercih verileri, bir kullanıcının belirli bir ürünü veya hizmeti neden beğendiğini veya beğenmediğini anlama fırsatını vererek öneri hizmet kalitesini iyileştirmeye yardımcı olsa da, aynı zamanda bireylerin gizliliğini daha çok tehlikeye atmaktadır. Bu tezde, kullanıcı verisine doğrudan ya da dolaylı yollarla erişim sağlayan kötü niyetli kişilerin ortaya çıkarabileceği mahremiyet riskleri çoklu-ölçütlü veri seti bakış açısıyla ele alınıp tanımlanmıştır.
- Çoklu-ölçütlü tercih verileri alanındaki gizlilik risklerini hafifletmek için *GKOF* sistemlerinde kullanılan *RKT* temelli gizlilik koruyucu yaklaşımlar çoklu-ölçütlü veri setlerine adapte edilmiştir. Böylece, çoklu-ölçütlü *OF* sistemleri için *RK* ve *RD* yaklaşımları ile elde edilen referans gizlilik ve doğruluk seviyeleri elde edilmiştir.
- Geleneksel veri maskeleme yaklaşımlarının çoklu-ölçütlü veri setlerinde kullanılmasıyla ortaya çıkan en büyük dezavantaj, veri mahremiyetini sağlarken önerilen tahminlerin doğruluğunda büyük kayıplara neden olmasıdır. Bu nedenle, elde edilen gizlilik seviyeleri ile tahmin doğruluğu arasında bir denge kurulması gerekmektedir. Bu amaçla, maskeleme

işlemi sırasında ortaya çıkacak öneri doğruluk kayıplarını hafifletmeyi hedefleyen, kullanıcı ve kullanıcının ölçütleri değerlendirme alışkanlıklarına göre maskeleme işleminde kullanılacak gizlilik parametrelerini belirleyen entropi tabanlı yeni gizlilik-koruma protokolleri geliştirilmiştir. Böylece, geleneksel veri maskeleme yaklaşımı ile elde edilen gizlilik seviyeleri muhafaza edilirken aynı zamanda öneri üretme doğruluğunda ortaya çıkan kayıplar hafifletilmiştir.

- Öneri sistemlerinde, öneri doğruluğunda ortaya çıkan kayıpların diğer bir nedeni de kullanıcıların sıra dışı oy verme eğilimleridir. Bu problemin maskelenmiş veri üzerinde ortaya çıkardığı öneri doğruluğu kayıplarını hafifletmek için ürün tabanlı sıra dışılık belirleme stratejileri sunulmuştur. Bu yaklaşımlarda, aktif kullanıcının oy verme eğilimi analiz edilip, kullanıcının oy verme eğilimine göre standart dışı oy değerleri ile derecelendirdiği ürün listesi belirlenmekte ve bu derecelendirmelere özgü olarak yeni veri maskeleme yaklaşımı uygulanarak öneri doğruluğunda ortaya çıkan kayıp hafifletilmektedir.

OF^k sistemlerinde gizliliği korumak için kullanılan veri maskeleme yaklaşımları gerçek kullanıcı verilerinde bozulmalara neden olmakta ve bunun sonucu olarak öneri doğruluğunda kayıplara neden olmaktadır. Buna karşılık, sunulan veri gizleme yaklaşımları bu kayıpları önemli ölçütle hafifletebilmektedir. Elde edilen sonuçlar genel olarak değerlendirildiğinde, uygulanan her bir maskeleme stratejisi, geleneksel yaklaşımlar ile benzer veri gizlilik seviyeleri elde etse de öneri üretme kalitesi bakımından önemli kazanımlar sağlamaktadır.

Yapılan çalışma kullanıcı gizliliğini sağlamak adına gelecek vaad etse de üzerinde çalışılması gereken zayıflıklar mevcuttur. Bunlardan ilki veri maskeleme işlemi ile ortaya çıkan öneri kayıplarının hafifletilmesi için öneri üretme işleminde yeni yaklaşımların geliştirilmesine ihtiyaç duyulmaktadır.

RKT'nin kullanıldığı veri maskeleme işlemlerinde, kullanıcı mahremiyet seviyesini belirlemek bireylerin ihtiyaçları doğrultusunda gerçekleştirilen sezgisel bir süreçtir. Yani kişi kendi ihtiyaç duyduğu mahremiyet düzeyine göre gizlilik seviyesini belirler ve gerçek oy değerleri o seviyeye göre maskelenir. Ancak, kullanıcıya bırakılan bu süreç beraberinde sisteme ekstra doğruluk kayıplarını da getirmektedir. Kullanıcı belirlediği mahremiyet seviyesinin karşılığında elde edebileceği gizlilik seviyesi ile ilgili bir

çıkarımda bulunamadığı için sezgisel olarak en yüksek seviyeyi belirlemeye çalışacak ve bu davranış sistemde fazladan doğruluk kayıplarına neden olacaktır. Bu nedenle, kullanıcı istekleri ve sistemin ihtiyaçları doğrultusunda gizlilik seviyesi belirleyecek ve bu işlemi otomatize edecek yeni yaklaşımlara ihtiyaç duyulmaktadır.

KAYNAKÇA

Ackerman, M. S., Cranor, L. F. ve Reagle, J. (1999). Privacy in e-commerce: examining user scenarios and privacy preferences. Proceedings of the first ACM conference on electronic commerce, 1–8.

Adomavicius, G. ve Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE transactions on knowledge and data engineering, 17(6), 734-749.

Adomavicius, G ve Kwon Y. (2007). New recommendation techniques for multi-criteria rating systems. IEEE Intelligent Systems, 22(3), 48-55.

Adomavicius, G ve Kwon, Y. (2015). Multi-criteria recommender systems. Ricci, Rokach ve Shapira, Recommender systems handbook, 847–880, Springer, US.

Aggarwal, C.C. (2016b). An introduction to recommender systems. Recommender Systems. Springer, 1–28.

Agrawal, D. ve Aggarwal, C. C. (2001). On the design and quantification of privacy preserving data mining algorithms. Proceedings of the 20th ACM SIGMOD-SIGACT-SIGART symposium on principles of database systems, ACM, 247–255.

Agrawal, R. ve Srikant, R. (2000). Privacy-preserving data mining. Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 439–450.

Badsha, S., Yi, X. ve Khalil, I. (2016). A practical privacy-preserving recommender system. Data Science and Engineering, 1(3), 161–177.

Berkovsky, S., Eytani, Y., Kuflik T. ve Ricci F. (2007). Enhancing privacy and preserving accuracy of a distributed collaborative filtering. In Proceedings of the 2007 ACM Conference on Recommender Systems. ACM, 9–16.

Berkovsky, S., Kuflik, T. ve Ricci, F. (2012). The impact of data obfuscation on the accuracy of collaborative filtering. Expert Systems with Applications, 39(5), 5033-5042.

Bilge, A., Kaleli, C., Yakut, İ., Güneş, İ. ve Polat, H. (2013). A survey of privacy-preserving collaborative filtering schemes. International Journal of Software Engineering and Knowledge Engineering, 23(08), 1085–1108.

Bilge, A. ve Yargıç, A. (2017). Improving accuracy of multi-criteria collaborative filtering by normalizing user ratings. Anadolu Üniversitesi Bilim ve Teknoloji Dergisi A - Uygulamalı Bilimler ve Mühendislik, 18 (1), 225-237.

Bilge, A. ve Polat, H. (2013). A scalable privacy-preserving recommendation scheme via bisecting k-means clustering. Information Processing & Management, 49 (4), 912–927.

Bobadilla, J., Ortega, F., Hernando, A. ve Guti´errez, A. (2013). Recommender systems survey. Knowledge-Based Systems, 46, 109–132.

Boutet, A., Frey, D., Guerraoui, R., Je´gou, A. ve Kermarrec, A.M. (2016). Privacy-reserving distributed collaborative filtering. Computing, 98 (8), 827–846.

Breese, J.S., Heckerman, D. ve Kadie, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering. In Proceedings of the 14th Annual Conference on Uncertainty in Artificial Intelligence, 43–52.

Burke, R., (2002). Hybrid recommender systems: Survey and experiments. User Modeling and User-Adapted Interaction, 12 (4), 331-370.

Canny J. (2002). Collaborative filtering with privacy. In Proceedings of IEEE Symposium on Security and Privac,. IEEE, 45–57.

Canny, J. (2002b), Collaborative filtering with privacy via factor analysis. In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere, Finland, 238-245.

Cantador, I., Fernández-Tobías, I., Berkovsky, S. ve Cremonesi, P. (2015). Cross-domain recommender systems. Ricci, Rokach ve Shapira, Recommender Systems Handbook. Springer, Boston, MA, 919-959.

Casino, F., Domingo-Ferrer, J., Patsakis, C., Puig, D. ve Solanas, A. (2015). A k-anonymous approach to privacy preserving collaborative filtering. Journal of Computer and System Sciences, 81 (6), 1000–1011.

Chaabane A., Acs G. ve Kaafar M. A. (2012). You are what you like! information leakage through users interests. In Proceedings of the 19th Annual Network & Distributed System Security Symposium (NDSS).

Chelishchev, P., Popov, A. ve Sørby, K. (2018). An investigation of outlier detection procedures for CMM measurement data. In MATEC Web of Conferences.

Chen R., Xie M. ve Lakshmanan L. V. (2014). Thwarting passive privacy attacks in collaborative filtering. In Proceedings of the International Conference on Database Systems for Advanced Applications, Springer, 218–233.

Chen, X. ve Huang, V. (2012). Privacy preserving data publishing for recommender system. In proceedings of the 36th computer software and applications conference workshops (COMPSACW), IEEE, 128–133.

Choi, K. ve Suh, Y. (2013). A new similarity function for selecting neighbors for each target item in collaborative filtering. Knowledge-Based Systems, 37, 146-153.

Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D. ve Sartin, M. (1999). Combining content-based and collaborative filters in an online newspaper. In Proceedings of ACM SIGIR Workshop on Recommender Systems, 60.

Cranor, L. F. (2004). I didnt buy it for myself. Designing Personalized User Experiences in eCommerce, Springer, 57–73.

Culnan, M. J. (1993). How did they get my name?: An exploratory investigation of consumer attitudes toward secondary information use. MIS quarterly, 341–363.

Deshpande, M. ve Karypis, G. (2004). Item-based top-N recommendation algorithms. ACM Transaction on Information Systems, 22(1), 143–177.

Dinev T., Xu H., Smith J. H. ve Hart P., (2013). Information privacy and correlates: an empirical attempt to bridge and distinguish privacy-related concepts. European Journal of Information Systems, 22(3), 295–316.

Dwork, C., McSherry, F., Nissim, K. ve Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. Theory of Cryptography, Lecture Notes in Computer Science, 3876, 265–284.

Ekstrand, M.D., Riedl, J.T. Ve Konstan, J.A. (2011). Collaborative filtering recommender systems. Foundations and Trends in Human-Computer Interaction 4(2), 81–173.

Elmisery, A. M. ve Botvich, D. (2017). An enhanced middleware for collaborative privacy in IPTV recommender services.

Erkin, Z., Veugen, T., Toft, T. ve Lagendijk, R. L. (2012). Generating private recommendations efficiently using homomorphic encryption and data packing. IEEE Transactions on Information Forensics and Security, 7 (3), 1053–1066.

Fan, J. ve Xu, L. (2013). A robust multi-criteria recommendation approach with preference-based similarity and support vector machine. Advances in Neural Networks–ISNN 2013, Springer, 385–394.

Friedman A., Knijnenburg B. P., Vanhecke K. , Martens L. ve Berkovsky S., (2015). Privacy aspects of recommender systems, Ricci, Rokach ve Shapira, Recommender systems handbook, Springer,649–688.

Fuchs, M. ve Zanker, M. (2012). Multi-criteria ratings for recommender systems: an empirical analysis in the tourism domain. In International Conference on Electronic Commerce and Web Technologies, Springer, Berlin, Heidelberg, 100-111.

Gao, L. ve Li, C. (2008). Hybrid personalized recommended model based on genetic algorithm. In Wireless Communications, Networking and Mobile Computing. WiCOM'08. 4th International Conference on IEEE, 1-4.

Ghazanfar, M. ve Prugel-Bennett, A. (2011). Fulfilling the needs of gray-sheep users in recommender systems, a clustering solution. In Proceedings of the 2011 International Conference on Information Systems and Computational Intelligence,18–20.

Ghazanfar, M.A. ve Prugel-Bennett, A. (2014). Leveraging clustering approaches to solve the gray-sheep users problem in recommender systems. Expert Systems with Application. 41(7), 3261–3275.

Ghorbani, H. ve Novin, A. H. (2016). An introduction on separating gray-sheep users in personalized recommender systems using clustering solution. International Journal of Computer Science and Software Engineering, 5 (2), 14–18.

Golbeck J. (2016). User privacy concerns with common data used in recommender systems. In International Conference on Social Informatics. Springer, 468–480.

Goldberg, D., Nichols, D., Oki, B. M., & Terry, D. (1992). Using collaborative filtering to weave an information tapestry. Communications of the ACM, 35(12), 61-70.

Gong, S. (2011). Privacy-preserving collaborative filtering based on randomized perturbation techniques and secure multiparty computation. International Journal of Advancements in Computing Technology, 3(4), 89–99.

Goodwin, C. (1991). Privacy: Recognition of a consumer right. Journal of Public Policy & Marketing, 149-166.

Gras, B., Brun, A. ve Boyer, A. (2016). Identifying grey sheep users in collaborative filtering: a distribution-based technique. In Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization, ACM, 17-26.

Grcar, M., Fortuna, B., Mladenic, D. ve Grobelnik, M. (2006). k-NN versus SVM in the collaborative filtering framework. Data Science and Classification, 251–260.

Gross, R. ve Acquisti, A. (2005). Information revelation and privacy in online social networks. In Proceedings of the 2005 ACM workshop on Privacy in the electronic society, ACM, 71-80.

Guerraoui, R., Kermarrec, A. M., Patra, R. ve Taziki, M. (2015). D2p: distance-based differential privacy in recommenders. Proceedings of the VLDB Endowment, 8(8), 862-873.

Herlocker, J. L., Konstan, J. A., Borchers, A. ve Riedl, J. (1999). An algorithmic framework for performing collaborative filtering. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 230-237.

Herlocker, J. L., Konstan, J. A., Terveen, L. G. ve Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems (TOIS), 22 (1), 5–53.

Herlocker, J., Konstan, J.A. ve Riedl, J. (2002). An empirical analysis of design choices in neighborhood-based collaborative filtering algorithms. Information Retrieval 5(4), 287–310.

Ho, Y., Fong, S. Ve Yan, Z. (2007). A Hybrid GA-based Collaborative Filtering Model for Online Recommenders. In International Conference on e-Business, 200-203.

Hou, M., Wei, R., Wang, T., Cheng, Y. ve Qian, B. (2018). Reliable medical recommendation based on privacy-preserving collaborative filtering. CMC-Computers Materials & Continua, 56 (1), 137–149.

Jannach, D., Karakaya, Z. ve Gedikli, F. (2012). Accuracy improvements for multi-criteria recommender systems. In Proceedings of the 13th ACM conference on electronic commerce, ACM, 674-689.

Jeckmans, A. J., Beye, M., Erkin, Z., Hartel, P., Lagendijk, R. L. ve Tang, Q. (2013). Privacy in recommender systems. In Social media retrieval, Springer, London, 263-281.

Jin, R. ve Si, L. (2004). A study of methods for normalizing user ratings in collaborative filtering. In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 568-569.

Kim, H. N., Alkhaldi, A., El Saddik, A. ve Jo, G. S. (2011). Collaborative user modeling with user-generated tags for social recommender systems. Expert Systems with Applications, 38(7), 8488-8496.

Leys, C., Ley, C., Klein, O., Bernard, P. Ve Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. Journal of Experimental Social Psychology, 49(4), 764-766.

Li, D., Chen, C., Lv, Q., Shang, L., Zhao, Y., Lu, T. ve Gu, N. (2016). An algorithm for efficient privacy-preserving item-based collaborative filtering. Future Generation Computer Systems 55, 311–320.

Li, D., Lv, Q., Shang, L. ve Gu, N. (2017a). Efficient privacy-preserving content recommendation for online social communities. Neurocomputing, 219, 440–454.

Li, Q., Wang, C. ve Geng, G. (2008). Improving personalized services in mobile commerce by a novel multicriteria rating approach. In Proceedings of the 17th international conference on World Wide Web, ACM, 1235-1236.

Li, Y., Liu, S., Wang, J. ve Liu, M. (2017b). A Local-Clustering-Based Personalized Differential Privacy Framework for User-Based Collaborative Filtering. In International Conference on Database Systems for Advanced Applications, Springer, Cham, 543-558.

Linden, G., Smith, B. ve York, J. (2003). Amazon.com recommendations: Item-to-item collaborative filtering. IEEE Internet Computing 7(1), 76–80.

Liu, X., Liu, A., Zhang, X., Li, Z., Liu, G., Zhao, L. ve Zhou, X. (2017). When differential privacy meets randomized perturbation: A hybrid approach for privacy-preserving recommender system. In International Conference on Database Systems for Advanced Applications, Springer, Cham, 576-591.

Luo, X., Xia, Y. ve Zhu, Q. (2012). Incremental collaborative filtering recommender based on regularized matrix factorization. Knowledge-Based Systems, 27, 271-280.

Malhotra, N. K., Kim, S. S. ve Agarwal, J. (2004). Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model. Information systems research, 15(4), 336-355.

McCrae, J., Piatek, A. ve Langley, A. (2004). Collaborative filtering. http:// www. imperialviolet. org.

McSherry, F. ve Mironov, I. (2009). Differentially private recommender systems: Building privacy into the netflix prize contenders. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 627-636.

Mekovec, R. ve Vrcek, N. (2011) Factors that influence internet users privacy perception, in Proceedings of the 33rd International Conference on Information Technology Interfaces,227–232.

Moore, A. (2008). Defining privacy. Journal of Social Philosophy, 39(3), 411-428.

Moreno, M. N., Segrera, S., López, V. F., Muñoz, M. D. ve Sánchez, Á. L. (2016). Web mining based framework for solving usual problems in recommender systems. A case study for movies′ recommendation. Neurocomputing, 176, 72-80.

Naak, A., Hage, H. ve Aimeur, E. (2009). A multi-criteria collaborative filtering approach for research paper recommendation in papyres. In International Conference on E-Technologies, Springer, Berlin, Heidelberg, 25-39.

Nilashi, M., Jannach, D., bin Ibrahim, O. ve Ithnin, N. (2015). Clustering-and regression-based multi-criteria collaborative filtering with incremental updates. Information Sciences, 293, 235-250.

Nilashi, M., bin Ibrahim, O., ve Ithnin, N. (2014). Hybrid recommendation approaches for multi-criteria collaborative filtering. Expert Systems with Applications, 41(8), 3879-3900.

Ning, X., Desrosiers, C. ve Karypis, G. (2015). A comprehensive survey of neighborhood-based recommendation methods. In Recommender systems handbook, Springer, Boston, MA, 37-76.

O'Connor, M. ve Herlocker, J. (1999). Clustering items for collaborative filtering. In Proceedings of the ACM SIGIR workshop on recommender systems, 128, UC Berkeley.

Parameswaran, R. ve Blough, D. M. (2007). Privacy preserving collaborative filtering using data obfuscation. In Granular Computing, GRC 2007. IEEE International Conference on IEEE, 380-380.

Parent, W. A. (1983). A new definition of privacy for the law. Law and Philosophy, 2(3), 305-338.

Park, M. H., Hong, J. H. ve Cho, S. B. (2007). Location-based recommendation system using bayesian user's preference model in mobile devices. In International Conference on Ubiquitous Intelligence and Computing, Springer, Berlin, Heidelberg, 1130-1139.

Parker R. B. (1973). A definition of privacy. Rutgers L. Rev., 27(275).

Phelps, J., Nowak, G. Ve Ferrell, E. (2000). Privacy concerns and consumer willingness to provide personal information. Journal of Public Policy & Marketing, 19(1), 27-41.

Polat, H. ve Du, W. (2005a). Privacy-preserving collaborative filtering. International journal of electronic commerce, 9(4), 9-35.

Polat, H. ve Du, W. (2005b). Privacy-preserving collaborative filtering on vertically partitioned data. In European Conference on Principles of Data Mining and Knowledge Discovery, Springer, Berlin, Heidelberg, 651-658.

Polat, H. ve Du, W. (2003). Privacy-preserving collaborative filtering using randomized perturbation techniques. Third IEEE International Conference on IEEE Data Mining, ICDM 2003, 625-628.

Polatidis, N., Georgiadis, C. K., Pimenidis, E. ve Mouratidis, H. (2017). Privacy-preserving collaborative recommendations based on random perturbations. Expert Systems with Applications, 71, 18-25.

Ramakrishnan, N., Keller, B. J., Mirza, B. J., Grama, A. Y. ve Karypis, G. (2001). Privacy risks in recommender systems. IEEE Internet Computing, (6), 54-62.

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P. ve Riedl, J. (1994). GroupLens: an open architecture for collaborative filtering of netnews. In Proceedings of the 1994 ACM conference on Computer supported cooperative work, ACM, 175-186.

Resnick, P. ve Varian, H. R. (1997). Recommender systems. Communications of the ACM, 40(3), 56-58.

Ricci, F., Rokach, L., and Shapira, B. (2015). Recommender systems: Introduction and challenges, Recommender Systems Handbook. Springer, pp. 1–34.

Roh, T. H., Oh, K. J. ve Han, I. (2003). The collaborative filtering recommendation based on SOM cluster-indexing CBR. Expert systems with applications, 25(3), 413-423.

Ruiz-Montiel, M. ve Aldana-Montes, J. F. (2009). Semantically enhanced recommender systems. In OTM Confederated International Conferences On the Move to Meaningful Internet Systems, Springer, Berlin, Heidelberg, 604-609.

Sahoo, N., Krishnan, R., Duncan, G. ve Callan, J. (2012). Research note—the halo effect in multicomponent ratings and its implications for recommender systems: The case of yahoo! movies. Information Systems Research, 23(1), 231-246.

Samatthiyadikun, P., Takasu, A. ve Maneeroj, S. (2013). Bayesian model for a multicriteria recommender system with support vector regression. In 2013 IEEE 14th International Conference on Information Reuse & Integration (IRI), IEEE, 38-45.

Sánchez-Moreno, D., González, A. B. G., Vicente, M. D. M., Batista, V. F. L. ve García, M. N. M. (2016). A collaborative filtering method for music recommendation using

playing coefficients for artists and users. Expert Systems with Applications, 66, 234-244.

Sarwar, B., Karypis, G., Konstan, J., Reidl, J.: Item-based collaborative filtering recommendation algorithms. In:WWW'01: Proc. of the 10th Int. Conf. onWorldWideWeb, pp. 285–295. ACM, New York, NY, USA (2001)

Schoeman, F. D. (1984). Philosophical dimensions of privacy: An anthology. Cambridge University Press.

Shannon, C. E. (2001). A mathematical theory of communication. ACM SIGMOBILE mobile computing and communications review, 5 (1), 3–55.

Shen, Y. ve Jin, H. (2014). Privacy-preserving personalized recommendation: An instance-based approach via di erential privacy. Proceedings of IEEE international conference on data mining,IEEE, 540–549.

Shmueli, E. ve Tassa, T. (2017). Secure multi-party protocols for item-based collaborative filtering. In Proceedings of the Eleventh ACM Conference on Recommender Systems, ACM, 89-97.

Shyong, K., Frankowski, D. ve Riedl, J. (2006). Do you trust your recommendations? An exploration of security and privacy issues in recommender systems. In Emerging trends in information and communication security, Springer, Berlin, Heidelberg, 14-29.

Si, L. ve Jin, R. (2003). Flexible mixture model for collaborative filtering. In Proceedings of the 20th International Conference on Machine Learning, 704-711.

Spiekermann, S., Grossklags, J. ve Berendt, B. (2001). E-privacy in 2nd generation E-commerce: privacy preferences versus actual behavior. In Proceedings of the 3rd ACM conference on Electronic Commerce, ACM, 38-47.

Su, X., Khoshgoftaar T. M. (2009). A survey of collaborative filtering techniques. Advances in artificial intelligence, 2009(4).

Takács, G., Pilászy, I., Németh, B. ve Tikk, D. (2008). Investigation of various matrix factorization methods for large recommender systems. In Data Mining Workshops, 2008. ICDMW'08. IEEE International Conference on . IEEE, 553-562.

Takács, G., Pilászy, I., Németh, B. Ve Tikk, D. (2009). Scalable collaborative filtering approaches for large recommender systems. Journal of machine learning research, 10, 623-656.

Tang, T. Y. ve McCalla, G. (2009). The pedagogical value of papers: a collaborative-filtering based paper recommender. Journal of Digital Information, 10(2).

Tufekci, Z. (2008). Can you see me now? Audience and disclosure regulation in online social network sites. Bulletin of Science, Technology & Society, 28(1), 20-36.

Waldo J. , Lin H. ve Lynette M. (2007) , Engaging privacy and information technology in a digital age. National Academies Press.

Wang, C., Caja, J. ve Gómez, E. (2018). Comparison of methods for outlier identification in surface characterization. Measurement, 117, 312-325.

Warren, S. D., & Brandeis, L. D. (1890). The right to privacy. Harvard law review, 193-220.

Wei, R., Tian, H. ve Shen, H. (2018). Improving k-anonymity based privacy preservation for collaborative filtering. Computers & Electrical Engineering, 67, 509-519.

Weinsberg, U., Bhagat, S., Ioannidis, S. ve Taft, N. (2012). BlurMe: Inferring and obfuscating user gender based on ratings. In Proceedings of the sixth ACM conference on Recommender systems, ACM, 195-202.

Westin, A. F. ve Maurici, D. (1998). E-commerce & privacy: What net users want. Hackensack, NJ: Privacy & American Business.

Westin, A. F. (1968). Privacy and freedom. Washington and Lee Law Review, 25(1), 166.

Yargıç, A. ve Bilge, A. (2017). Privacy Risks for Multi-Criteria Collaborative Filtering Systems. Proceedings of the 26th International Conference of Computer Communication and Networks (ICCCN), IEEE, 1-6.

Yager, R. R. (2003). Fuzzy logic methods in recommender systems. Fuzzy Sets and Systems, 136(2), 133-149.

Zhang, Y., Zhuang, Y., Wu, J. ve Zhang, L. (2009). Applying probabilistic latent semantic analysis to multi-criteria recommender system. Ai Communications, 22(2), 97-107.

Zheng, Y., Agnani, M. ve Singh, M. (2017). Identification of Grey Sheep Users by Histogram Intersection in Recommender Systems. In International Conference on Advanced Data Mining and Applications, Springer, Cham, 148-161.

Zheng, Y., Agnani, M. ve Singh, M. (2017). Identifying grey sheep users by the distribution of user similarities in collaborative filtering. In Proceedings of the 6th Annual Conference on Research in Information Technology, ACM, 1-6.

Zou, J. ve Fekri, F. (2015). A belief propagation approach to privacy-preserving item-based collaborative filtering. IEEE Journal of Selected Topics in Signal Processing, 9(7), 1306-1318.