

EGE ÜNİVERSİTESİ FEN BİLİMLERİ ENSTİTÜSÜ

(DOKTORA TEZİ)

**BİYOMEDİKAL BİLGİ ÇİZGELERİ İÇİN
MAKİNE ÖĞRENMESİ TABANLI ANLAMSAL
BAĞ TAHMİNİ**

Remzi ÇELEBİ

Tez Danışmanı: Dr. Öğr. Üyesi Özgür GÜMÜŞ

Bilgisayar Mühendisliği Anabilim Dalı

Sunuş Tarihi: 13.07.2018

**Bornova-İZMİR
2018**

Remzi ÇELEBİ tarafından **DOKTORA TEZİ** olarak sunulan "**BİYOMEDİKAL BİLGİ ÇİZGELERİ İÇİN MAKİNE ÖĞRENMESİ TABANLI ANLAMSAL BAĞ TAHMİNİ**" başlıklı bu çalışma E.Ü. Lisansüstü Eğitim ve Öğretim Yönetmeliği ile E.Ü. Fen Bilimleri Enstitüsü Eğitim ve Öğretim Yönergesi'nin ilgili hükümleri uyarınca tarafımızdan değerlendirilerek savunmaya değer bulunmuş ve **13.07.2018** tarihinde yapılan tez savunma sınavında aday oybirliği/oyçokluğu ile başarılı/başarısız bulunmuştur.

Jüri Üyeleri:

Jüri Başkanı: Dr. Öğr. Üyesi Özgür GÜMÜŞ

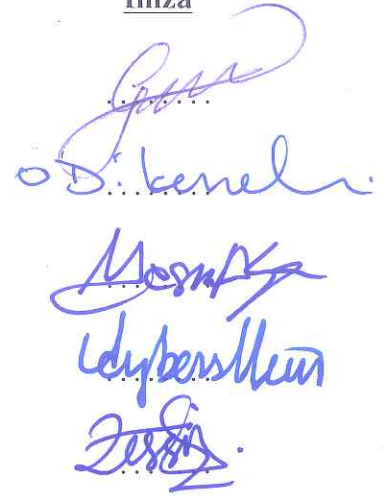
Raportör Üye: Prof. Dr. Oğuz DİKENELLİ

Üye: Doç. Dr. Yeşim AYDIN SON

Üye: Prof. Dr. Aybars UĞUR

Üye: Dr. Öğr. Üyesi Zerrin IŞIK

İmza



EGE ÜNİVERSİTESİ FEN BİLİMLERİ ENSTİTÜSÜ**ETİK KURALLARA UYGUNLUK BEYANI**

EÜ Lisansüstü Eğitim ve Öğretim Yönetmeliğinin ilgili hükümleri uyarınca Yüksek Lisans Tezi / Doktora Tezi olarak sunduğum “BİYOMEDİKAL BİLGİ ÇİZGELERİ İÇİN MAKİNE ÖĞRENMESİ TABANLI ANLAMSAZ BAĞ TAHMİNİ” başlıklı bu tezin kendi çalışmam olduğunu, sunduğum tüm sonuç, doküman, bilgi ve belgeleri bizzat ve bu tez çalışması kapsamında elde ettiğimi, bu tez çalışmasıyla elde edilmeyen bütün bilgi ve yorumlara atıf yaptığımı ve bunları kaynaklar listesinde usulüne uygun olarak verdiğimi, tez çalışması ve yazımı sırasında patent ve telif haklarını ihlal edici bir davranışımın olmadığını, bu tezin herhangi bir bölümünü bu üniversite veya diğer bir üniversitede başka bir tez çalışması içinde sunmadığımı, bu tezin planlanmasından yazımına kadar bütün safhalarda bilimsel etik kurallarına uygun olarak davrandığımı ve aksinin ortaya çıkması durumunda her türlü yasal sonucu kabul edeceğimi beyan ederim.

27 / 07 / 2018

İmzası:



Adı-Soyadı: Remzi ÇELEBİ

ÖZET

BİYOMEDİKAL BİLGİ ÇİZGELERİ İÇİN MAKİNE ÖĞRENMESİ TABANLI ANLAMSAK BAĞ TAHMINİ

ÇELEBİ, Remzi

Doktora Tezi, Bilgisayar Mühendisliği Anabilim Dalı

Tez Danışmanı: Dr. Öğr. Üyesi Özgür GÜMÜŞ

13.07.2018, 86 sayfa

Son zamanlarda, gerçek dünyadaki varlıklar ve bunların birbiriyle ilişkileri hakkında milyonlarca gerçekler (ifadeler) içeren büyük bilgi tabanlarının oluşturulması oldukça ilgi görmektedir. Günümüzde biyomedikal alandaki bir çok verinin Anlamsal Web teknolojileri ile erişebilir olması sayesinde bu bilgi tabanları Bağlı Veri formunda olan bilgi çizgeleri olarak sunulmaktadır. Bilgi çizgeleri, veriyi tanımlamada güçlü bir model sunmakta ve aynı zamanda altında yatan çizge yapısı sayesinde çizge madenciliği algoritmalarının uygulanmasını mümkün kılmaktadır. Bu tezde büyük bilgi çizgelerinde eksik bağlantıları keşfetmek ve yeni bağlantıları tahmin etmek için çeşitli yaklaşımlar sunulmuştur. Bu tez kapsamında biyolojik ve biyomedikal bilgi ağlarında varlıklar arasında yeni bağları keşfetmek için bilgi çizgeleri kullanarak makine öğrenmesi temelli melez bir yaklaşım önerilmiştir. Yeni ilişkilerin tahmini için çizgenin yapısal ve anlamsal özelliklerine dayanan iki öznitelik grubu, yerel ve global öznitelikler, kullanılmıştır. Yerel öznitelikler, ağ yakınlıklarına ve global öznitelikler ise anlamsal çizgenin vektör temsiline dayanır. Bu iki öznitelik grubu ile eğitilen makine öğrenmesi modelleri, ayrı ayrı ve bütünleştirilerek bağ tahmini için kullanılmıştır. Ayrıca bağ tahmini yöntemlerinin değerlendirilmesinde göz ardı edilen durumlar için test senaryoları geliştirilmiş ve bu test senaryoları için önerilen yöntemin başarısı denenmiştir. Önerilen yaklaşımların yararlılığı biyomedikal alanda ilaç keşfi ve halk sağlığı için önemli olan iki problem, yeni ilaç-ilaç etkileşimi tahmini ve yeni ilaç endikasyonu tahmini, için başarılı bir şekilde uygulanarak gösterilmiştir. Yöntemin mevcut yaklaşımlara üstünlüğü böylece kanıtlanmıştır.

Anahtar Sözcükler: Anlamsal Web, Bağ Tahmini, Biyomedikal Ağlar, Bağlı Açık Veri, Bilgi Çizgeleri, Vektör tabanlı Çizge Gömmesi, Makine Öğrenmesi.

ABSTRACT**MACHINE LEARNING BASED SEMANTIC
LINK PREDICTION FOR BIOMEDICAL KNOWLEDGE
GRAPHS**

ÇELEBİ, Remzi

PhD. in Computer Engineering

Supervisor: Asst. Prof. Üyesi Özgür GÜMÜŞ

13.07.2018, 86 pages

The construction of large knowledge bases which contain large volumes of data about real world objects and their relationships has been an object of great interest in recent times. Nowadays, many databases in the biomedical field are accessible through Semantic Web technologies, so these databases have been presented as knowledge graphs in the form of Linked Data. The knowledge graphs are powerful models for defining data which also enable the application of graph mining algorithms thanks to the underlying graph structure. A number of approaches for the discovery of missing links and prediction of new links in large knowledge graphs is presented within this thesis. A hybrid approach based on machine learning which uses knowledge graphs to discover new links between entities in biological and biomedical information network is proposed. Two groups of features, local and global features, based on structural and semantic properties of the knowledge graphs, are used for the link prediction. Local features are based on network proximity and global features are based on vector representation of the semantic graph. The machine learning models trained with these two feature groups were evaluated separately and jointly. In addition, test scenarios were developed for cases that were often ignored in evaluating link prediction methods, and the proposed methods applied to these scenarios were tested. The utility of the proposed approaches was demonstrated by successfully applying them on two important problems for the drug discovery process in the biomedical field; predicting new drug-drug interactions and predicting novel drug indications. The method's superiority over existing approaches were thereby demonstrated.

Keywords: Semantic Web, Link Prediction, Biomedical Networks, Machine Learning, Linked Open Data, Knowledge Graphs, Vector based Graph Embedding.



TEŞEKKÜR

Doktora sürecinde her zaman desteğini yanımda hissettiğim danışmanım Dr. Öğr. Üyesi Özgür Gümüş başta olmak üzere; tez çalışmasının şekillenmesinde değerli görüş ve önerileri ile katkıda bulunan jüri üyesi hocalarım Prof. Dr. Oğuz Dikenelli ve Doç. Dr. Yeşim Aydın Son'a en içten teşekkürlerimi sunarım.

Doktora sürecinde beni destekleyen ve yalnız bırakmayan tüm arkadaşlarıma özellikle de Araş. Gör. Erkan Yaşar'a teşekkürü bir borç bilirim.

Hayatım boyunca karşılaştığım her zorlukta beni destekleyen, bugünlere gelmemde emeği geçen aileme çok teşekkür ederim.

Son olarak, 2214-A-Yurt Dışı Doktora Sırası Araştırma Burs Programı kapsamında beni destekleyen Türkiye Bilimsel ve Teknik Araştırma Kurumu'na (TÜBİTAK) teşekkür ederim.

İÇİNDEKİLER

	<u>Sayfa</u>
ÖZET	vii
ABSTRACT	ix
TEŞEKKÜR	xi
ŞEKİLLER DİZİNİ	xxiii
ÇİZELGELER DİZİNİ	xxv
SİMGELER VE KISALTMALAR DİZİNİ	xxvii
1 GİRİŞ	1
1.1 Motivasyon ve Problem Tanımı	1
1.2 Yaklaşım ve Önerilen Çözüm	4
1.3 Çalışmanın Literatüre Katkıları	4
1.4 Tezin Yapısı	5
2 TEMEL BİLGİLER	6
2.1 Anlamsal Web	6
2.2 Anlamsal Web Teknolojileri ve Standartları	7
2.2.1 RDF	7
2.2.2 RDFS	8
2.2.3 OWL	9
2.2.4 SPARQL	11



İÇİNDEKİLER (Devam)

	<u>Sayfa</u>
2.3 Bağlı Veri	12
2.4 Anlamsal Veri Bütünleştirme.	14
2.5 Bilgi Çizgesi	15
2.6 Biyomedikal Ağlarda Anlamsal Çizge Madenciliği	19
2.6.1 Anlamsal Derecelendirme	19
2.6.2 Sık Tekrarlanan Alt-Çizge Bulma	21
2.6.3 Çizge Kümelenmesi / Topluluk Tespiti	22
2.6.4 Anlamsal İlişki Keşfi	23
2.7 Makine Öğrenmesi.	24
3 ANLAMSAL İLİŞKİ KEŞFİ YÖNTEMLERİ	26
3.1 Sosyal Ağlarda Bağ Tahmini	26
3.1.1 Komşuluk Tabanlı Bağ Tahmini Yöntemleri	27
3.1.2 Patika Tabanlı Bağ Tahmini Yöntemleri	27
3.2 Çoklu-İlişkisel Bağ Tahmini	30
3.2.1 Mantıksal Çıkarım Tabanlı Bağ Tahmini Yöntemleri	30
3.2.2 Vektör Gömme Tabanlı Bağ Tahmini Yöntemleri	30
4 LİTERATÜR ÖZETİ.	35
4.1 Yönteme İlişkin Literatür Özeti.	35



İÇİNDEKİLER (Devam)

	<u>Sayfa</u>
4.2 İlaç-İlaç Etkileşimi Tahmini Durum Çalışmasına İlişkin Literatür Özeti . . .	37
4.3 İlaç Endikasyon Tahmini Durum Çalışmasına İlişkin Literatür Özeti	40
5 ÖNERİLEN YÖNTEM	42
6 ÖNERİLEN YÖNTEMİN KULLANIMINA YÖNELİK DURUM ÇALIŞMALARI	47
6.1 İlaç-İlaç etkileşimlerinin tahmini	47
6.1.1 Veri seti	49
6.1.2 Bilgi Çizgesi Oluşturma.	49
6.1.3 Öznitelik Çıkarma	49
6.1.4 Tahmin ve Değerlendirme	50
6.1.5 Deneysel Sonuçlar ve Tartışma	52
6.2 İlaç Endikasyon Tahmini	56
6.2.1 Dayanak Yöntem	57
6.2.2 Eğitim ve Doğrulama.	62
6.2.3 Deneysel Sonuçlar ve Tartışma	63
7 SONUÇLAR	67
7.1 Özet ve Katkıları	67
7.2 İleriki Çalışmalar	69



İÇİNDEKİLER (Devam)

	<u>Sayfa</u>
KAYNAKLAR DİZİNİ	72
ÖZGEÇMİŞ	83
TÜRKÇE-İNGİLİZCE TERİMLER SÖZLÜĞÜ	84





ŞEKİLLER DİZİNİ

<u>Şekil</u>	<u>Sayfa</u>
1.1 Heterojen biyolojik bilgi ağlarındaki etkileşimler ve ilişkiler	3
2.1 Anlamsal web mimarisi	7
2.2 Bir RDF üçlü tanımı ve örneği gösterimi.	8
2.3 RDFS ile tanımlanmış bir RDF çizgesi.	10
2.4 Örnek sorunun cevabı için yazılan alt SPARQL sorguları.	12
2.5 Örnek sorunun cevabı için yazılan birleşik SPARQL sorgusu.	13
2.6 Verilen örnek için RDF çizge gösterimi	18
2.7 Hastalık ile ilişkili yollar vasıtasıyla hastalık ile ilişkili genlerin öncelendirmesi	20
2.8 Sık görülen anlamsal alt-çizge örneği	22
3.1 RDF2Vec yönteminin adımları	33
5.1 Önerilen yöntem genel bakış	43
5.2 Homojen ilaç-ilac etkileşim ağı	45
5.3 İki parçalı ilaç-hastalık etkileşim ağı	46
6.1 İlaç çiftlerinin test sınıflarına göre eğitim ve test kümelerine ayrılması	51
6.2 Geleneksel çapraz doğrulama için öznelik veri kaynaklarının F puan karşılaştırılması	53
6.3 Modellerin test sınıflarına göre karşılaştırılması (DB + KEGG + PHARMGKB birleşimi, bilgi çizgesi olarak kullanılmıştır.)	54



ŞEKİLLER DİZİNİ (Devam)

Şekil	Sayfa
6.4 Modellerin öznitelik gruplarına göre karşılaştırılması (DB + KEGG + PHARMGKB birleşimi, bilgi çizgesi olarak kullanılmıştır.)	55
6.5 Bio2RDF alt kümesinde Clozapine için bilgi ağının görselleştirilmesi	59
6.6 İlaç kimyasal bilgilerini sorgulamak için yazılan SPARQL sorgusu .	60
6.7 İlaç hedef bilgileri sorgulamak için yazılan SPARQL sorgusu	61
6.8 İlaç yan etkileri bilgilerini sorgulamak için yazılan SPARQL sorgusu	61
6.9 Hastalık tanımlarını sorgulamak için yazılan SPARQL sorgusu . . .	62
6.10 Eğitim ve test kümesi ayrımının sentetik veri ile gösterimi	63
6.11 Özniteliklerin önem dereceleri	64
6.12 Dayanak Yöntem için farklı çapraz doğrulama düzenekleri için elde edilen AUC ve F-puan değerleri	65
7.1 Bireye Özgü Tedaviye Genel Bakış	69
7.2 Bağ tahmini yöntemlerinin bireye özgü ilaç önerilmesinde kullanılmasına yönelik bir örnek ağ	70



ÇİZELGELER DİZİNİ

<u>Çizelge</u>	<u>Sayfa</u>
2.1 SQL ve SPARQL sorgulama dillerinin karşılaştırılması	11
2.2 R2R dili ile yazılmış bir kural örneği	15
2.3 Silk dili ile yazılmış bir kural örneği	16
3.1 RDFS ve OWL-DL temel yapıları ve mantık yapıları	31
6.1 Gömme temsilini öğrenmek için kullanılacak her bir bilgi kaynağına ilişkin üçlü, düğüm ve bağ tipi sayısı	49
6.2 Drugbank Bilgi Çizgesi kullanılarak Drugbank 5.0 veri seti üzerinde farklı gömme yöntemlerinin tahmin performans değerleri .	52
6.3 Birleştirilmiş (DB + KEGG + PHARMGKB) Bilgi Çizgesi kullanılarak Drugbank 5.0 veri seti üzerinde farklı gömme yöntemlerinin tahmin performans değerleri	53
6.4 Farklı çapraz doğrulama düzenekleri kullanılarak birleştirilmiş RDF bilgi çizgesinin performansı	54
6.5 Ağ yakınlık ölçütleri kullanılarak elde edilen tahmin sonuçları . . .	55
6.6 10-kat çapraz doğrulama kullanarak yöntemlerin tahmin sonuçları .	64
6.7 Rastgele Orman sınıflandırıcısının Reboksetin ilacı için ürettiği ilk 10 endikasyon tahmini	66



SİMGELER VE KISALTMALAR DİZİNİ

Kısaltma	Açılım
ADR	Adverse Drug Reaction (İstenmeyen İlaç Etkileşimi)
DDI	Drug-Drug Interaction (İlaç-İlaç etkileşimi)
DL	Description Logic (Betim Mantık)
KB	Knowledge Base (Bilgi Tabanı)
KG	Knowledge Graph (Bilgi Çizgesi)
KGE	Knowledge Graph Embedding (Bilgi Çizge Gömmesi)
LOD	Linked Open Data (Bağlı Açık Veri)
LR	Logistic Regression (Lojistik Regresyon)
NB	Naive Bayes (Naive Bayes)
OWL	Web Ontology Language (Web Ontoloji Dili)
RDF	Resource Description Framework (Kaynak Betim Çerçevesi)
RDFS	RDF Schema (RDF Şeması)
SPARQL	SPARQL Protocol and RDF Query Language (SPARQL Protokolü ve RDF Sorgu Dili)
URI	Uniform Resource Identifier (Tek Biçimli Kaynak Tanımlayıcı)
URL	Uniform Resource Locator (Tek Biçimli Kaynak Konumlayıcı)



1 GİRİŞ

Bu bölümde tezin motivasyonu ve problem tanımı yapılarak tezin ele aldığı araştırma konuları, önerilen çözüm ve çalışmanın literatüre katkıları sunulacaktır.

1.1 Motivasyon ve Problem Tanımı

Bilgi tabanları (Knowledge Bases -KBs) kişi, eşya, yer ve bunun gibi bir çok varlık ve ilişkiler hakkında gerçekler içerir. Bilgi tabanları günümüzde bir çok görev için kullanılmaktadır. Örneğin arama motorları bilgi tabanı kullanarak arama sonuçlarını iyileştirebilir, soru-cevap sistemleri bir soruyu bilgi tabanı üzerinde çalışacak bir sorguya çevirebilir, bilgi çıkarım sistemleri model eğitimi için bilgi tabanlarını kullanabilir . Bilgi tabanlarındaki veriler çizge olarak tanımlanırsa bu bilgi tabanlarına bilgi çizgeleri (knowledge graph) denir. Bilgi çizgeleri, veriyi tanımlamada güçlü bir model sunar ve aynı zamanda altında yatan çizge yapısı nedeniyle çizge madenciliği algoritmalarının uygulanmasını mümkün kılar (Chen et al., 2009).

Anlamsal Web (Semantic Web) (Berners-Lee et al., 2001), web üzerinde makinelerin anlayabileceği ve işleyebileceği tutarlı, mantıklı ve anlamlı bir çizge veri modeli sunarak birleşik bir veri ağı oluşturmayı hedeflemiştir. Web'in Anlamsal Web'e doğru evrilmesiyle beraber, Web'teki bir çok veri, bilgi çizgeleri formunda sunulmaya başlanmıştır. Bilgi çizgeleri, anlamsal çizge veya ağ olarak da adlandırılır. Son zamanlarda birçok biyolojik ve biyomedikal veri tabanı Anlamsal Web teknolojileri üzerine inşa edilmekte ve verinin sorgulanması, entegrasyonu ve federasyonu için bilgi çizgelerini kullanmaktadır (Katayama et al., 2014) . Bu veri tabanları Anlamsal Web altyapısına dayalı bir çizge veri modeli olan Bağlı Veri (Linked Data) (veya herkese açık olması halinde Bağlı Açık Veri (Linked Open Data - LOD)) ilkelerini kullanarak verilerini kullanıcılara sunmaktadırlar (Smith et al., 2005; Jupp et al., 2014). Bağlı Veri, URI (Uniform Resource Identifier- Tek Biçimli Kaynak Tanımlayıcısı)'ler ile tanımlanan <özne, yüklem, nesne> üçlülerinden oluşan bilginin Kaynak Betim Çerçevesi (Resource Description Framework - RDF) kullanarak tanımlanmasını önerir. Yüklem, özne ve nesneyi birbirine bağlayarak RDF üçlülerini oluşturur ve bu üçlülerin birleşimi doğal bir bilgi çizgesi meydana getirir.

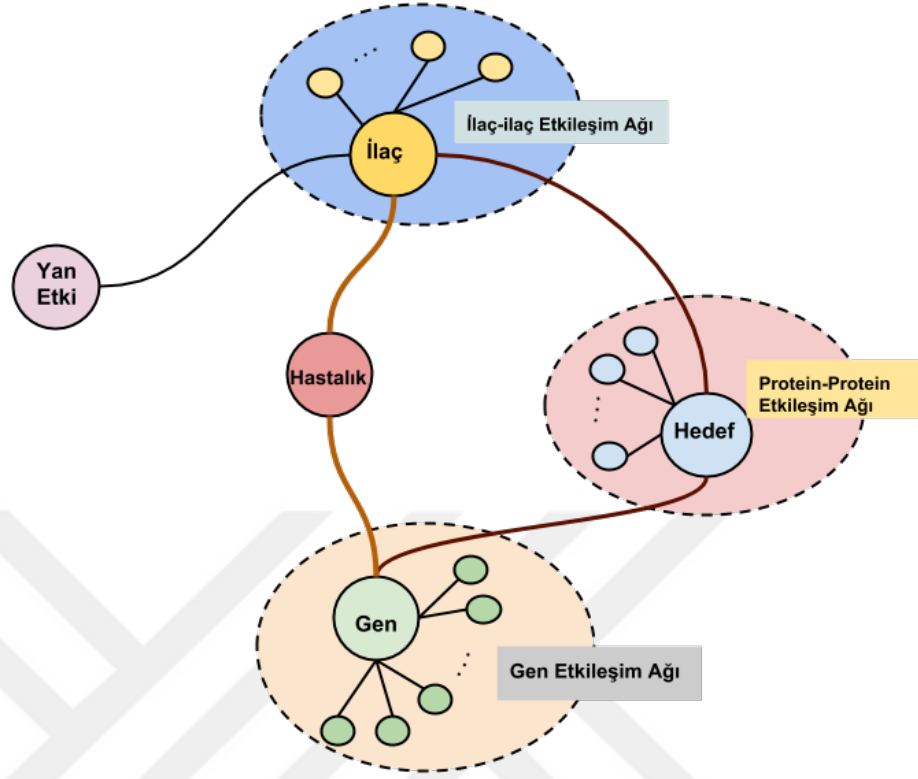
Bilgi çizgeleri çok kullanışlı olmakla beraber, en büyük kapsamlı bilgi çizgeleri bile tam değildir ve eksik bilgiler içerir. Bu eksik bilgiler biyolojik bilgi çizgeleri için çok daha yaygındır ve çalışmalar genellikle varlıklar arasındaki eksik

ilişkilerin bulunmasına yoğunlaşmıştır. Protein etkileşim ağları, besin etkileşim ağları, ilaç etkileşim ağları ve gen ifadesi ağları gibi biyolojik bilgi ağları içindeki eksik ilişkiler genellikle maliyeti yüksek laboratuvar veya klinik deneyler sonucu bulunmaktadır. Örneğin ilaç-ilaç etkileşimleri, Faz-4 klinik çalışmaları veya ilaç piyasaya sunulduktan sonraki ilaç yan etki bildirimi raporları sonucu bulunduğundan çok sayıda etkileşim bilgisi eksiktir ya da kesin değildir, bu da maliyeti ve beklenmedik ilaç etkileşim riskini artırır (Percha and Altman, 2013).

Bilgi çizgesi tamamlama (knowledge graph completion), var olan bilgi çizgesinden yararlanarak varlıklar arasındaki bağı tahmin etmeyi amaçlar. Bilgi çizgesi tamamlama, sosyal ağ analizindeki bağ tahminine benzer, ancak aşağıdaki nedenlerden dolayı daha zordur: 1) bilgi çizgelerindeki düğümler farklı tip ve niteliklere sahip varlıklardır; 2) bilgi çizgelerindeki kenarlar ise farklı tiplere sahip ilişkilidir. Sosyal ağ analizinde kullanılan geleneksel bağ tahmini yöntemleri literatürde homojen çizgeler (tek tip düğüm ve bağa sahip) için uygulanmıştır (Lü et al., 2011) . Bilgi çizgesi, çok tipli düğüm ve bağa sahip heterojen bir yapıda olduğundan geleneksel bağ tahmin yaklaşımları, bilgi çizgeleri üzerinde bağ tahmin için yetersizdir (Lin et al., 2017) . Bu tür heterojen biyolojik bilgi çizgelerinde bağ tahmini uygulanarak hastalıklar, genler, ilaçlar ve klinik bulgular arasındaki bilinmeyen veya potansiyel bağlantıları keşfedebilir. Bu çalışmalar, biyolojik bilgiyi ve klinik standartları iyileştirebilir ve dolayısıyla da olarak sağlık hizmetlerinin kalitesini yükseltebilir.

Şu an biyolojik çalışmalar çoğunlukla, ilaç, gen, hastalık arasındaki etkileşimleri tek başlarına ayrı ayrı incelenmektedirler; araştırmacılar genellikle ilaç ve gen etkileşimlerinin tahmini için ilaç ve gen etkileşim ağlarını veya ilaç ve hedef arasındaki ilişki tahmini için ilaç-hedef etkileşimlerini kullanırlar (Şekil 1.1'e bakınız). Bu ağların farklı veriler eklenerek zenginleştirilmesi bilinmeyen etkileşimlerin tahmini için bütüncül bir bakış açısı kazandırır. Örneğin ilaç-ilaç etkileşimi için ilaç hedef bilgileri, ilaç yan etkileri ve hastalık (endikasyon) bilgileri dahil ederek bilginin bütüncül olarak değerlendirilmesini sağlar ve ilaç-ilaç etkileşim tahminlerini iyileştirir.

Bilgi tamamlama görevi için yaklaşımlar iki ana kategoriye ayrılabilir: 1) eksik bilgileri tahmin etmek için mantıksal çıkarımlar kullanan yöntemler, 2) bilgi çizgesindeki varlıkların ve ilişkilerin bir vektör uzay modelini bulan yöntemler. Örneğin, mantıksal çıkarım kullanan yöntemler, Muslera ile Selçuk İnan'ın takım arkadaşı olduğu, Muslera'nın Galatasaray için oynadığı bilgisi biliniyorsa Selçuk İnan'ın da Galatasaray için oynadığı çıkarımını yapar. Buna karşılık bir vektör uzay metodu, Muslera ve Selçuk İnan için vektörlerin birbirine yakın olduğunu ve



Şekil 1.1: Heterojen biyolojik bilgi ağlarındaki etkileşimler ve ilişkiler

oynadıkları takımların birbirine vektörel olarak yakın olduğunu bulabilir. Mantıksal çıkarım yöntemleri genellikle daha kesindir ve uzay vektör modelinden daha güçlü çıkarımlar yaparlar fakat ölçeklenebilir değildirler. Bilgi çizge gömme yöntemleri, bilgi çizgesini çizge yapısını bozmadan, belli özelliklerini koruyarak vektör uzayına yerleştirir. RDF2Vec (Ristoski and Paulheim, 2016) ve TransE (Bordes et al., 2013) gibi gömme yöntemleri, bilgi çizgelerinin etkin ve ölçeklenebilir şekilde gömme temsillerini oluşturabildiklerini ve bu temsilleri kullanarak farklı makine öğrenmesi görevleri için başarılı olabildiklerini göstermişlerdir. Bunlardan RDF2Vec, Word2Vec'in dil modelleme yaklaşımını RDF Çizge Gömme'ye uyarlayan son zamanlarda ortaya çıkmış bir yöntemdir. Word2Vec (Mikolov et al., 2013), kelime yerleştirme olarak adlandırılan kelimelerin vektör gösterimini öğrenmek için bir sinir ağı modelini eğitir. Her sözcüğü, anlamsal ve söz dizimsel olarak daha yakın sözcüklerin vektör uzayında daha yakın olacağı, bir vektör uzayına yerleştirir. Bu yaklaşımın altında yatan hipotez, metin içinde sıralı olarak birbirine yakın kelimeler, istatistiksel olarak da daha fazla bağımlı olacaklardır. RDF2Vec, çizgedeki varlıklar ve varlıklar arasındaki ilişkilerin vektör temsilini öğrenmek için çizgede rastgele yürüyüşler yaparak bir yapay sinir ağı modeline girdi olacak yürüyüşler veya yollar dizisine dönüştürür.

1.2 Yaklaşım ve Önerilen Çözüm

Bu tezde büyük bilgi tabanları üzerinde çıkarım yapmak için yaklaşımlar sunulmuştur. Sunduğumuz yaklaşımlar bilgi tabanını bir çizge olarak görmekte/ele almakta ve bu çizgenin karakteristiğini elde etmek için öznitelikler oluşturmaktadır. Ve bu öznitelikler makine öğrenmesi modellerini eğitmek için kullanılmaktadır. Bu tez kapsamında biyolojik ve biyomedikal bilgi ağlarında varlıklar arasında yeni ilişkileri keşfetmek için bilgi çizgeleri kullanarak makine öğrenmesi temelli melez bir yaklaşım önerilmiştir. Yeni ilişkilerin tahmini için çizgenin yapısal özelliklerine dayanarak yerel ve global öznitelikler kullanılmıştır. Bu amaçla ağ yakınlıklarına ve çizgenin vektör temsiline dayanan yaklaşımlar bir arada kullanılarak, varlıklar arasında bağ tahmini yapılmıştır.

1.3 Çalışmanın Literatüre Katkıları

- Farklı şematik yapıda ve (aynı varlık için) farklı isimlendirmeye sahip biyomedikal verileri Anlamsal Web teknolojiler kullanılarak bütünleştirilmiş ve meydana gelen heterojen bilgi ağını RDF çizge veri modeli ile ifade ederek bu bütünleşik bilgi ağı için bir bağ tahmin yöntemi sunulmuştur.
- Bağlı veri için bağ tahmini yöntemleri incelenerek, global ve yerel öznitelikleri içeren melez bir yaklaşım önerilmiştir. Yerel öznitelikler, sosyal ağ bağlantı analizinde kullanılan yakınlık ölçümlerini içerir. Global öznitelikler ise bir çizgenin düğüm ve bağlantılarını çizge yapısını koruyarak bir vektör uzayına gömen çizge gömme yöntemleri kullanılarak çıkarılmıştır.
- Biyomedikal veriler üzerinde iki farklı araştırma durumu (ilaç-ilaç, ilaç-hastalık ilişki tahmini) ele alınmış, önerilen yaklaşım, var olan son teknolojik yaklaşımlar ile karşılaştırılmıştır.
- Potansiyel ilaç-ilaç etkileşimlerini tahmin etmek için Açık Veri (LOD) kullanılarak öznitelik vektörü çıkarmak amacıyla gömme yöntemleri uygulanmıştır. Farklı bağlı veri kaynaklarının ve gömme yöntemlerinin, ilaç-ilaç etkileşimi tahminine etkisi araştırılmıştır.
- İlaç ve hastalığı karakterize eden öznitelik vektörleri çıkarılmış ve yeni ilaç-ilaç etkileşimi ve ilaç-hastalık ilişkisi tahmini gibi makine öğrenmesi görevleri için nasıl kullanılabileceği gösterilmiştir.
- İlaç-ilaç ve ilaç-hastalık gibi çiftleri girdi olarak alan yöntemlerin (çift girdili- paired-input) tahmin performansını doğru ve gerçekçi bir şekilde değerlendirmek için değerlendirme düzenekleri sunulmuştur.

1.4 Tezin Yapısı

Tezin yapısı ve her bölümde içeriği aşağıdaki gibi düzenlenmiştir:

- Bölüm 2’de, Anlamsal Ağ ve Bağlı Veri ve bu teknolojiler hakkında ön bilgi verilmiş ve ayrıca bu teknolojileri veri bütünleştirme bağlamında nasıl kullanıldığından bahsedilmiştir. Ayrıca biyomedikal ağlarda anlamsal çizge madenciliğinde kullanılan yöntemler incelenmiş ve kullanım durumlarına örnekler verilerek genel bir resim ortaya konmuştur.
- Bölüm 3’te, anlamsal ilişki keşfi için sosyal ağlarda ve anlamsal ağlardaki var olan son teknoloji yöntemler incelenmiştir.
- Bölüm 4’te literatürde yöntemimize benzeyen diğer çalışmalar özetlenerek sunulmuştur. Bunun yanısıra durum çalışmalarına ilişkin ilgili çalışmalar da irdelenmiştir.
- Bölüm 5’te tez kapsamında önerilen makine öğrenmesi tabanlı anlamsal bağ tahmini yöntemi sunulmuştur.
- Bölüm 6’da, önerilen yöntem bağlamında kullanım durumları, yöntemin elde ettiği sonuçlar ve diğer yöntemler ile karşılaştırması verilmiştir. Her durum için sonuçların iyimser bir şekilde yorumlanmasını önlemek için değerlendirme düzenekleri detaylı anlatılmış ve buna göre sonuçlar tartışılmıştır.
- Son olarak Bölüm 7’de, yapılan çalışma özetlemiş ve kilit sonuçlar/çıktılar maddeler halinde verilmiştir. Ayrıca ileriki çalışmalar için bu yöntemin kişiye özgü sağlık kapsamında kullanımı hakkında önerilerde bulunulmuştur.

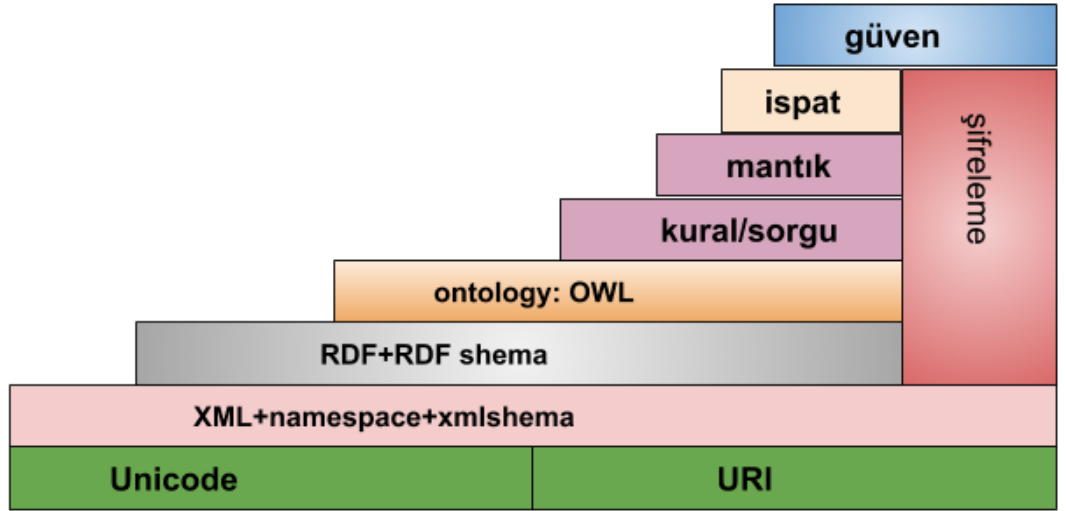
2 TEMEL BİLGİLER

2.1 Anlamsal Web

Web teknolojisi resim, ses ve metinden oluşan dokümanları tanımlayan ve bu dokümanların diğer kullanıcılar ile paylaşımını ve sunumunu sağlayan bir teknolojiler bütünüdür. Web'te yer alan içeriğin sözdizimsel ve anlamsal olarak heterojen yapıda olması ve bu içeriklerin doğal dille yazılmış yapısal olmayan verilerden oluşması, bunların bilgisayar tarafından işlenmesini ve anlaşılmasını zorlaştırmaktadır.

Web, iki temel fikir üzerinde kurulmuştur; dokümanları standart bir format, HTML, ile web sayfası olarak sunmak ve bu dokümanlar arasındaki geçişler (hyperlink) için bağlantılar vermek. Anlamsal Web (Berners-Lee et al., 2001), şuan ki Web'in genişletilmiş bir hali olarak önerilen bir teknoloji olup, Web'in doküman paylaşım yaklaşımını verilere de uyarlayarak Web'te yer alan bilginin birbirine bağlı açık bir devasa veri ağı olarak/şeklinde sunumunu hedeflemiştir (Ding et al., 2012). Aynı Web'te olduğu gibi Anlamsal Web'te de bağlı veri ağı organik bir biçimde, hiç bir merkezi otorite olmadan, yeni veriler var olan verilere bağlanarak büyümektedir. Anlamsal Web'in günlük hayatta kullanımına bir örnek olarak arama motorları istenilen bilgiyi, anahtar kelimeleri taramak yerine kişilerin sorularını cevaplayarak bulabileceklerdir. Örneğin bir kullanıcı 2006 model 1.4 motor ikinci el dizel bir otomobili ararken bir çok web sitesinden bilgiler bulunur, birleştirilir ve sonuç listelenir.

Anlamsal Web, yeni nesil ağ, mevcut internetin uzantısı olarak kabul edilir ve heterojen kaynaklardan veri entegrasyonu için bir çerçeve sağlar. Anlamsal Web, makinelerin internet üzerinde bilginin bulunması, birleştirilmesi ve çıkarılması gibi emek yoğun işleri yapmasını sağlar. Açık bağlı veri internet üzerinde açık halde olan verinin, yayınlanması, entegre edilmesi ve analizi için anlamsal ağ teknolojisini kullanan yeni bir yaklaşımdır. Açık bağlı verinin en temel standardı olan RDF, farklı kaynaklardan alınan verilerin bütünleştirilme ve bağlanma zorluklarını azaltmak için birçok doğal özelliğe sahiptir. Örneğin, açık URI adreslerinin kullanımı farklı alanlardaki varlıkların ve ilişkilerin birleştirilmesi için formal ve standart bir mekanizma sunar. Ayrıca RDF üçlü modeli, farklı alanlar arasındaki heterojenliği çözmede anahtar bir rol oynayan anlamsal ontoloji dilleri ile modellenebilen ortak ve formal sözlükler için de uygundur (Chen et al., 2009).



Şekil 2.1: Anlamsal web mimarisi

2.2 Anlamsal Web Teknolojileri ve Standartları

Anlamsal Web'in temelleri ağda yer alan bilginin kendisinin ve üst-verisinin açık olarak web standartları ile yayınlanması, bağlanabilirliği ve erişimi üzerine kurulmuştur. Bu temellerin tümü, WWW Konsorsiyumu tarafından geliştirilen standartlar ile desteklenmektedir. Anlamsal Web'in temelini oluşturan standartlar, Şekil 2.1'de görüldüğü üzere en alt basamakta yer alan URI ve Unicode, onun bir üst basamağı XML, RDF (Manola et al., 2004), RDFS (McBride, 2004), OWL (Horrocks et al., ; McGuinness et al., 2004) ve SPARQL (Hancock, 2004; Harris et al., 2013) olarak sayılabilir ¹.

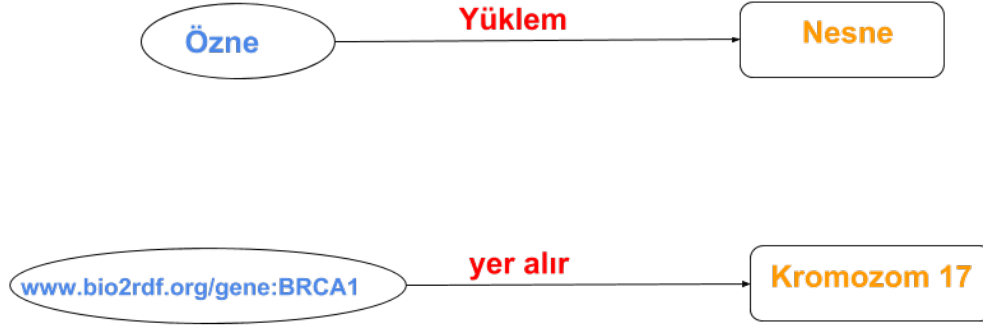
Bu standartlardan en önemli olanları sonraki alt bölümlerde detaylıca açıklanmıştır.

2.2.1 RDF

RDF, kaynakları ve arasındaki ilişkileri tanımlamak için önerilen bir kaynak tanım standardıdır. RDF, gerçekleri veya ifadeleri <özne, yüklem, nesne> üçlüleri ile ifade eder (Şekil 2.2). Bu üçlüler kendi başına bir anlam taşıdıkları gibi başka üçlüler ile birleşerek daha geniş bir anlam ortaya koyabilir. RDF, XML veya JSON gibi birçok formatta gösterilebilir.

Bir RDF ifadesinde özne URI ile tanımlanmış bir kaynağı gösterir. Yüklem ise hem bir kaynak hem de özneye ait bir özelliği tanımlar. Son olarak nesne

¹Tim Berners Lee'nin Semantic Web konuşması: <https://www.w3.org/2000/Talks/1206-xml2k-tbl/>



Şekil 2.2: Bir RDF üçlü tanımı ve örneği gösterimi.

ise bir kaynak veya bir özelliği gösteren alfanümerik bir değerdir. Günümüzde web sayfalarını tanımlamak için URI kullanılırken, anlamsal web’te ise varlıkları tanımlamak için URI kullanılır. URL bir web kaynağında lokasyonu tanımlarken, URI ise bu web kaynağının kendisini tanımlar. Bunun yanında Çince ve Arapça gibi farklı dil desteklerine sahip bir standart olan IRI (Internationalized Resource Identifier - Uluslararası Biçimli Kaynak Tanımlayıcısı) ise URI’nin yerine geçmeye başlayan bir standarttır.

2.2.2 RDFS

RDFS (RDF Schema - RDF Şema), RDF veri modelinin genişletilmiş halidir ve RDF verisi için bir üst-veri (şema) tanımı yapar. Bunun yanında RDF modeli için kısıtlama ve genelleme tanımlaması sağlar. Ayrıca bir alan hakkında kavramları ve bu kavramlardan türemiş varlıkları ve bunlar arasındaki ilişkileri formal olarak tanımlamayı sağlar. RDFS, RDF kaynaklarının özelliklerini ve sınıflarını tanımlamak için oluşturulmuş bir kelime tanımlama dili olarak tanımlanabilir (Antoniou et al., 2012). RDFS yapıları aşağıdaki gibidir:

- **Sınıflar**

- **rdfs:Resource** her kaynağın sınıfıdır. RDF tarafından tanımlanan her şey bir kaynaktır.
- **rdfs:Class**, bir kaynağı diğer kaynakların bir sınıfı olarak tanımlar.
- **rdfs:Literal**, sayısal ve karakter dizisi gibi değerleri gösterir.

- **rdfs:Datatype**, bir veri tipi tanımlamak için kullanılır. **rdfs:Class**'ın hem bir örneği hem de bir alt sınıfıdır.
- **rdfs:Property**, özelliklerin sınıfını tanımlar.

- **Özellikler**

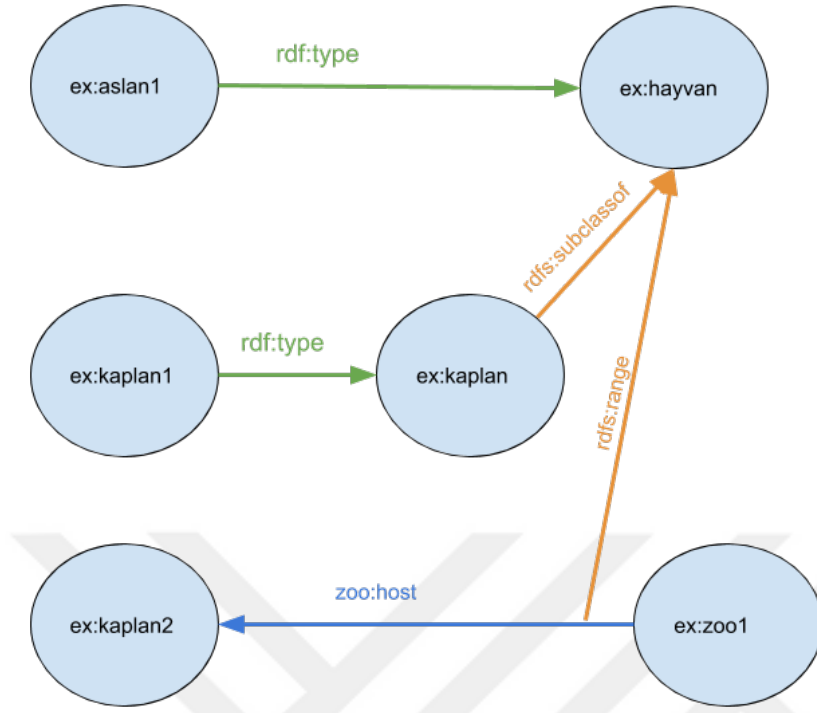
Özellikler **rdfs:Property** sınıfının örnekleridir. Özne kaynakları ile nesne kaynakları arasındaki ilişkiyi açıklar.

- **rdfs: domain**: Öznenin sınıfını belirtir.
- **rdfs:range**: Nesnenin veri tipini veya sınıfını belirtir.
- **rdfs:type**, kaynağın bir sınıfın örneği olduğu göstermek için kullanılır.
- **rdfs:subClassOf**, sınıflar arasında hiyerarşi kurmayı sağlar. Bir sınıfı, diğer bir sınıfın alt sınıfı olarak tanımlar.
- **rdfs:subPropertyOf**, bir özelliği taşıyan bütün kaynakların, başka bir özelliği taşıdığı durumları göstermek için kullanılır.
- **rdfs:label** insan tarafından okunabilir bir kaynak ismi vermek için kullanılır.

RDFS'in varlığı, RDF sözcüklerinin hiyerarşik olarak düzenlenmesi için modelleme yapıları sağlar. RDFS kullanılarak, sözlük bilgisi tanımlanabilir; hangi özelliklerin hangi nesne ve hangi değerler için geçerli olabileceği belirtilebilir; ve nesneler arasındaki ilişkiler açıklanabilir. Bu nedenle, RDFS, Anlamsal Web vizyonuna uygun olarak anlamsal bilgiyi makineye erişebilir kılar. Şekil 2.3'te yer alan RDF çizgesi bu ilişkileri açıklamak için iyi bir örnektir.

2.2.3 OWL

OWL (Web Ontology Language - Web Ontoloji Dili), bir ontoloji tanımlamak için W3C tarafından nesneler, nesne grupları ve bunlar arasındaki ilişkiler hakkında zengin ve karmaşık bilgiyi göstermek için tasarlanan bir ontoloji dilidir. Ontoloji, temel kavramlar ve aralarındaki ilişkileri tanımlayan bir bilgi temsili yaklaşımıdır. Ontoloji, hem insanlar hem de makineler tarafından bir alan bilgisinin ortak bir şekilde anlaşılmasını sağlar. OWL, RDFS gibi üst bilgiyi tanımlayan yapılar içerir, RDFS'ten farklı olarak Betim Mantık (Description Logic) tarafından devralınan özellikleri zenginleştirerek RDFS modelinin anlamsallığını ve ifade edilebilirliğini genişletir. OWL ontolojileri RDF verisi ile birlikte tutarlılık kontrolü, çıkarsama ve akıl yürütme yeteneklerine sahip bir bilgi tabanı oluşturur. OWL üç alt dile sahiptir. Bunlar;



Şekil 2.3: RDFS ile tanımlanmış bir RDF çizgesi.

- **OWL Lite:** öncelikli olarak bir sınıflandırma hiyerarşisine ve basit kısıtlamalara ihtiyaç duyan kullanıcıları desteklemektedir. Örneğin, kardinalite kısıtlamalarını desteklerken, yalnızca 0 veya 1'in asallık değerlerine izin verir.
- **OWL DL:** hesaplama bütünlüğünü korurken (tüm sonuçların hesaplanabilirliği garanti edilir) ve karar verilebilirliği (tüm hesaplamalar sonlu zamanda bitecek) maksimum açıklık isteyen kullanıcıları desteklemektedir. OWL DL tüm OWL dil yapılarını içerir, ancak sadece belirli kısıtlamalar altında kullanılabilirler (örneğin, bir sınıf birçok sınıfın bir alt sınıfı olabilirken, bir sınıf başka bir sınıfın örneği olamaz). OWL DL, OWL'nin resmi temelini oluşturan mantıkları araştıran bir araştırma alanı olan Betimleme Mantığıyla olan benzerliğinden dolayı adlandırılmıştır.
- **OWL Full:** hesaplama güvencesi olmadan maksimum ifade etme yeteneği ve RDF cümle özgürlüğünü isteyen kullanıcılar içindir. Örneğin, OWL Full'de bir sınıf eş zamanlı olarak hem bireylerin bir topluluğu olarak, hem de kendi başına bir birey olarak davranabilir. OWL Full, bir ontolojinin önceden tanımlanmış (RDF veya OWL) kelime hazinesinin anlamını arttırmasına izin verir. Herhangi bir akıl yürütme yazılımının OWL Full'in her özelliği için tam akıl yürütmeyi destekleyebilmesi olası değildir.

Çizelge 2.1: SQL ve SPARQL sorgulama dillerinin karşılaştırılması

SQL	SPARQL
SELECT [ilac_endikasyon], [ilac_yan_etki] FROM [endikasyon_bilgi],[yanetki_bilgi] WHERE endikasyon_bilgi.ilac_id=yanetki_bilgi.ilac_id ilac_ismi="aspirin"	SELECT ?ilac_endikasyon, ?ilac_yanetki WHERE { ?ilac :yanetki ?ilac_yanetki ?ilac :endikasyon ?ilac_endikasyon FILTER regex(str(?ilac), "aspirin") }

2.2.4 SPARQL

SPARQL, RDF formatındaki verinin sorgulanması ve getirilmesi için geliştirilen bir sorgu dilidir. Bir RDF verisinde, verilen bir üçlü ile eşleşen bütün ifadeleri bulmak için aşağıdaki SPARQL sorgusu yazılabilir:

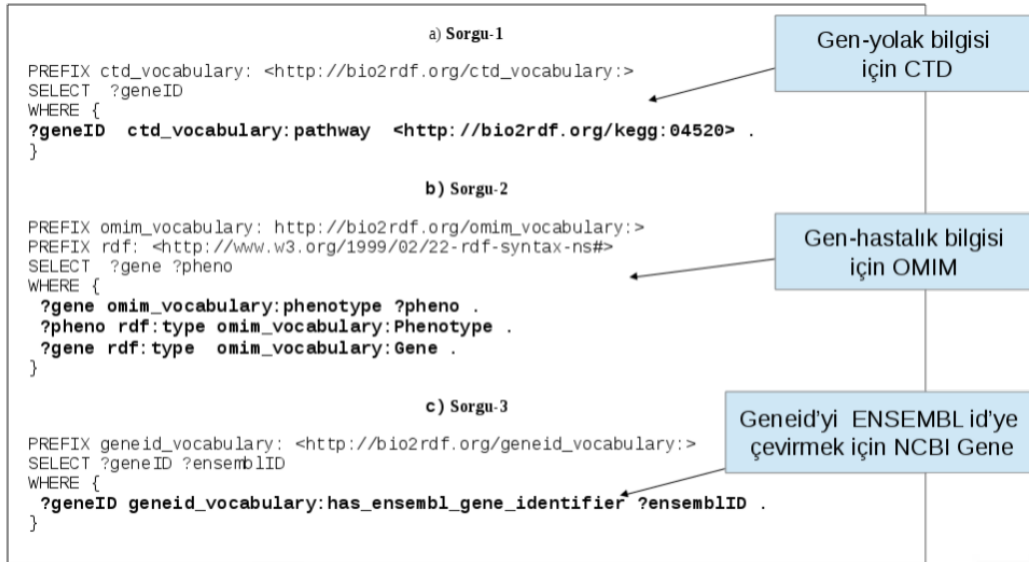
*SELECT * WHERE ?ozne ?yuklem ?nesne .*

SPARQL, SQL diline çok benzemekle beraber, doğal dildeki sorgulama mantığına daha yakın ve esnek bir yapıya sahiptir. Çizelge 2.1’de sorgu dillerini karşılaştırmak için aynı sorgu için SQL ve SPARQL dillerinde sorgunun nasıl yazılabildiği örneklendirilmiştir.

Bir üçlü deposunda (RDF store) tutulan üçlülere SPARQL kullanarak sorgulayabilmek için SPARQL uç noktası (SPARQL endpoint) adı verilen servisler bugün webde oldukça yaygındır. Örneğin şu soru ile ilgilendiğimizi düşünelim: *"Belirli bir yolak için, yolaktaki genlerle ilişkili hastalıklar nelerdir?"*

Bu sorunun cevabını almak için bir dizi veri kaynağını kullanmak gerekir: gen-hastalık ilişkisi için OMIM, gen yolak bilgisi için CTD, gen kimlik numarasının ENSEMBL id’ye dönüştürülmesi için NCBI Gene. Bu veri setleri, Bio2RDF (Callahan et al., 2013) projesinin bir parçası olarak web üzerinde ayrı ayrı Şekil 2.4’teki gibi sorgulanabilir.

Her sorgu, farklı ve uzak uç noktalardan veri toplanmasını gerektirir. Sorgu-1, Sorgu-2 ve Sorgu-3 sorguları birleştirildiğinde, üç uç noktadan herhangi birine karşı çalıştırılırsa herhangi bir cevap üretmeyecektir. Yukarıda tanımlanan



Şekil 2.4: Örnek sorunun cevabı için yazılan alt SPARQL sorguları.

soruyu yanıtlamak için bu veri kümeleri arasındaki bağlantıları dolaşmamız gerekmektedir. Bu nedenle, sorgu her bir sorgunun sonucunun kendi veri kümesinde gerçekleştirilmesi ve daha sonra birleştirilmesi için birleşik (federated) olmalıdır. SPARQL sorguları birleştirimi, sorguların hangi sırada ve hangi veri kümesi üzerinden yürütüleceğini tanımlama işlemidir. SPARQL 1.1 federasyon uzantısı ², bir alt sorgunun gerçekleştirileceği uç noktanın URL'sini belirtmek için kullanılır. Birleşik sorgu, SPARQL uç noktalarını sorgulamak için SPARQL 1.1 tarafından tanımlanan bir SERVICE deyimini kullanır. Yukarıda tanımlanan sorgular, Şekil 2.5'te görüldüğü gibi birleşik halde çalıştırılabilir.

2.3 Bağlı Veri

Makinaların webdeki veriyi daha etkin kullanarak daha akıllı işler yapabilmesi için tasarlanan Anlamsal Ağ'ın gerçekleşmesinin ilk adımı olarak web üzerinde birbirine bağlı veri kümelerinden oluşan bir bilgi tabanı yaratmanın gerekliliği kabul edilmiştir (Heath and Bizer, 2011) . Bu doğrultuda Bağlı Veri İlkeleri yayınlanarak, Anlamsal Web'in dönüşümü için bu aşamadan sonra Bağlı Veri (Linked Data) adı kullanılmaya başlanmıştır. Bu ilkeler (Bizer et al., 2009) tarafından şu şekilde tanımlanmıştır:

1. Varlıkları tanımlamak için URI'ler kullanılması
2. Varlıkları arayıp bulmak için HTTP URI kullanılması

²<https://www.w3.org/TR/sparql11-federated-query/>

```

PREFIX omim_vocabulary: <http://bio2rdf.org/omim_vocabulary:>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX ctd_vocabulary: <http://bio2rdf.org/ctd_vocabulary:>
PREFIX geneid_vocabulary: <http://bio2rdf.org/geneid_vocabulary:>
SELECT ?ensemblID ?pheno
WHERE {
  SERVICE <http://cu.ctd.bio2rdf.org/sparql>
  {
    ?geneID ctd_vocabulary:pathway <http://bio2rdf.org/kegg:04520> .
  }
  SERVICE <http://cu.omim.bio2rdf.org/sparql>
  {
    ?gene omim_vocabulary:xref ?geneID .
    ?gene omim_vocabulary:phenotype ?pheno.
    ?pheno rdf:type omim_vocabulary:Phenotype .
    ?gene rdf:type omim_vocabulary:Gene .
  }
  SERVICE <http://cu.gene.bio2rdf.org/sparql>
  {
    ?geneID geneid_vocabulary:has_ensembl_gene_identifier ?ensemblID
  }
}

```

Şekil 2.5: Örnek sorunun cevabı için yazılan birleşik SPARQL sorgusu.

3. Bir URI istendiğinde, standartları (RDF(S), SPARQL) kullanarak yararlı bilginin sağlanması
4. Diğer URI'lere bağlantı yapılarak daha fazla bilginin keşfedilmesi

Bağlı Veri'nin temeli, web üzerinde nesneler için tekil tanımlayıcılar olarak HTTP URI (Sauermann et al., 2007) kullanmaktır. Böylece veri erişimi bir standarda oturtularak klasik ağa benzer olarak her yazılım tarafından veriye HTTP protokolü ile erişilir. HTTP ile erişilen verinin yazılımlar tarafından yorumlanabilmesi için RDF biçimine uygun olarak paylaşılması ilkesi tanımlanmıştır. RDF kullanımı beraberinde üst-verinin RDFS/OWL standardı ile tanımlanması ve sorgulama için HTTP protokolüne dayanan SPARQL standardının kullanılmasını getirmiştir. RDF, esnek bir biçimde verinin web üzerindeki herhangi bir kaynakta bulunan veri ile bağlanabilmesine olanak tanır. RDFS/OWL üst-veri dili de sözlüklerin anlamsal olarak tanımlanması ve mümkünse ağdaki sözlükler arasındaki bağlantıların yaratılması için gerekli yapıları içerir. Bu doğrultuda son ilke, ilgili veriler arasında RDF bağlantıları yaparak eldeki veriden başka kaynaklara ulaşılmasını ve böylece verinin zenginleşmesini amaçlar. Bağlanabilirlik yeteneği sayesinde dağıtık bilgi modelleri bütünleştirilmekte ve bütünleşmiş bilgi yazılımlar tarafından karar destek, arama, öneri yapma gibi birçok farklı amaçla kullanılabilmektedir.

2.4 Anlamsal Veri Bütünleştirme

Yaşam bilimleri alanında birçok biyolojik veri tabanı kullanılır hale gelmiştir. Bu veri tabanlarının heterojenliği veri entegrasyonunu zorlaştırmaktadır. Bazı entegrasyon problemleri kaynakların, ilişkilerin, farklı veri formatlarının ve eş anlamlarının bulunmasından, anlam belirsizliğinden kaynaklanır. Anlamsal Web standartları ve teknolojileri, yaşam bilimleri gibi dinamik ve dağıtık sistemlerin bütünleştirilmesi (entegrasyonu) için son yıllarda kullanılmaya başlanmıştır ve bir biyomedikal bilgi ağının modellenmesi için birlikte çalışabilir bir veri modeli sunar.

Veri tümleştirmenin ana problemlerinden biri aynı kaynağı tanımlayan farklı veri setlerindeki varlıkların belirlenmesidir. Literatürde bu konuda farklı anahtar kelimelerle benzer çalışmalar yapılmıştır (Clark, 2004): kayıt bağlama (record linkage), kimlik çözümleme (identity resolution), varlık eşleme (entity matching). Küçük veri setleri için bu bağlama işlemi, insan yargısı ile doğru şekilde yapılabilir. Fakat büyük veri setleri için “manuel” yaklaşımların kullanılması pratik değildir. Bu yüzden bu çaptaki verilerin bir bilgisayar tarafından eşlenmesini veya bağlanmasını sağlamak için otomatik yaklaşımlar kullanmak daha mantıklıdır. En basit bilgisayar destekli yaklaşım, aynı kimlik numaralarını veya bir varlığı tanımlayan bazı özellikleri kullanarak tanımlanmış kurallara dayanır. Var olan birçok yöntem böyle varlıkları belirlemek için kurallar kullanır veya bu kuralları öğrenmeyi amaçlar (Isele and Bizer, 2012).

Web ortamındaki (bağlı) veriye erişmek ve sorgulamak için SPARQL 1.0 ve SPARQL1.1 sorgulama araçları, farklı sözlük ve ontolojiler arasında anlam bütünlüğünü sağlamak için ise normalizasyon araçları mevcuttur. Bu araçlardan OWL, RDFS üst-ontolojisi tanımlayarak, SWRL (Horrocks et al., 2004) ve RIF (Kifer, 2008) gibi diller kural tanımlayarak, SPARQL Construct veri dönüşümü yaparak ve R2R (Bizer and Schultz, 2010) gibi ontoloji eşleme dilleri ise ontolojiler arasındaki kavramları eşleyerek ontoloji normalizasyonu yapar. Kimlik çözümleme için SILK (Volz et al., 2009) ve LINES (Ngomo and Auer, 2011) gibi kural tabanlı varlık bağlama/eşleme araçları mevcuttur. Ayrıca veri ağının filtrelenmesi için Named Graph (Carroll et al., 2005) yaklaşımı, birçok RDF saklama aracı tarafından desteklenmektedir. LDIF (Schultz et al., 2011) gibi anlamsal veri entegrasyonu araçları, veri entegrasyonunun her bir aşaması için bir çözüm kümesi sunmaktadır.

Veri bütünleştirme amacıyla, veri entegrasyonunun birçok aşaması için kullanılan LDIF aracının kullanılması mümkündür. RDF’e çevrilen verinin farklı veri setlerinde ontolojilerin eşlenmesi için veya hedef bir ontoloji dönüşümü

Çizelge 2.2: R2R dili ile yazılmış bir kural örneği

```

@prefix r2r: <http://www4.wiwiwiss.fu-berlin.de/bizer/r2r/>.
@prefix owl: <http://www.w3.org/2002/07/owl#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix mp: <http://www.example.org/smw-lde/smwTransformations/>.
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix pharmgkb_voc: <http://bio2rdf.org/pharmgkb_vocabulary:>.
@prefix omim_voc: <http://bio2rdf.org/omim_vocabulary:>.
@prefix pmkb_voc: <http://bio2rdf.org/pmkb_vocabulary:>.

mp:PharmGene
a r2r:ClassMapping;
r2r:prefixDefinitions ""pmkb_voc: <http://bio2rdf.org/pmkb_vocabulary:>.
pharmgkb_voc: <http://bio2rdf.org/pharmgkb_vocabulary:>.
xsd: <http://www.w3.org/2001/XMLSchema#>.""";
r2r:sourcePattern "?SUBJ a pharmgkb_voc:Gene";
r2r:targetPattern "?SUBJ a pmkb_voc:Gene" .

mp:OMIMGene
a r2r:ClassMapping;
r2r:prefixDefinitions ""pmkb_voc: <http://bio2rdf.org/pmkb_vocabulary:>.
omim_voc: <http://bio2rdf.org/omim_vocabulary:>.
xsd: <http://www.w3.org/2001/XMLSchema#>.""";
r2r:sourcePattern "?SUBJ a omim_voc:Gene";
r2r:targetPattern "?SUBJ a pmkb_voc:Gene" .

```

için R2R dilli ile kurallar yazılabilir. Örnek olarak iki ayrı veri setinde (Griffith and Griffith, 2004) , PharmGKB (P-Star et al., 2011) tanımlanmış farklı gen kavramlarını hedef bir sınıfa çevirmek için Çizelge 2.2’de verilen bir kural yazılmıştır.

Ayrıca farklı veri setlerinde tanımlanmış aynı varlıkların eşlemesi için LDIF için bulunan SILK çerçevesi varlık eşlemesini sağlar. Çizelge 2.3’te SILK dili ile yazılmış iki farklı veri setinde aynı adı taşıyan genlerin eşlemesi için yazılan bir kural verilmiştir.

2.5 Bilgi Çizgesi

Bilgi çizgesi, gerçek hayattaki varlıkların ve arasındaki ilişkilerin çizge olarak modellenmesi ile oluşan bilgiyi tanımlayan çeşitli ve heterojen gerçeklerin toplamına denir. Bir bilgi çizgesi varlıkların düğümler, ilişkilerin ise kenarlar olarak temsil edildiği çok-ilişkisel (multi-relational) bir veri modelidir. Bir kenar örneği

Çizelge 2.3: Silk dili ile yazılmış bir kural örneği

```

<Silk>
....

<Interlinks>
<Interlink id="genes">
<LinkType>owl:sameAs</LinkType>
<SourceDataset dataSource="pharmgkb" var="a">
<RestrictTo>?a rdf:type pharmgkb_voc:Disease</RestrictTo>
</SourceDataset>

<TargetDataset dataSource="omim" var="b">
<RestrictTo>?b rdf:type omim_voc:Disease</RestrictTo>
</TargetDataset>

<LinkageRule>
<Aggregate type="max">
<Compare metric="equality">
<Input path="?a/rdfs:label" />
<Input path="?b/rdfs:label" />
</Compare>

</Aggregate>
</LinkageRule>

<Filter threshold="1.0" />
</Interlink>

</Interlinks>
</Silk>

```

bir gerçeğin üçlü (<baş varlık, ilişki, son varlık> - <head, relation, tail>) olarak ifadesidir. Bazı gerçekler, varlıkları birbiriyle ilişkilendirirken (John, Mary'nin arkadaşıdır, John Jurassic Park'ı sever.), bazıları da varlıkların özelliklerini tanımlar (Jurassic Park, 9 Haziran 1993'te gösterime girdi).

Çizge veri modeli, verideki bağlantıları (ilişkileri) kullanarak yeni veri ögelerinin kolayca var olan bilgiye bağlanmasını ve çizgedeki bağlantıları dolaşarak diğer dolaylı bilgilere erişimi sağlar. Ayrıca bilgi çizgeleri, bilgiyi bir ontoloji ile beraber çizge olarak kodlayıp çıkarsama yaparak yeni bilgiler edinebilirler. Bilgi çizgeleri sahip oldukları ontoloji sayesinde kendi kendini tanımlayıcı bir üst-veriyi de içerirler. Bu üst-veriler verinin kendisini tanımlarken aynı zamanda verinin kaynağını ve versiyon bilgisini de içerebilirler. Bilgi çizgeleri, kullanıcılara neyi bildiklerini, nasıl bildiklerini ve bilginin hangi kaynaktan alındığını cevaplayabilir. Bu sayede kullanıcılar, sürekli değişen bilgilerin güvenilirliğini ve kredibilitesini değerlendirebilir. Bilgi çizgesindeki ontolojiler, barındırdıkları sözlük vasıtasıyla veri paylaşımını ve ortak bir terminoloji kurulmasını sağlar. Aynı zamanda bilgi çizgesindeki sözlük kullanılarak, doğal dile yakın bir tarzda, varlıkların ve ilişkilerin ismi cinsiden ifade edilen sorgular yazılabilir.

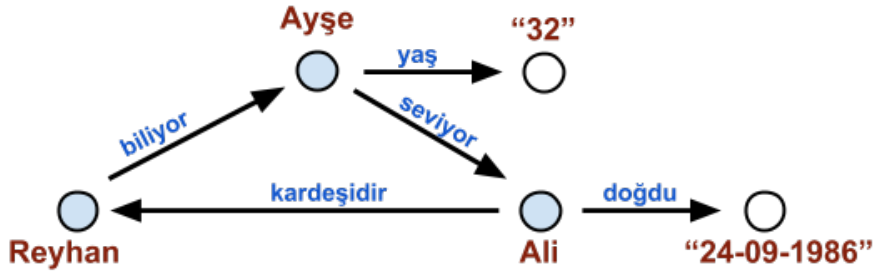
Yapay Zeka araştırmacıları tarafından veri gösterimi/temsili için geliştirilen en büyük çabalardan olan Anlamsal Ağ, veriyi bilgi çizgesi olarak kodlar. Wilcke ve arkadaşları (Wilcke et al., 2017) Anlamsal Ağ'da kullanılan bilgi çizge veri modelini, üç temel ilkeye dayandırmışlardır:

1. Bilgi, gerçekler (ifadeler) olarak kodlanmalı

Anlamsal Ağ'ın ardındaki temel fikir, bilginin gerçekler kullanılarak ifade edilmesi gerektiğidir. Aşağıdaki örnek göz önünde bulundurulduğunda:

- *Reyhan biliyor Ayşe.*
- *Ayşe seviyor Ali.*
- *Ayşe yaş "32".*
- *Ali kardeşidir Reyhan.*
- *Ali doğdu "24-09-1986".*

Yukardaki örnekler, Anlamsal Web'in veri modeli olan RDF ile tanımlanabilen ifadelerdir. Bu model, her bir ifadenin soldan sağa sırayla iki varlığı (özne ve nesne) ilişkilendiren tek bir özellik (fil) içermesi gerektiğini belirtir. Bu üçü birlikte bir RDF üçlüsü olarak adlandırılır. Yukarıdaki örneği Şekil 2.6'da gösterildiği gibi yönlü bir çizge olarak da gösterebiliriz. Kenarlar



Şekil 2.6: Verilen örnek için RDF çizge gösterimi

ilişkileri temsil eder. Düğümler in şekilleri rollerini yansıtır: içi dolu daireler varlıkları temsil eder, içi boş daireler ise nitelikleri temsil eder.

Kaynaklar metin, sayı veya tarih gibi varlıklar (şeyler) veya sabitler olabilirler. Üçlüler, iki kaynak varlık olduğunda iki kaynak arasındaki ilişkileri, sağdaki kaynak varlık ve soldaki kaynak sabit olduğunda ise nitelikleri ifade edebilirler. Örneğin örnekteki üçlülerin son satırı, Ali'nin doğum tarihini "24-09-1986" sabit değeri ile ifade eder.

RDF veri modeli yukarıda verilen kurallar dışında bilgi mühendislerinin bilgilerini nasıl modellemesi gerektiği konusunda bir kısıtlama getirmez. Örneğimizi farklı bir şekilde modelleyebilirdik; örneğin tarihleri kaynak olarak. Modelleme seçimleri genel olarak veri kümesinin uygulama alanına ve amaçlarına bağlıdır.

2. Arka plan bilgisi (background knowledge), ontoloji içinde ifade edilmeli

RDF veri modeli, modelleme tercihleri üzerinde özgürce kontrol sağlarken, ontolojiler bir alandaki bilginin topluluk tarafından ifade edilmesi için bir yol sunar. Bu amaçla ontolojiler bir alanı tanımlayan sınıfları (varlık tipleri) ve niteliklerini, bu sınıflar ve nitelikleri üzerindeki kısıtları ve çıkarımları içerir. Örneğin, bir ontolojide "Person" sınıfı tüm bireyler için bir tip (type) olarak tanımlanarak varlıklar ile sınıflar ilişkilendirilebilir.

Reyhan type Person.

Ayşe type Person.

Ali type Person.

Reyhan, Ayşe ve Ali Person sınıfının örnekleri olduğu söylenir. Person sınıfı "Human" sınıfı ile eşit ve "Animal" sınıfından ayrık ve "Agent" sınıfının

alt sınıfı olması gibi özellikler taşıyabilir. Bu son özellik (alt sınıf olma durumu) özyinelemeli bir özellik örneğidir ve bu özellik RDF yalın modelini genişleten RDF Schema (RDFS) ontolojisi ile tanımlanabilir. Diğer ilişkiler ise daha karmaşık yapıları modelleyebilen ve daha ifade edilebilir bir model olan OWL ontolojisi ile tanımlanabilir.

3. Veri Kümeleri arasında bilgi yeniden kullanılmalı

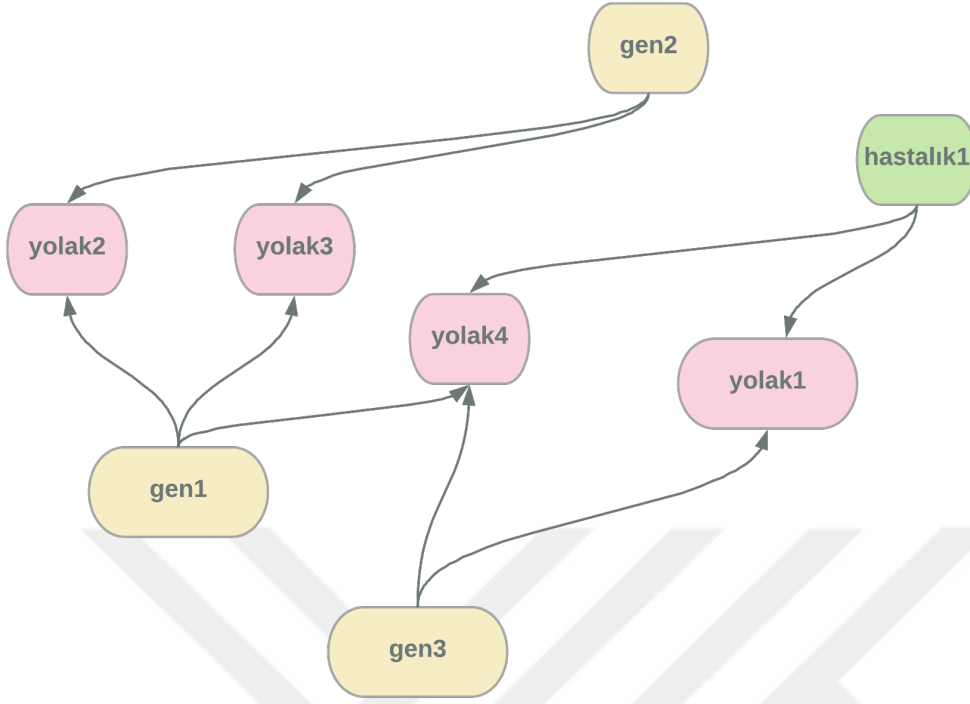
Bilgiyi yeniden kullanmak için kaynakların isimlerini kullanmak yerine benzersiz bir tanımlayıcı kullanılabilir. Bu tanımlayıcılar, URI veya IRI ile adlandırılır ve bir web adresi biçiminde olurlar. Örneğin Ege Üniversitesinde görevli olan Remzi ve Erkan’a atıfta bulunmak için “<http://bilmuh.ege.edu.tr/staff/RemziCelebi>” ve “<http://bilmuh.ege.edu.tr/staff/ErkanYasar>” IRI’lerini kullanabiliriz. Bunları genellikle daha kısa bir formda yazmak için (ege:RemziCelebi ve ege:ErkanYasar) bir namespace (ege: <http://bilmuh.ege.edu.tr/staff/>) tanımlanabilir. Ayrıca varolan bir ontolojinin sözcükleri (örneğin FOAF’ın knows) kullanılarak, RDF üçlülere ortak ve paylaşımlı bir bilgi üretirler.

2.6 Biyomedikal Ağlarda Anlamsal Çizge Madenciliği

Birçok biyomedikal ağ, tipik olarak, farklı ilişki tiplerini ve bağlantılı biyomedikal varlıkları içerir. Anlamsal Web’in bilgi çizge modeli, bu ağlardaki bilginin temsili için elverişli bir modeldir. Bilgi çizgeleri, tüm düğümlere ve tüm bağlantılara bir etiket atanmış özel bir çizgedir. Bilgi çizgeleri üzerinde anlamsal çizge madenciliği şu şekilde uygulanabilir; 1) bilgi çizgesini, geleneksel çizge yapısına dönüştürerek klasik çizge/ağ analiz yöntemlerinin uygulanması veya ii) anlambilimsellik (semantics) kullanılarak çıkarsama yapılması. Chen ve arkadaşları (Chen et al., 2009) çalışmalarında, biyomedikal veri ağları için anlamsal çizge madenciliğini 4 ana başlık altında incelemiştir:

2.6.1 Anlamsal Derecelendirme

Anlamsal Derecelendirme (Semantic Ranking), çizgenin elemanlarını (düğüm, ayırıt, alt-çizge) merkezlik, bağlantı analizi gibi yöntemler ile sıralamayı hedefler. Anlamsal derecelendirmenin bir çok uygulama alanı olabilir; örneğin belli bir hastalık için gen önceliklendirmesi için kullanılabilir. Şekil 2.7’de gösterilen bir hastalık ile ilişkili yolaklar vasıtasıyla, hastalık ile ilişkili genleri öncelendirmek için hastalığa ilişki gen-yolak bilgilerini içeren bir bilgi ağı kullanılabilir.



Şekil 2.7: Hastalık ile ilişkili yolaklar vasıtasıyla hastalık ile ilişkili genlerin öncelendirilmesi

Çizgenin önemli elemanlarını belirlemek için derece merkezliği (degree centrality), öz vektör merkezliği (eigenvector centrality), yakınlık (closeness) ve arada bulunma (betweenness) gibi klasik ağ analiz metotları yaygın olarak kullanılır. Web bilgi geri-getirimi bağlamında objeleri derecelendirmek için webin bağ yapısını kullanan en önemli yaklaşımlar, PageRank (Page et al., 1999) ve HITS (Kleinberg, 1999) algoritmalarıdır. Bu yaklaşımlar tek tip düğüm ve bağa sahip çizgeler için düğümleri derecelendirirler ve bu yapıdaki herhangi bir veri ağı için kullanılabilirler.

Arama motoru Google'ın kullandığı PageRank algoritması, web sayfaları için otorite/popülerite (rank) değeri hesaplar. PageRank, bir web gezgin kullanıcısının bir web sayfasından başlayarak o sayfanın bağlantılarından birini rastgele seçip dolaştığı ve ara sıra başka bir başlangıç sayfasına atladığı rastgele bir yürüyüşü (random walk) benzetir. Bu rastgele yürüyüş sonsuza dek tekrarlanırsa, verilen bir sayfanın rütbesi/derecesi rastgele gezen kullanıcının o web sayfasına uğrama sıklığı olarak hesaplanır. Bu rastgele sürecin kararlı durum dağılımı hesaplanarak bir sayfanın popülerite değeri belirlenebilir.

HITS biraz daha karmaşık bir süreci varsayar ve web sayfalarını iki tür olarak modeller: hub ve otorite (authority). Hublar birçok otorite sayfaya bağlantısı olan web sayfalarıdır. Otoriteler ise birçok hub tarafından bağlantı alan web sayfalarıdır.

Her sayfanın bir hub ve otorite skoru atanır. Bu skorlar, bir sayfanın skorlarını hemen yanındaki komşularının skorlarına bağlı olarak güncelleyen iteratif bir algoritma tarafından hesaplanır. Bu yaklaşım bir yerine iki farklı rastgele yürüyüş gerçekleştirerek (hub ve otorite geçişleri ile) PageRank'e benzerlik gösterir.

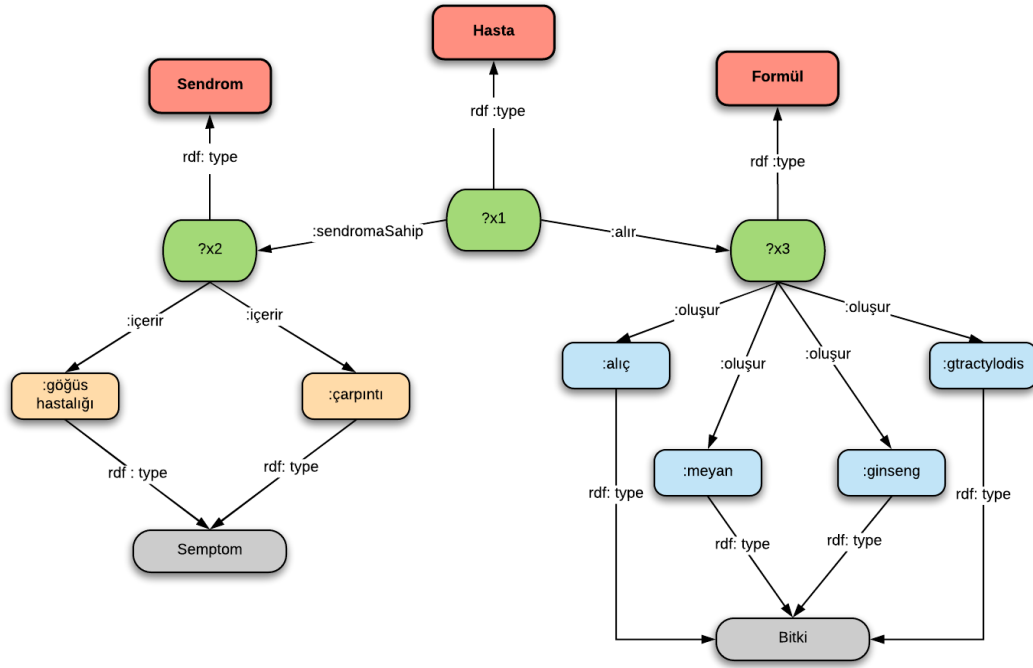
Bunların yanında birden fazla düğüm ve bağ tipleri içeren veri ağları için ObjectRank (Balmin et al., 2004) ve PopRank (Nie et al., 2005) gibi anlamsal derecelendirme metotları mevcuttur. ObjectRank, veri tabanını çizge olarak modelleyerek verinin bağlı yapısını kullanır ve çizgenin düğümlerini sorgulanan kelimelere göre sıralayabilmektedir. ObjectRank, PageRank'in yöntemini veri tabanındaki varlıkların derecelendirilmesi için kullanmış ve genişletmiştir. ObjectRank, kavramlar arasındaki ilişkilerin alan uzmanları tarafından ağırlıklandırılmasını gerektirir ve PageRank algoritmasını genişleterek objenin sorgulanan anahtar kelimelere olan yakınlık değerini hesaplar. ObjectRank verinin üst verisini (şemasını) bir çizge olarak ele alıp bu şema üzerindeki sınıfların otorite transfer oranını hesaplamak için kullanılır.

PopRank, PageRank modelini genişleterek bir ilişki ağının içindeki bir objeye işaret eden her bağlantı türü için ayrı popülerite yayılım faktörü-PYF (popularity propagation factor) ekler. Her farklı bağlantı türü için farklı PYF değeri kullanır. PopRank PYF'leri otomatik belirlemek için öğrenme tabanlı bir yaklaşım önermiştir. Bu yaklaşım, alan uzmanından objelerin kısmi sıralamasını kullanarak farklı bağlantılar için PYF değerlerini öğrenir. Benzetilmiş Tavlama (Simulated Annealing) algoritması, olası bütün PYF kombinasyonlarının tarama uzayını gözden geçirerek en uygun PYF kombinasyonunu bulmaya çalışır.

2.6.2 Sık Tekrarlanan Alt-Çizge Bulma

Medikal veri ağı içinde sık tekrar eden örüntülerin bulunması ağ analizinin diğer önemli bir problemidir. Sık tekrarlanan örüntülerin tespit edilmesi, bir hasta grubu için çok uygulanan tedavi yöntemi, ilaçları ve bu ilaçların etkileşimlerini bulmak için kullanılabilir. Şekil 2.8'de dört bitkiyi içeren formül (sağ), iki semptomla (solda) kendini gösteren bir sendroma sahip hastayı tedavi etmek için sıklıkla kullanıldığı, alt-çizge olarak ifade edilmiştir. Bu örnekte $?x_1$, $?x_2$ ve $?x_3$ değişkenleri gösterir.

Sık tekrarlanan alt-çizgeleri bulmak için Apriori özelliği (Agrawal et al., 1994) kullanan yöntemlerde, tipik olarak aday alt-çizge oluşturma fazını takiben eşleştirme fazı gelir. Doğal eşleştirme için alt-çizge izomorfizm testi gereklidir



Şekil 2.8: Sık görülen anlamsal alt-çizge örneği

ki bunun için de etkin algoritmalar gereklidir. Inokuchi ve arkadaşları (Inokuchi et al., 2000), AGM algoritmasını tanımlamışlardır. Bu algoritma Apriori tabanlı bir yöntem olup çizge veritabanı içinde en az destek ile bütün alt-çizgeleri bulur. Kuromachi ve arkadaşları (Kuramochi and Karypis, 2001) ise AGM yöntemini, çizgenin komşuluk matrisini kullanarak ve aday alt yapılarını üretmek için bir optimizasyon tanımlayarak geliştirmişlerdir. Yan ve arkadaşları (Yan and Han, 2002), gSpan algoritmasını tanımlamışlar ve her çizgeyi bir derinlik-öncelikli arama koduna çevirerek ve bunu alfabetik olarak sıralayarak aday üretme maliyetinden kaçınmışlardır. Daha sonra bu arama ağacının üzerinde derinlik öncelikli arama (depth-first search) çalıştırmışlardır. Dehaspe ve arkadaşları (Dehaspe et al., 1998), Inductive Logic Programming tekniklerinden faydalanan toksikoloji alanındaki sık tekrarlanan örüntüleri bulmak için kullanmışlardır.

2.6.3 Çizge Kümelenmesi / Topluluk Tespiti

Çizgenin düğümlerinin özelliklerine göre çizge içindeki düğümlerin gruplandırılmasına dayanır. Bu özellikler, düğümlerin benzerliğine göre ya da düğümlerin bağlantılarına göre yapılabilir. Örneğin belli bir fenotip gösteren bir canlıdaki proteinlerin gruplanması için çizge kümelemesi yöntemleri kullanılabilir.

Topluluk/grup belirlemek için var olan tekniklerin çoğu ya birleştirici

(agglomerative) veya bölücü (divisive) yöntemler olarak sınıflandırılabilir. Spektral çizge bölümlendirme (spectral graph partitioning) yöntemleri, basit olması ve lineer cebir temelli olması sebebiyle çizge kümeleme algoritmaları arasında çok önemli bir role sahiptir. Bu yöntemler ile silinecek en az bağlantıyı belirleyerek verilen sayıda grup oluşturabilir (Newman, 2004; Ding et al., 2001). İlgili bir biçimde, Gibson ve arkadaşları (Gibson et al., 1998) HITS yetki matrisinin baskın öz vektörlerinin, webin topluluk yapısı için doğal bir ayrışma sağladığını göstermiştir. Grup tespiti için diğer yaklaşımlar, grupları birbirine bağlayan bağlantıları belirlemek için kenar arada bulunma (edge betweenness) ölçütünü kullanır (Wang et al., 2005). Kenar arada bulunma değeri en yüksek olan bağlantılar, çizgeyi bölmek için aşamalı şekilde kaldırılır.

Akış tabanlı olan Markov Kümeleme Algoritması (Markov Clustering Algorithm-MCL) (Enright et al., 2002), rastgele yürüme benzetimi yaparak şişme (inflation) işlemi ile ağ içindeki zayıf bağlantıları azaltıp güçlü bağlantıları artırmayı sağlar, genişleme (expansion) işlemi ile ise farklı düğümlere sıçramayı sağlar. Bu algoritma sonunda akışın koptuğu bölgeler meydana getirilerek grupların tespiti gerçekleşir.

Bir başka akış tabanlı olan Tekrarlı Rastgele Yürüme (Random Walk With Restart) (Macropol et al., 2009) Kümeleme algoritması, PageRank algoritması gibi bir merkezi düğümden başlayarak düğümler için bir derecelendirme yapar ve bir birbirine yakın dereceye sahip düğümleri bir kümeye yerleştirir. Bu merkezi düğümü belirlemek için de bütün düğümlerin eşit popülerite (rank) katsayısı ile başlayarak, yayılım sonucunda popülerite değerlerinin her bir adımda yakınsaması sağlanır. Bu yakınsama işlemi bittikten sonra en büyük popülerite değeri en merkezi düğüm olarak belirlenir.

Birleştirmeli yaklaşımlardan olan aşağıdan yukarı hiyerarşik kümeleme algoritmaları, bir düğümden başlayarak en yakın düğümleri veya kümeleri hiyerarşik bir yapının içine yerleştirerek kümelere ayırır. En yakın küme veya düğümleri belirlemek için benzerlik veya uzaklık ölçülür. UVCLUSTER ve Süper Mıknatıs Çeken Kümeleme (Super Paramagnetic Clustering) (Pan et al., 2013) gibi yaklaşımlar, birleştirmeli yaklaşımlara örnektir.

2.6.4 Anlamsal İlişki Keşfi

İki düğüm arasındaki ilişkiyi, aralarındaki bağlantıların türü, uzunluğu ve ağırlığı ile keşfeder. Anlamsal ilişki keşfi, tez kapsamında Bölüm 3'te detaylı bir

şekilde incelenmiştir.

2.7 Makine Öğrenmesi

Makine öğrenmesi, matematiksel ve istatistiksel işlemler ile veriler üzerinden tahminlerde bulunan sistemlerin modellenmesidir. Bu tezde bağ tahmini yapmak için denetimli makine öğrenmesi türlerinden sınıflandırıcılar kullanılmıştır. Sınıflandırıcılar, etiketlenmiş gözlemlerden yeni gözlemlerin etiketinin (hangi sınıfa dahil olacağını) tahmin edilmesini sağlar. Çıkış değerleri kategoriktir. Bu tezde kullanılan sınıflandırıcılar şunlardır: Lojistik Regresyon, Naive Bayes ve Rastgele Orman.

- **Lojistik Regresyon:** Sınıflandırma problemleri için lineer modellerin genelleştirilmiş bir halidir. Lineer regresyon, gözlemlerin devamlı (continuous) etiketlerinin ($y \in \mathbb{R}$) tahmini için kullanırken, lojistik regresyon, gözlemlerin sınıf etiketlerinin ($y \in 0, 1$) tahmini için kullanılır. Lojistik fonksiyon, hedef değişkenini, bağımsız kestirici değişkenlerin (özniteliklerin) lineer kombinasyonu şeklinde ifade ederek gözlemler için sınıf olasılıklarını hesaplar. Lojistik regresyon bağ tahmini problemi için bağın bulunmasını 1, bulunmamasını 0 olarak kodlayarak, özniteliklerin bağ olma durumunu lineer olarak ne kadar açıklayabildiğini belirler. Lojistik regresyon, verilen öznitelikler ile bağ oluşma olasılığını şu şekilde tanımlar:

$$\beta X = \ln\left(\frac{p}{1-p}\right) \quad (2.1)$$

X , öznitelik vektörünü ve β , her bir özniteliğe karşılık gelen katsayılardan meydana gelen vektörü temsil eder. p ise bağ olma olasılığını ifade eder.

Lojistik regresyon, özelliklerin katsayılarını ($\beta_0, \beta_1, \dots, \beta_n$) öğrenmek için katsayıların uygunluğunu ölçen maksimum olabilirlik (maximum likelihood) yöntemini kullanır. Lojistik regresyon, veri kümesindeki bütün gözlemlerin doğru sınıfta olma olasılığını maksimize etmek için en uygun katsayıları aşağıdaki maliyet fonksiyonuna göre öğrenmeye çalışır:

$$J = \prod_{y_i=1} p(y|x) = \prod_{y_i=1} h(X) \prod_{y_i=0} (1-h(X)) \quad (2.2)$$

$h(X)$ lojistik fonksiyonunu ifade eder:

$$h(X) = \frac{1}{1 + e^{-(\beta X)}} \quad (2.3)$$

- **Naive Bayes:** Naive Bayes (Domingos and Pazzani, 1997), Bayes teoremine dayanan olasılıksal bir sınıflandırma yaklaşımıdır. Naive Bayes; basit yapısı, hesaplama etkinliği ve yüksek doğru sınıflandırma başarımı nedeniyle metin madenciliği başta olmak üzere pek çok sınıflandırma çalışmasında başarı ile kullanılmaktadır. Bayes teoremi olayların birbirinden bağımsız olduğunu varsayar ve bir koşullu olasılığı şu şekilde tanımlar:

$$P(A | B) = \frac{P(B | A)P(A)}{P(B)} \quad (2.4)$$

Naive Bayes, verilen bir sınıf için özniteliklerin birbirinden bağımsız olduğu (sınıf koşullu bağımsızlık- class conditional independence) varsayımı yaparak, $P(X|C_i)$ koşullu olasılığını maksimize etmeye çalışır:

$$P(X|C_i) = \prod_{k=1}^n P(X_k | C_i) \quad (2.5)$$

$X = (x_1, x_2, \dots, x_n)$ gözlemler kümesini ve $C = (c_1, c_2, \dots, c_m)$ ise sınıflar kümesini gösterir.

- **Rastgele Orman:** Rastgele Orman (Random Forest) (Ho, 1995), sınıflandırma başarısından dolayı son zamanlarda çok tercih edilen bir topluluk öğrenme yöntemidir. Rastgele seçilen bir alt öznitelik kümesi ve eğitim verisi ile oluşturulan n tane karar ağacının sonucunu birleştirir. Bu topluluk sınıflandırıcı, farklı karar ağaçlarının oluşmasına imkan verir ve bir gözlem için her bir ağacın verdiği sınıflandırma etiketlerini oylayarak hedef etiketini bulur. Rastgele Orman oluşturulurken, budama (pruning) ve durdurma kuralı (stopping condition) uygulanmaz. Bu, tahminin doğruluğunu ve gücünü artırır. Rastgele Orman, sınıflandırma yapmak için şu adımları izler:

1. Eğitim veri kümesi içinden N adet rastgele örnek, yer değiştirmeli olarak (with replacement) seçilir.
2. M tane toplam kestirici değişkeni içinden $m \leq M$ olacak şekilde m adet değişken seçilir.
3. Seçilen veri kümesi ve değişkenler ile karar ağacı oluşturulur.
4. Bu karar ağaçlarının bir örnek için yaptığı sınıf tahminleri arasında en yüksek oya sahip olan sınıf etiketi, o örneğin sınıfı olarak seçilir.

3 ANLAMSAK İLİŞKİ KEŞFİ YÖNTEMLERİ

Anlamsal ilişki keşfi için literatürde birçok yöntem olmasına karşın, homojen ağlar için sosyal ağ analizinde kullanılan bağ tahmini ve heterojen ağlar için ise çoklu-ilişkisel bağ tahmini adı altında kapsamlı ve metodolojik olarak incelenmiştir.

3.1 Sosyal Ağlarda Bağ Tahmini

Sosyal ağda bağ tahmini (Lü et al., 2011) bir ağın var olan bağlantıları ve topolojik özelliklerinden yararlanarak olması beklenen veya eksik olduğu düşünülen bağlantıların tahmin edilmesine denir. Bu bağlamda bağ tahmini yöntemleri öğrenme türüne göre ikiye ayrılır; denetimli (supervised) ve denetimsiz (unsupervised). Denetimli bağ tahmini, problemi, tipik ikili (binary) 0/1 sınıflandırma problemine çevirerek ele alır (Hasan, Al Hasan, and Zaki 2011). Bunun için makine öğrenmesi yöntemleri kullanarak herhangi iki düğüm arasında bağlantı olup olmadığı tahmin edilir. Eğitim kümesinde (training set), iki düğümün kendine has özellikleri veya yapısal benzerlikleri girdi olarak, bağ olma durumu (1) veya olmama durumu (0) ise çıktı olarak kullanılır. Bir diğer yaklaşım olan denetimsiz bağ tahmini ise iki düğüm arasındaki olası bağ için düğümlerin topolojik özelliklerini kullanarak bir benzerlik skoru hesaplar. Ağ içinde var olmayan fakat olası bağlar bu skorlara göre sıralanarak, bağların sınıflarına (0/1) karar vermek için bir eşik değeri belirlenir. Denetimli bağ tahmini, sınıfların benzerlik skorları arasındaki sınırı öğrendiğinden, denetimsiz öğrenmeden daha iyi performans gösterme potansiyeline sahiptir.

Benzerlik skoru için ortak komşuluk ölçütünü kullanan Common Neighbors (Liben-Nowell and Kleinberg, 2007) , Jaccard (Jaccard, 1912), Adamic/Adar (Adamic and Adar, 2003) gibi lokal yaklaşımların yanında; patika tabanlı en kısa patika uzaklığı (shortest path distance), Katz, Rooted PageRank, PropFlow (Liben-Nowell and Kleinberg, 2003; Lichtenwalter et al., 2010) gibi yaklaşımlar da mevcuttur. Sosyal ağ analizinde kullanılan yöntemler, eğitim çizgesinde var olmayan bir bağın bulunma veya ileride görülme olasılığını hesaplar. Her $\langle x,y \rangle$ düğüm çifti için hesaplanan bu olasılık değerleri, benzerlik veya yakınlık skoru olarak da görülebilir.

3.1.1 Komşuluk Tabanlı Bağ Tahmini Yöntemleri

- **Ortak komşular:** Bu yöntemde, x ve y düğümlerinin ortak komşu sayısı, x ve y arasındaki benzerlik değeri olarak hesaplanır. $\Gamma(x)$, x düğümünün komşularını $\Gamma(y)$, y düğümünün komşularını gösterebilir, bu iki düğümün ortak komşuları $|\Gamma(x) \cap \Gamma(y)|$ olarak tanımlanır. Ortak komşuların sayısının kullanma fikri, ağ geçişlilik özelliğine dayanır. Basitçe sosyal ağlarda bir x düğümü z düğümüne bağlı ve y düğümü de z düğümüne bağlı ise, x düğümünün y düğümüne bağlanma olasılığı çok yüksektir. Böylelikle ortak komşuların sayısı artarsa bu iki düğümün arasında bağ olma şansı da yükselir. Newman (Newman, 2001) bu ölçütü, iş birliği ağlar bağlamında x ve y düğümlerinin ortak komşu sayısı arttıkça bunların zaman içinde iş birliği yapma olasılıkları arasındaki pozitif ilişkiyi göstermek için kullanmıştır.
- **Jaccard katsayısı:** Ortak komşular sayısı normalize edilerek hesaplanan bir benzerlik ölçütüdür, x ve y düğüm çifti için Jaccard katsayısı şöyledir:

$$Jaccard(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \quad (3.1)$$

Kavramsal olarak x ve y 'nin komşularının birleşimi içinden rastgele bir seçim yapıldığında ortak komşuların seçilme olasılığını tanımlar.

- **Adamic/Adar:** Adamic ve Adar (Adamic and Adar, 2003) bu skoru iki web sayfası arasında benzerlik ölçütü olarak önermişlerdir. Bağ tahmini için ortak komşularının komşu sayıları temel alınarak özelleştirilmiştir. İki düğüm arasındaki benzerliği, ortak komşularının daha az komşusu olanı daha fazla ağırlıklandırarak hesaplar. Yapılan hesaplama şöyledir:

$$Adamic/Adar(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log |\Gamma(z)|} \quad (3.2)$$

3.1.2 Patika Tabanlı Bağ Tahmini Yöntemleri

Bu yaklaşım tekniğinde ise iki düğüm arasındaki yollar dikkate alınır. Komşuluk tabanlı yaklaşımlara göre çok daha karmaşıktır. En sık kullanılan patika tabanlı yaklaşımlar; En kısa uzaklık, Katz , PropFlow, Ortalama gidiş-geliş süresi ve Rooted PageRank.

- **En kısa uzaklık:** Bu yöntem, $\langle x, y \rangle$ bağlantısını tahmin etmek için benzerlik ya da bir olasılık olarak, x ve y düğümleri arasındaki en kısa yol uzunluğunu

kullanır. Bir ağdaki iki düğüm arasındaki uzunluk azaldıkça bu iki düğüm arasında bağ oluşturma olasılığı artar. Eğer iki düğüm arasında hiç bir yok/bağlantı yoksa bu olasılık 0 olarak hesaplanır.

- **Katz:** Leo Katz (Katz, 1953) tarafından önerilen bu ölçüt, en kısa uzaklığın bir türevidir fakat genellikle bağ tahmini için daha iyi çalışır. X ve y düğümleri arasındaki var olan tüm yolların sayısını toplar. Uzun yolların katkısını düşürmek için yolları uzunluğuna göre β^l olarak ağırlıklandırır. Katz değeri şöyle hesaplanır:

$$Katz(x, y) = \sum_{l=1}^{\infty} \beta^l \cdot |paths_{x,y}^{<l>}| \quad (3.3)$$

$|paths_{x,y}^{<l>}|$ uzunluğu l olan x ve y arasındaki tüm yolların kümesi ve β ise yol ağırlıklarını düzenleyen serbest parametredir. Küçük β değeri için en kısa uzaklık ölçütüne benzer şekilde davranır. Bu ölçütün bir dezavantajı hesaplama açısından pahalı olmasıdır.

- **Ortalama gidiş-geliş süresi:** Bir çizgede rastlantısal yürüyüş yapılarak x düğümünden başlayıp y düğümüne ulaşmak için gerekli adımların ortalama sayısı $H(x, y)$ ile gösterilsin, x ve y arasındaki ortalama gidip gelme süresi aşağıdaki gibi tanımlanır:

$$C(x, y) = H(x, y) + H(y, x) \quad (3.4)$$

Kısa gidiş-geliş zamanı bu iki düğümün benzer olduğunu gösterir. Bu ölçütün hesaplanması kolay olmasına rağmen, değişken (varyansı yüksek) olduğu için bu özellik kullanılarak yapılan tahminler zayıf olabilmektedir.

- **Rooted PageRank:** PageRank çizgenin yapısını dikkate alarak her bir düğüm için tek bir popülerite değeri hesaplar. Ancak PageRank bir düğümün sahip olduğu bir özelliktir, bir düğüm çifti $\langle x, y \rangle$ arasındaki benzerliği gösterecek şekilde değiştirilmelidir. PageRank'ın orijinal tanımı iki varsayım altında bir düğüm için önem derecesini belirtir; bir gezgin α olasılıkla bir web sayfasından rastgele bir sayfaya atlar ve $1 - \alpha$ olasılık ise sayfadaki bir bağlantıya tıklayarak gezintisine devam eder. Bu rastgele yürüyüş sonsuza kadar devam etmesi halinde gezgin sayfaları önem derecesi oranında ziyaret etmiş olacaktır. Rooted PageRank algoritması (Liben-Nowell and Kleinberg, 2007), PageRank (Page et al., 1999) algoritmasının genişletilmiş bir versiyonudur ve bir kök düğümünden başlayarak her bir düğüme bir popülerlik skoru atar. Böylece o kök düğüm ile diğer her bir düğüm arasında bir benzerlik skoru hesaplanır:

$$R_{i+1} = (1 - \alpha)M \times R_i + \alpha P \quad (3.5)$$

Denklem 3.5’da görüldüğü üzere R_i düğümlerin i . yinelemesi için sıralama vektörü iken R_{i+1} sonraki yineleme için sıralama vektörüdür. α geçerli düğümden kök (başlangıç) düğüme atlamak için rastlantısal yürüyüş olasılığıdır ve $(1-\alpha)$ rastgele bir komşuya olan bir atlama olasılığıdır. $P \llbracket_{N \times 1}$ vektörü rastlantısal bir yürüyüşle başlangıç düğümlerini belirler ve başlangıç düğümlerine karşılık gelen değerler 1 ve α olasılığıyla atlayan diğerleri için ise 0’dır. M çizgeye karşılık gelen bir kare stokastik matriksidir ve böylece her matris girdisi aşağıdaki şekilde atanır:

$$m_{ij} = \begin{cases} \frac{1}{N_j}, & \text{j'den i'ye bir bağlantı varsa} \\ 0, & \text{aksi durumda} \end{cases} \quad (3.6)$$

N_j , j düğümünün çıkış derecesini belirtir.

- **PropFlow:** PropFlow yaklaşımı (Lichtenwalter et al., 2010), rastlantısal yürüyüş ile bir ağ içinde bağların ağırlıklarına göre bilgi yayılımını hesaplayan bir algoritma sunar. Rooted PageRank yaklaşımına benzer ancak bir düğümünün yerel komşularıyla sınırlandırılmış bir yürüyüş yapar. Rooted PageRank’in aksine PropFlow yürüyüşü yeniden başlatma veya yakınsama gerektirmediği için Rooted PageRank’den hesaplama olarak çok daha etkindir. Algoritma, basitçe 1 uzunluğuna sınırlandırılmış derinlik-öncelikli tarama yapar. Bu yaklaşım sınırlı bir rastlantısal yürümenin l adımda ve daha az adımda x ’ten başlayıp y ’de sona erme olasılığını hesaplar.

PropFlow, x ve y düğümleri arasında yeni bağ oluşma olasılığını aşağıdaki gibi hesaplanır:

$$PropFlow(x, y) = Score(x) \frac{\omega(x, y)}{|\Gamma(x)|} \beta^{h-1} \quad (3.7)$$

h x ’ten y ’e kaç adımda vardığını (zıplama sayısı), $\omega(x, y)$ x ve y arasındaki bağın ağırlığını, β aşınma faktörünü, $\Gamma(x)$ x ’in komşularını gösterir ve y düğümü rastlantısal yürümede x ’ten sonraki düğümdür.

3.2 Çoklu-İlişkisel Bağ Tahmini

3.2.1 Mantıksal Çıkarım Tabanlı Bağ Tahmini Yöntemleri

Anlamsal ilişki keşfinin diğer bir yöntemi ise bilgi çizgesi/ağı üzerinde otomatik çıkarım yöntemlerinin uygulanmasıdır. Çıkarsama (reasoning) araçları tutarlılık kontrolü, sınıfların yeni örneklerinin bulunması, örneği olan sınıfın bulunması, bir kavramın kavram hiyerarşisinde yerinin bulunması gibi çıkarsama işlemlerini yapar. Çıkarsama araçları iki gruba ayrılabilir; Betim Mantık tabanlı ve kural tabanlı çıkarsama araçları.

Anlamsal Web’de OWL-DL dilini kullanan Betim Mantık tabanlı çıkarsama araçları daha popüler bir kullanıma sahiptir. Bu araçlar OWL-DL ile tanımlanan kuralları ve Betim Mantık çerçevesini kullanarak çıkarım görevini gerçekleştirirler. OWL-DL, RDFS’in ifade edilebilirlik açısından genişletilmiş halidir, RDFS Plus olarak adlandırılır. OWL-DL’in temel ve mantık yapıları, RDFS ile karşılaştırmalı olarak Çizelge 3.1’de verilmiştir. Betim Mantık tabanlı araçlardan bazıları olan FaCT (Horrocks, 1999), FaCT++, Racer (Haarslev and Möller, 2001) ve Pellet (Sirin et al., 2007) farklı mantık türlerini çalıştırmalar.

Bir çıkarsama örneği vermek gerekirse, bilgi tabanımızda aşağıdaki örnekler bulunsun.

A :ebeveyn C (A, C’nin ebeveynidir.)

B :ebeveyn C (B, C’nin ebeveynidir.)

A rdf:type Erkek (A bir Erkektir)

B rdf:type Kadın (B bir Kadındır)

Bu bilgi tabanından A ve B’nin karı-koca olduğu çıkarımı yapılabilir.

3.2.2 Vektör Gömme Tabanlı Bağ Tahmini Yöntemleri

Makine öğrenmesi gibi işlerde varlıkları bağlamak, varlığın farklı ilişkilerinin enginleştirilmesinde ve daha fazla öznitelik yaratılmasında yardımcı olur. Son yıllarda LOD kullanılarak veri madenciliği ve makine öğrenmesi için öznitelik

Çizelge 3.1: RDFS ve OWL-DL temel yapıları ve mantık yapıları

	RDFS	OWL-DL
Temel Yapılar	subClassOf subPropertyOf domain range	ObjectProperty DatatypeProperty FunctionalProperty InverseFunctionalProperty equivalentClass equivalentProperty sameAs inverseOf TransitiveProperty SymmetricProperty
Mantık Yapıları	Union Intersection Equivalent Classes Equivalent Properties Equivalent Individuals	-

oluşturan yaklaşımlar önerilmiştir (Selbas et al., 2011; Cheng et al., 2011). Bu yaklaşımların çoğu öznitelik seçmeyi kullanıcının manuel olarak, çoğu durumda SPARQL sorgusu olarak, tanımlanılmasını gerektirir. FeGeLOD (Paulheim and Fümkrantz, 2012) ve RapidMiner Linked Open Data Extention (Ristoski et al., 2015) gibi yaklaşımlar ise LOD'dan elde edilen öznitelikler ile zenginleştirmek için 6 farklı denetimsiz öznitelik oluşturma stratejisi kullanmışlardır.

RDF çizgesi için kullanılan diğer yaklaşımlar ise çekirdek (kernel) tabanlı yaklaşımlardır. Çekirdek tabanlı yaklaşımlar çizgedeki yürüyüş, patika, ağaç gibi aynı altyapıları sayarak çizgenin elemanları arasındaki uzaklığı hesaplar ve her bir eleman için bir öznitelik vektörü oluşturur.

Löscher (Löscher et al., 2012), kesişim çizgesi (intersection graph) ve kesişim ağacı (intersection tree) tabanlı iki genel RDF çizge çekirdeği önermiştir. Vries kesişim ağacı çekirdeğini basitleştirmiş ve Weisfeiler-Lehman ağacı kullanarak etkin bir çekirdek geliştirmiştir (de Vries, 2013; de Vries and de Rooij, 2015).

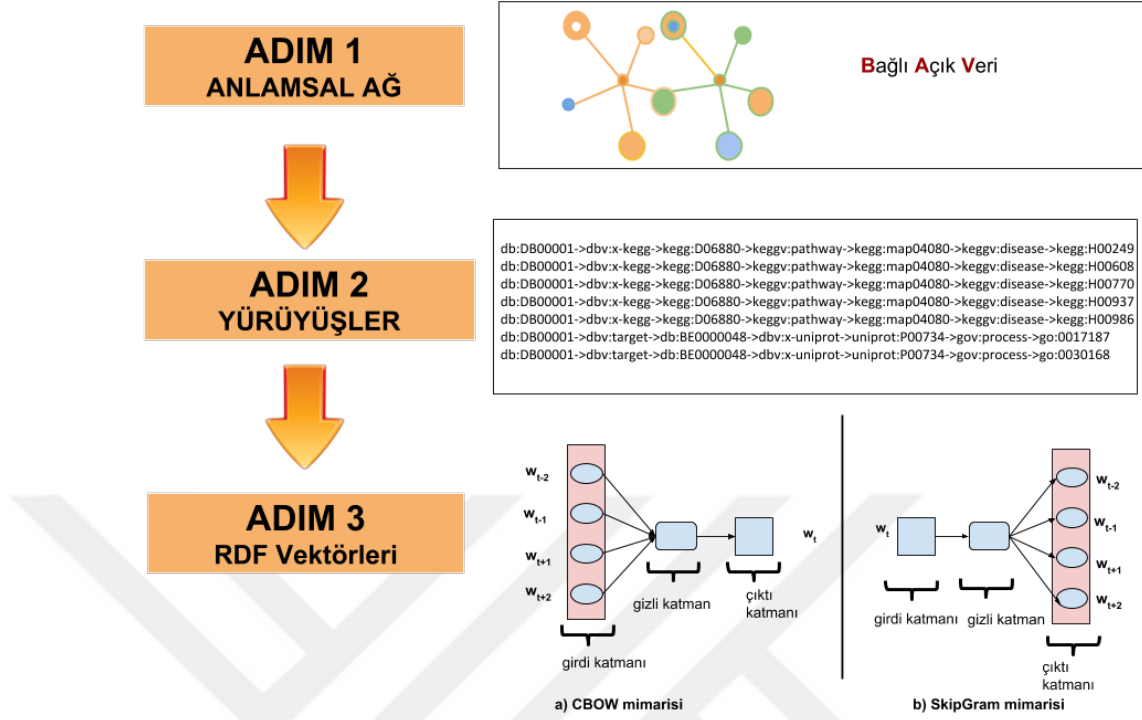
Ayrıca RDF2Vec (Ristoski and Paulheim, 2016), TransE (Bordes et al., 2013) ve TransD (Ji et al., 2015) gibi bilgi çizge gömmesi için başarılı yaklaşımlar tanıtılmıştır. Bu yaklaşımlar alt bölümlerde detaylı olarak açıklanmıştır.

- **RDF2Vec:** RDF2Vec (Ristoski and Paulheim, 2016), RDF ağ temsili için Word2Vec tabanlı bir yeni metodoloji önermektedir. RDF ağının çizge modeli üzerinde yürüyüşler yaparak veya belli yapıları çıkararak bunları sekansa çevirir. Bu sekansı Word2Vec ile eğiterek bağlama ilişkin bir vektöre yaklaştırır. Bu vektörler, tek sıcak kodlama (one-hot encoding) ile kodlanan ikili vektörlere göre tahmin gücü daha yüksek yoğun bir vektör gösterimi oluşturur. RDF2Vec yöntemi şu adımları izlemektedir; i) anlamsal ağ oluşturur, ii) oluşturulan ağ üzerinde yürüyüşler gerçekleştirir, daha sonra bu yürüyüşleri, <varlık->ilişki->varlık> şeklinde sekanslara dönüştürür ve son aşamada iii) bu sekansları CBOW veya Skip-Gram yapay sinir ağı modelleri ile eğiterek her bir varlık ve ilişki için bir vektör yaratır. (Şekil 3.1'e bakınız). RDF2Vec yürüyüşleri için üç başarılı ağırlıklandırma stratejisi önerilmiştir; Tekdüze, PageRank, PageRank Split.

- **Tekdüze:** Tekdüze ağırlık RDF2Vec tarafından kullanılan standart yaklaşımdır, burada her kenarın takip edilme olasılığı eşittir.
- **PageRank:** Bir düğümün Pagerank popülaritesini, tüm gelen kenarlarına (in-edge) atar.
- **PageRank Split:** Bir düğümün Pagerank popülaritesini, tüm gelen kenarlarına eşit olarak paylaştırır.

Metin Gömme - word2vec

Metinlerdeki kelimeleri temsil etmek için kelime torbası (bag of word) gibi alternatif doğal dil modelleri geliştirilmiştir. Kelime torbası, her kelime için bir ikili öznitelik vektörü kullanırken, her vektör indeksi bir kelimeyi gösterir. Bu yöntemler basit ve gürbüz (robust) olmasına rağmen, vektörlerin büyük boyutlu (high dimensional) ve seyrek (sparse) olması gibi sakıncaları barındırması nedeniyle performansları sınırlıdır. Bu sınırlandırmaları aşmak için kelimelerin düşük boyutta (low dimensional) ve dağıtık gömme temsiline (distributed embedding representation) olanak veren doğal dil modelleri önerilmiştir. Word2vec, iki katmanlı yapay sinir ağı modeli eğiterek hesaplama yönünden etkin bir dil modeli geliştirmiştir. Ayrıca Word2vec varlıkları otomatik olarak öğrenebilir ve kelimeler arasındaki ilişkileri vektörler üzerinde cebirsel işlemler kullanarak tahmin edebilir. Her kelime, kelimenin sabit büyüklükteki bir pencere içinde kalan bağlam kelimelerine göre olasılıklarının logunu maximize etmek için eğitilir. Sözlükteki her kelime, girdi ve çıktı olmak üzere iki vektör ile temsil edilir. Girdi vektörlerini öğrenmek ucuz iken, çıktı vektörlerini öğrenmek maliyetlidir. Etkin eğitim için hiyerarşik softmax ve negatif örnekleme (sampling) gibi yaklaşma



Şekil 3.1: RDF2Vec yönteminin adımları

(approximation) teknikleri geliştirilmiştir. Word2vec kelimelerin gömme temsilini oluşturmak için iki mimari sunmaktadır; CBOW ve Skip-Gram.

Continuous Bag-of-Words Model (CBOW)

CBOW modeli bir hedef kelimeyi çıktı ve hedef kelimenin etrafındaki diğer kelimeleri ise girdi olarak eğitim yapan iki katmanlı bir yapay sinir ağı modelidir. CBOW hedef kelimenin etrafındaki diğer kelimeler ile bulunma olasılığını (log probilitesini) artırmaya çalışır. Aşağıdaki objektif fonksiyonuna sahiptir:

$$\frac{1}{T} \sum_{t=1}^T \log p(w_t | w_{t-c} + \dots + w_{t+c}) \quad (3.8)$$

c bağlam (context) penceresinin büyüklüğüdür, p ise şöyle tanımlanır:

$$p(w_t | w_{t-c} + \dots + w_{t+c}) = \frac{\exp(\bar{v}^T v'_{w_t})}{\sum_{w=1}^W \exp(\bar{v}^T v'_w)} \quad (3.9)$$

v'_w, w kelimesinin çıkış vektörünü, V sözlükteki kelimelerin tümünü ve v ise bağlam kelimelerinin ortalama vektörünü göstermektedir.

Skip-Gram Model

CBOW kelimeyi verilen bağlama göre tahmin ederken, Skip-gram ise verilen kelimenin bağlamını tahmin eder. Skip-gram modeli, bir eğitim dokümanı veya cümlesi içindeki hedef kelimenin etrafındaki kelimeleri tahmin etmek için yararlı kelime temsillerini bulmaya çalışır. Skip-gram, verilen $w_1, w_2, w_3, \dots, w_T$ kelime sekansında c bağlam penceresi büyüklüğü olmak üzere, ortalama log olasılığını aşağıdaki şekilde artırmaya çalışır:

$$\frac{1}{T} \sum_{t=1}^T \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{t+j} | w_t) \quad (3.10)$$

p ise softmax fonksiyonu kullanılarak aşağıdaki şekilde tanımlanır:

$$p(w_{t+j} | w_t) = \frac{\exp(v'_{w_0} v_{w_1})}{\sum_{w=1}^W \exp(v'_w v_{w_1})} \quad (3.11)$$

v_w ve v'_w w kelimesinin girdi ve çıktı vektörlerini, V ise sözlükteki kelimelerini tümünü gösterir.

- **TransE:** TransE (Bordes et al., 2013) bilgi çizgesinde her varlık ve ilişki için ilişkiyi ön uç varlıktan (özne), arka uç (nesne) varlığa dönüşümü olarak gösteren düşük boyutlu bir vektör olarak kodlar. Bilgi çizgesinde bir üçlü, aynı zamanda bir ifadeyi gösterir, (h, r, t) için h gömme vektörüne r gömme vektörü eklendiğinde t gömme vektörüne yaklaşması beklenir ($h + r = t$). TransE, her üçlü için h ve t arasındaki farkı minimize edecek bir yapay sinir ağı modeli eğitir. TransE 1-1 ilişkileri modellemek için elverişlidir, ancak 1-N, N-1 ve N-M ilişkileri için yetersizdir.
- **TransD:** Bu yöntemde (Ji et al., 2015) her varlık ve ilişki iki vektör ile tanımlanır; gömme vektörü ve projeksiyon vektörü. Projeksiyon vektörü bir varlık vektörünü, ilişki vektörüne yansıtılmasını temsil eder ve haritalama matrislerini oluşturmak için kullanılır. Her varlık-iliski çiftinin benzersiz haritalama matrisi vardır. Bu sayede 1-N, N-1 ve N-M ilişkilerini ele alınabilmektedir. Ayrıca, TransD'nin vektör işlemleri ile değiştirilebilen matris-vektör geçiş işlemleri yoktur, böylelikle daha etkin çalışmaktadır.

4 LİTERATÜR ÖZETİ

Bu bölümde tezde önerilen yöntemle ilişkin ve önerilen yöntemin kullanımına yönelik iki durum çalışmasına ilişkin literatür özeti sunulacaktır.

4.1 Yönteme İlişkin Literatür Özeti

Protein-protein ve ilaç-ilac gibi biyolojik ağlar, tek tip düğüm içeren homojen ağlardır. Bu ağlarda bağ tahmini için sosyal ağlarda bağ tahmini yöntemlerinde kullanılan benzerlik ölçütlerinin, öznitelik olarak değerlendirilmesi mümkündür. Biyolojik ağlar, sosyal ağlardan derecelerin dağılımı (degree distribution), ağ geçişkenliği (network transitivity) gibi özellikler açısından farklılık göstermesine rağmen bağ tahmininde kullanılan özniteliklerin bu gibi ağlarda iyi birer kestirici olduğu önceki çalışmalarda belirtilmiştir (Lü and Zhou, 2011; Newman and Park, 2003). Hastalık-gen, hastalık-ilac gibi bazı diğer biyolojik ağlarda ise iki farklı tip düğüm bulunmaktadır ve yalnız farklı tip düğümler arasında bağlantılar mevcuttur. Bu ağlar, çift modlu (two mode)/iki parçalı (bipartite) bir yapıya sahiptir. Sosyal ağlardaki patika tabanlı yöntemler, bu ağlara doğal olarak uygulanabilir iken, ortak komşuluk yöntemlerinin uygulanabilmesi için değişikliğe ihtiyaç vardır. İki parçalı ağlarda bağlantılar sadece farklı tipteki düğümler arasında bulunduğundan bu iki tip düğümler arasında ortak komşu bulunmaz. Chen ve arkadaşları (Chen et al., 2005)'nın iki parçalı ağlarda ortak komşuluk yöntemlerini uygulayabilmek için bir düğümün komşuları $\Gamma(x)$ yerine o düğümün komşularının ortak komşularını $\bigcap_{c \in \Gamma(x)} \Gamma(c)$ kullanmayı önermişlerdir. Önerilen benzerlik ölçütleri her zaman simetrik değildir ($puan(s, t) \neq puan(t, s)$). Simetrik bir ölçüt hesaplamak için Davis ve arkadaşları (Davis et al., 2011; Davis et al., 2012) iki düğüm arasında 3 uzunluğundaki patikalara dayalı bir komşuluk yaklaşımı önermişlerdir. Bunun için kaynak (s) ve hedef (t) düğümleri arasındaki 3 uzunluğundaki benzersiz bütün patikaların sayısı, ortak komşu sayısı olarak alınmıştır. Tek modlu (single mode) ağlarda bir s düğümü ile t düğümü arasındaki 2-uzunluktaki patikalar, ortak komşuları içerir. Çift modlu ağlarda ise her bir ortak komşu düğümü tek sayılı uzunluktaki patikalarda yer alır. Bu nedenle bir s ve t düğümü arasındaki komşu sayısı için 3 uzunluktaki yolların sayısı kullanılır. Benzer bir şekilde Jaccard kat sayısını, s'den t'ye 3-uzunluklu benzersiz yolların sayısını s veya t'den başlayarak üç uzunluktaki benzersiz tüm yolların toplam sayısına bölümü olarak formüle edilmiştir. Adamic/Adar benzerliği için ise log terimi olarak $\log(|\Gamma(s) \cup \Gamma(t)|)$ kullanılmıştır.

Heterojen ağlarda her farklı tür bilgi, doğal olarak bağ tahmin etmede farklı

miktarda kanıt sunar. Sosyal ağ analizinde kullanılan yöntemlerin, her düğüme ve bağa eşit muamele ettiklerinden, heterojen ağlarda doğrudan uygulanabilirliği kısıtlıdır. Bu özellikle frekans ve dağılımı farklı olan farklı tipteki düğümler ve bağlar için sorun teşkil eder. Bu gibi sorunlar göz ardı edildiğinde dahi tüm bağların eşit ele alınması tahmin edilen bağ hakkında ek bir bilgi sunmayabilir.

Davis ve arkadaşları (Davis et al., 2011) heterojen ağlarda farklı bağlar için uygun ağırlıklandırma şemasına sahip bir bağ tahmini yöntemi olan MRLP'yi tanıtmışlardır. Ağırlıkların, geleneksel olarak üçlü sayımı (triad census) olarak ve daha yakın zamanda 3 düğümlü çizgecik (graphlet) sayımı olarak ta adlandırıldığı yöntem, ağdaki benzersiz 3 düğümlü alt yapılarının görülme sıklığı ile tanımlanır. Üçlü sayımı, heterojen ağlara neredeyse aynı şekilde uygulanır. Tek fark benzersiz yapıların sayısıdır. Üçlü sayısı her bir yapının olasılığını verir, bu her kısmi üçlünün bir bağ türü ile kapanma olasılığı olarak ta nitelendirilebilir. Verilen bir üç düğüm (s,n,t) ve bir bağ tipi (x) için x bağ tipinin (s,n,t) üçlü yapıları içinde bulunma olasılığı hesaplanır. Bunun için (s,n,t) örüntüsüne sahip yapıların sayısı, bu örüntüler içinden s ve t arasında x tipli bir bağın görülme sayısına bölünür. Bu olasılık, x tipli bağın potansiyel yokluğu durumu dışında gözlenen örüntünün doğru olduğunu varsayar. Bu değer, bir ağırlıklandırma planına çevrilerek heterojen ağlar için ağırlıklı, olasılıksal bir Adamic/Adar ölçütü elde edilir. MRLP'nin, diğer klasik bağ tahmin yöntemlerine göre heterojen ağlarda başarılı sonuçlar elde ettiği bildirilmiştir. Fakat yöntem, bütün üçlü yapıları bulmayı gerektirdiğinden büyük veriler için ideal bir çözüm değildir.

Yang ve arkadaşları (Yang et al., 2012), Kempe ve arkadaşlarının (Kempe et al., 2003) sosyal ağlarda etki maksimizasyonu (influence maximization) problemi için önerdiği etki yayılım (influence propagation) yaklaşımını, heterojen ağlara uyarlayarak ağırlıklı bir akış hesaplaması yöntemi önermişlerdir. Önerilen yöntemde, bir bağın diğer bir bağa göre koşullu olasılığı $P(i | j)$, bağlar arasındaki korelasyonuna tekabül eder. Bu koşullu olasılık, i ve j tipindeki bağlar ile oluşmuş ağlar hizalandığında, aynı düğümler arasında eşleşen bağ oranına denktir. Etki yayılımı yaklaşımı, bu olasılık değerleri ile ağırlıklandırılarak genişletilmiştir.

MRPF yaklaşımı (Johnson et al., 2013), orijinal PropFlow yaklaşımını ve Katz yaklaşımını birleştirerek çoklu-ilişkisel veri ağına uyarlamıştır. Bu algoritmada u ve v düğümleri arasındaki benzerlik hesaplanırken, cezalandırma için kullanılan katsayısı iki düğüm arasındaki uzaklığın yanında, düğüm zıplama (hop) sayısına bağlıdır. Genetik (hastalık-gen ilişkisi) ve protein-protein ağları gibi bazı biyomedikal ağlar için nispeten başarılı sonuçlar rapor edilmiştir. Yöntemin zaman karmaşıklığı, $O(|V|.|E|)$ olarak verilmiştir. Küçük ve ortalama veri setleri için ideal

olmasına rağmen büyük veri setleri için elverişli değildir.

Birçok çalışma bağ tahmini için düğümler arasındaki farklı kavram şeması patikalarını (concept schema path), öznitelik çıkarmak için kullanmaktadır (Zhang et al., 2017; Sun et al., 2011; Cao et al., 2014; Lao and Cohen, 2010). (Yu et al., 2012). Şema patikaları, iki düğüm arasındaki ilişkilerin, hangi tip bağlardan oluştuğunu tanımlarlar ve ayrıca üst patika (meta path) olarak ta adlandırılırlar (Zhang et al., 2014). Sun ve arkadaşlarının PathPredict yaklaşımı (Sun et al., 2011), ortak yazar ilişkisini kestirmek için üst patikalardan çıkarılan öznitelikleri kullanarak bir lojistik sınıflandırıcı eğitir. Üst patika tabanlı yaklaşımların en büyük kısıtlarından biri, iki düğüm arasında çok fazla sayıda üst patika bulunmasıdır. PRA yaklaşımı (Lao and Cohen, 2010), önemli patikaları seçmek için $< v, u >$ düğümleri arasında arama gerçekleştirerek en sık görülen k tane üst patikayı bulur. Yöntem, daha sonra bu üst patikalardaki bulunan veri patikası (data path) sayısını, öznitelik vektörüne çevirerek bir lojistik sınıflandırıcı eğitir.

Son zamanlarda popüler olan vektör tabanlı yöntemlerden DeepWalk (Perozzi et al., 2014), RDF2Vec'e benzer bir şekilde bilgi çizgesini rastgele dolaşarak, elde edilen patika sekanslarının iki katmanlı özel bir yapay sinir ağı ile eğitilmesine dayanır. DeepWalk, RDF2Vec yaklaşımından farklı olarak yalnız düğümlerden meydana gelen patikaları kullanırken, RDF2Vec düğüm ve bağ bilgilerini birlikte içeren patika sekanslarını eğitim için kullanır. Alshahrani ve arkadaşları (Alshahrani et al., 2017) çalışmalarında DeepWalk yaklaşımını genişleterek, ağdaki bağları patika sekanslarına dahil etmişlerdir. Ayrıca bilgi çizgesinde mantıksal çıkarsama yaparak, yeni ilişkiler edinerek, bilgi çizgesini zenginleştirmişlerdir. Çıkarsama yapıldıktan sonra ilaç-hastalık ilişkisi dahil birçok görevde DeepWalk'ın tahmin başarısını artırmışlardır.

4.2 İlaç-İlaç Etkileşimi Tahmini Durum Çalışmasına İlişkin Literatür Özeti

Araştırmacılar ilaç-ilaç etkileşim tahmininde birçok yaklaşım ve yöntem kullanılmaktadır. Daha önceki çalışmalar literatür araştırmasını farklı ana başlıklar altında toplamaya çalışmışlardır; (benzerlik tabanlı ve sınıflandırma tabanlı, benzerlik tabanlı ve öznitelik tabanlı) (Zhang et al., 2017; Abdelaziz et al., 2017). Bu literatür sınıflandırmaları birçok açıdan eksik kalmakta ve yaklaşımlar arasındaki ayrımı yeterince açıklayamamaktadır. Bu nedenle literatürdeki çalışmalar, Tavsiye Sistemlerinin (Recommender System) literatür sınıflandırması baz alınarak hafıza tabanlı (memory based) ve model tabanlı (model

based) olarak sınıflandırılmıştır (Bobadilla et al., 2013). Hafıza tabanlı yaklaşım, benzerlik skorlarının belleğe yüklenip bu verileri kullanarak en benzer komşuları önermeye dayanır. Model tabanlı yaklaşım ise verilerden bir model üretilip bu modeli kullanarak önerme yapmaktadır.

Hafıza tabanlı yaklaşımlar, aday bir ilaç çiftini bilinen ilaç çiftleri ile karşılaştırarak bunun etkileşip etkileşmediğini tahmin eder. Bir ilaç çiftine çok benzer bilinen etkileşen ilaçlar bulmak, bu aday ilaçlar arasında bir etkileşim olduğunu destekler. Bu yöntemlerden bazıları şunlardır:

Ferdousi ve arkadaşları (Ferdousi et al., 2017) ilaç-ilaç etkileşimi tahmini yapmak için Drugbank veri tabanından elde ettikleri taşıyıcı, transporter, enzim ve hedef (carriers, transporters, enzymes and targets - CTET) bilgilerini kullanmışlardır. Bu çalışmada onaylanmış 2189 ilaç, 45.530 bilinen ilaç etkileşimi ve 2.394.766 bilinmeyen ilaç çifti incelenmiştir. DDI'ları belirlemek için her bir ilaç için bağlantılı tüm CTET'leri toplayıp ikili vektörler oluşturmuşlardır. Daha sonra bu birleşik vektörler ile beraber birçok benzerlik yöntemini kullanarak DDI'ları tespit etmeyi amaçlamışlardır. Bu benzerlik yöntemlerinden iç çarpım tabanlı benzerlik (inner product-based similarity measures - IPSMs) kullanarak 250.000'den fazla potansiyel yeni DDI tahmin etmişlerdir. Ferdousi ve arkadaşları sınıflandırıcıyı eğitmek için toplam 2004 öznitelik kullanmışlardır.

Villar ve arkadaşları (Vilar et al., 2014) ilaçların moleküler yapısal benzerliklerine dayalı bir model geliştirmişlerdir. Bu çalışmada veri seti olarak DrugBank v.3.0'da belirtilen toplam 6624 ilaç ve 9454 DDI kullanılmıştır ve buradan çıkarılan etkileşimler altın standart olarak değerlendirilmiştir. Drugbank ilaç-ilaç etkileşimleri ve moleküler parmak izleri modellemenin kombinasyonu ile DDI benzerlik modeli oluşturulmuştur. İlaç çiftlerinin benzerliği, Tanimoto katsayısı ile moleküler parmak izlerine dayalı olarak hesaplanmıştır. Metot sonuç olarak 0,68 duyarlılık (sensitivity), 0,96 özgüllük (specificity) ve 0,26 kesinlik (precision) başarıları göstermiştir ve 58.403 yeni DDI tahmin edilmiştir.

Model tabanlı yaklaşımlar, daha çok ilaç benzerlik ölçütlerini öznitelik olarak kullanmışlardır (Celebi et al., 2015). Bunlardan Gottlieb ve ark. (Gottlieb et al., 2012) farklı ilaç benzerlik ölçütlerini kullanarak yeni tahmin modeli olan INDI çerçevesini geliştirmişlerdir. INDI, 7 farklı ilaç benzerliği kullanarak, bunların bilinen ilaç ilaç etkileşimleri ile maksimum benzerliğini hesaplayarak 49 öznitelik oluşturmuş, bunları kullanarak bir lojistik sınıflandırıcı eğitmiştir. Cheng ve arkadaşları (Cheng and Zhao, 2014) ilaçların fenotipik, terapötik, yapısal ve genomik benzerliklerini kullanarak ilaç etkileşimlerini tahmin etmek için

HNAI çerçevesi sunmuşlardır. Cami ve ark. (Cami et al., 2013) ilaçlar arasındaki farmakolojik ve çizge/ağ özniteliklerini çıkararak bir lojistik sınıflandırıcı eğitmişlerdir. Zhang ve ark. (Zhang et al., 2015) etiket yayılım (label propagation) yöntemini ilaç kimyasal altyapısı, ilaç yan etkisi ve ilaç beklenmeyen yan etkileri üzerinde uygulamışlardır. Li et al. (Li et al., 2015) ilaçların kombinasyon etkisini tahmin etmek için ilaç moleküler benzerliğini ve ilaç fenotipik (yan etki) benzerliğini birleştiren bir Bayes ağı geliştirmişlerdir. Zhang ve ark. (Zhang et al., 2017) çeşitli ilaç verilerini toplamışlar ve böylece kimyasal, biyolojik, fenotipik ve ağ verilerini bütünleştirerek ilaç-ilac etkileşimlerini öngörmüşlerdir.

Makine öğrenmesi yöntemlerine girdi olarak öznitelik vektörleri alan çalışmalar da vardır. Luo ve arkadaşları (Luo et al., 2014) moleküler altyapı kullanarak öznitelik vektör bazlı bir metod önermişlerdir. Moleküler yapısı verilen bir ilacın 611 insan proteinine karşı her insan protein hedefinin bağlanma (docking) skoru hesaplanmıştır. Bu da 611 boyutunda bir docking vektörü oluşturmuştur. Daha sonra bir ilaç çifti için bu vektörlerin birleşim yada konsantrasyonu için toplama ve farklarının mutlak değeri 1222 boyutunda bir öznitelik vektörü oluşturmuştur. Daha sonra bu öznitelik vektörü ile lojistik model eğitilerek 2515 ilaç molekülü için potansiyel ilaç-ilac etkileşimleri bulunmuştur.

Abdelaziz ve arkadaşları (Abdelaziz et al., 2017) DDI'ları tahmin etmek için benzerlik tabanlı bir çerçeve olan Tiresias'i sunmuşlardır. Farmakolojik benzerlikleri, ilacın metin ve RDF gömme vektörü tabanlı benzerlik bilgilerini kullanarak elde edilen benzerliklerden türetilen 1014 öz nitelik kullanılmıştır. Bu öznitelikler her ayrı benzerlik için bir ilaç çiftine en çok benzeyen bilinen etkileşen ilaç çiftine benzerlik değerini gösterir. RDF Veri ağı olarak DrugBank, UMLS, DailyMed, Uniprot ve CTD veri setlerinden oluşan birleşik bir bilgi tabanı oluşturulmuştur. Böylece bu entegre bilgi ağı, enzimler, kimyasal yapılar ve yollar gibi varlıklar ile ilaç özellikleri ve ilişkilerini içerir. Bu bilgi ağını ilaçlar arasındaki global benzerlik ölçütünü hesaplamak için kullanmışlardır. Değerlendirme ölçütü olarak kesinlik, duyarlılık, F-puanı ve AUPR kullanmışlardır.

Hameed ve arkadaşları (Hameed et al., 2017) DDI çıkarımında ikili sınıflandırma yöntemleri için gerekli olan potansiyel negatif DDI örneklerini tahmin etmek amacıyla Büyüyen Özdüzenleyici Haritalar (Growing Self Organizing Map - GSOM) kümelenmesine dayalı Pozitif Etiketsiz Öğrenme (Positive-Unlabeled Learning - PUL) yaklaşımını sunmuşlardır. Drugbank'tan elde ettikleri 6036 DDI içerisinde 589 ilaç çiftinin birbiriyle etkileşmediğini öngörmüşlerdir. Bunları ikili sınıflandırma yönteminde negatif sınıf olarak ele almışlardır. Önerilen yaklaşım 5-kat çapraz doğrulama kullanarak sırasıyla 0,97 kesinlik, 0,98 duyarlılık ve 0,97 F

puan değerlerini elde etmiştir.

4.3 İlaç Endikasyon Tahmini Durum Çalışmasına İlişkin Literatür Özeti

Bilinen ilaçların yeni endikasyonları tahmin etmek için farklı yöntemler önerilmiştir. Önceki büyük ölçekli yeni ilaç endikasyonu kestirimi için yapılan çalışmalar gen ifadesi profilinin eşlenmesi (Lamb et al., 2006; Hu and Agarwal, 2009) ve bilinen ilaç-hastalık ilişkilerine dayalı benzer ilaç veya hastalıkların önerilmesine (Chiang and Butte, 2009) odaklanmıştır. Chiang ve Battle (Chiang and Butte, 2009) ağ tabanlı olan *ilişki yüzünden suçlama* (guilt-by-association) yaklaşımı ile potansiyel ilaç-hastalık ilişkilerini tahmin etmişlerdir. Bu yaklaşımın altındaki temel varsayım, iki hastalık aynı tedavi profiline sahip ise bu iki hastalıktan biri için kullanılan ilacın diğer hastalık için de kullanılmasıdır. Ancak bu yöntem, birçok hastalığın tedavisinde kullanılan ilaçlar veya çok farklı tedavi seçeneklerine sahip hastalıklar için bir miktar yanılmaya (bias) sebep olmaktadır.

Wu ve arkadaşları (Wu et al., 2013), ilaç-hedef ve hastalık-gen ilişkilerini kullanarak hastalık ve ilaç bilgilerinden oluşan ağırlıklı bir heterojen ağ oluşturmuşlardır. Bu ağ üzerinde kümeleme yaparak ilaç adaylarını belirlemişlerdir. Wang ve arkadaşları (Wang et al., 2013) ise benzer şekilde oluşturdukları heterojen ağ modeli üzerinde özyinelemeli bir algoritma geliştirmişlerdir. Bu algoritma, bir hastalık için potansiyel ilaç adaylarını derecelendirmektedir. Martinez ve arkadaşları (Martínez et al., 2015) DrugNet adını verdikleri ağ tabanlı önceliklendirme yöntemi ile hastalık, ilaç ve hedef bilgilerinden oluşan bütünleştirilmiş bilgiyi ilaç-hastalık ve hastalık-ilac önceliklendirme için kullanmışlardır. Luo ve çalışma arkadaşları (Luo et al., 2016) ise ilaç ve hastalık benzerliklerini, bilinen ilaç-hastalık ilişki ağı ile birleştirerek Bi-Random Walk(BiRW) algoritmasını önermiş ve ilaçların yeni endikasyonlarını tahmin etmek için uygulamışlardır.

Makine öğrenmesinin, diğer ilaç endikasyon tahmini yöntemlerine göre en büyük avantajı, farklı ilaç ve hastalık özelliklerinin birleştirilerek tahmin modelinde kullanılabilir olmasıdır. Ayrıca bir ilacın endikasyonu olmasını sağlayan belirleyici özelliklerin bulunmasını sağlayabilir. Makine öğrenmesi temelli daha önceki ilaç endikasyon tahmin çalışmaları daha çok benzerlik ölçütlerini kullanmışlardır. Bunlardan en çok atıf yapılan yöntemlerden biri, PREDICT yöntemidir (Gottlieb et al., 2011). Bu yöntemde 5 ilaç-ilac benzerlik ve 2 hastalık-hastalık benzerlik ölçütü kullanılarak, ilaç-hastalık ilişkilerini öngörmek için lojistik sınıflandırıcı

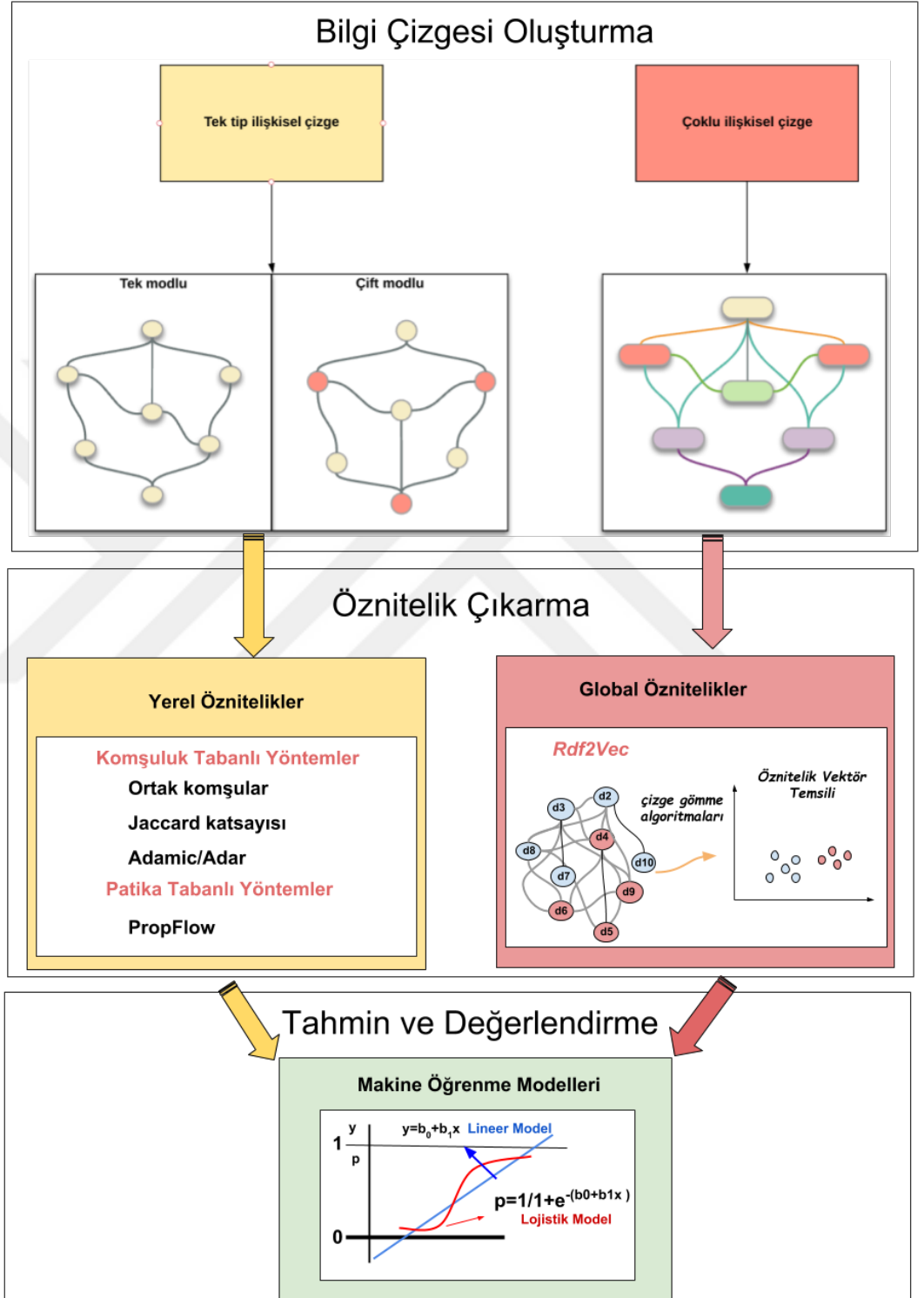
eđitilmiřtir. Yapılan deneylerde bu yöntemin 0.92 AUC başarısına yaklařtıđı bildirilmiřtir. Zhang ve alıřma arkadaşları ise farklı kaynaklardan elde edilmiř ila verileri ile ila benzerlik skoru hesaplayarak bir hastalık iin en yakın k ilacı önermiřlerdir (Zhang et al., 2013). Guney ise bu alıřmayı arařtırmacıların tekrarlayabilmesi iin bir yazılım aracı geliřtirerek halka aık olarak sunmuřtur (Guney, 2016).



5 ÖNERİLEN YÖNTEM

Bağ tahmini için farklı bilgilere dayanan öznitelikler veya modeller mevcuttur, tahmin performansını güçlendirmek için bunların birleştirilmesi, doğal çözümlerden biridir. Topluluk Öğrenme (ensemble learning), modelin daha iyi bir genelleme yapmasını ve yüksek doğruluğa sahip olmasını sağlamak için makine öğrenmesi modellerini bir araya getiren bir tekniktir (Polikar, 2006). Bir topluluk öğrenme sistemi, iki bileşene sahiptir; temel kestiriciler (predictors) ve birleştirme kuralları. Topluluk sistemlerde her kaynak için elde edilen öznitelikler veya modeller, temel kestirici olarak tanımlanır. Bu temel kestiricileri birleştirmek için ağırlıklı ortalama kuralı ve sınıflandırıcı kuralı olmak üzere iki popüler kural mevcuttur. Ağırlıklı ortalama birleştirme kuralı, temel kestiriciler için ağırlıklı ortalama değerini hesaplayarak bir etiket değerine dönüştürür. Çoğu zaman bu kestiriciler için ağırlıklar bilinmediğinden bu ağırlıkların genetik algoritma gibi yöntemler ile öğrenilmesi gerekir (Zhang et al., 2017). Sınıflandırıcı birleştirici kuralı ise temel kestiricileri bir etikete eşleyecek bir sınıflandırıcı fonksiyonunu öğrenmeye çalışır.

Şekil 5.1’de genel bakışı verilen önerilen yöntem, tahmin edilmesi istenen bağ tipi için iki ayrı öznitelik grubu üzerinde topluluk öğrenme gerçekleştirir. Bunun için iki farklı bilgi çizgesi oluşturur; tek tip ilişkisel çizge, çoklu ilişkisel çizge. Tek tip ilişkisel çizge, yalnız tahmin edilecek bağ tipini içeren bir çizgedir. Bu çizgede bağlar tek tip düğümler arasında tanımlanmışsa tek modlu çizge, iki farklı tipteki düğümler arasında tanımlanmışsa çift modlu (iki parçalı) çizge olarak adlandırılır. Bu gibi çizgelerde bağ tahmini için çizge yapısını karakterize eden ağ yakınlık yöntemleri kullanılmaktadır. Sosyal ağlar için kullanılan geleneksel komşuluk ve patika tabanlı yöntemler, tek modlu çizgelere uygulanmıştır. Bu yöntemlerden Ortak Komşu Sayısı, Jaccard, Adamic/Adar ve PropFlow yakınlık ölçütleri, öznitelik vektörü olarak kullanılmıştır. Çift modlu çizgelerde ise ortak komşu olmadığından, alternatif komşuluk tanımlamaları değerlendirilmiş ve (Davis et al., 2011; Davis et al., 2012) tarafından önerilen komşuluk yöntemleri, yakınlık ölçütleri olarak kullanılmıştır. Bu tanımlamada, ortak komşu sayısı yerine 3 uzunluğundaki patikaların sayısı kullanılmıştır.



Şekil 5.1: Önerilen yöneme genel bakış

Önerilen yöntemde iki parçalı ağlar için komşuluk yakınlıkları aşağıdaki şekilde tanımlanmıştır.

$$\text{Ortak Komşular}(\mathbf{x}, \mathbf{y}) = |Path_{\langle s, n_1, n_2, t \rangle}| \quad (5.1)$$

$$\text{Jaccard}(\mathbf{x}, \mathbf{y}) = \frac{|Path_{\langle s, n_1, n_2, t \rangle}|}{|Path_s^3|} \quad (5.2)$$

$$\text{Adamic/Adar}(\mathbf{x}, \mathbf{y}) = \sum_{n_1, n_2 \in Path_{\langle s, n_1, n_2, t \rangle}} \frac{1}{\log |\Gamma(n_1) \cap \Gamma(n_2)|} \quad (5.3)$$

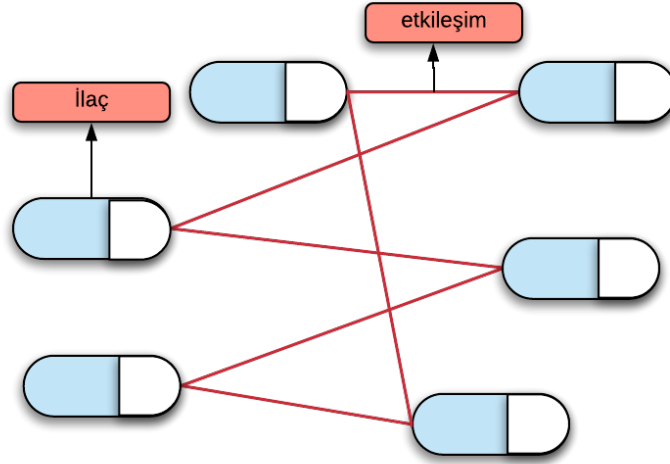
$Path_{\langle s, n_1, n_2, t \rangle}$, s ve t düğümleri arasında, ara düğümleri n_1 ve n_2 olan üç uzunluklu patikaları göstermektedir. $Path_s^3$ ise s düğümünden başlayarak 3 uzunluklu olası bütün patikaları temsil etmektedir.

Bunun yanısıra tahmin edilmek istenen bağ tipi dışında diğer tüm bağ tiplerini içeren bilgi çizgesi kullanılarak global öznitelikler çıkarılmıştır. Bilgi çizgesinin global özniteliklerini oluşturmak için vektör tabanlı yöntemlerden RDF2Vec uygulanmıştır. RDF2Vec, bilgi çizgesindeki anlamsal yapıyı koruyarak her düğüm ve bağ için bir vektör temsili oluşturur. Önerilen yöntemde RDF Çizge Gömme vektör temsili kullanarak bağ tahmini yapmak için ilk adım olarak RDF formatında veriler birbirlerine bağlanarak bilgi çizgesi oluşturulur. İkinci adımda düğümlerin RDF Çizge Gömme öznitelik vektörü çıkarılır. Son adımda, bir düğüm çifti için çıkarılan öznitelikler birleştirilerek (art arda sıralanarak) tahmin modelleri eğitilir.

Bu iki öznitelik grubundaki öznitelikler birleştirilerek modellerin tahmin gücünün artırılması amaçlanmıştır. Bunun için ise birleştirme sınıflandırıcıları kullanılmıştır.

Tez kapsamında incelenen bilgi ağları şunlardır:

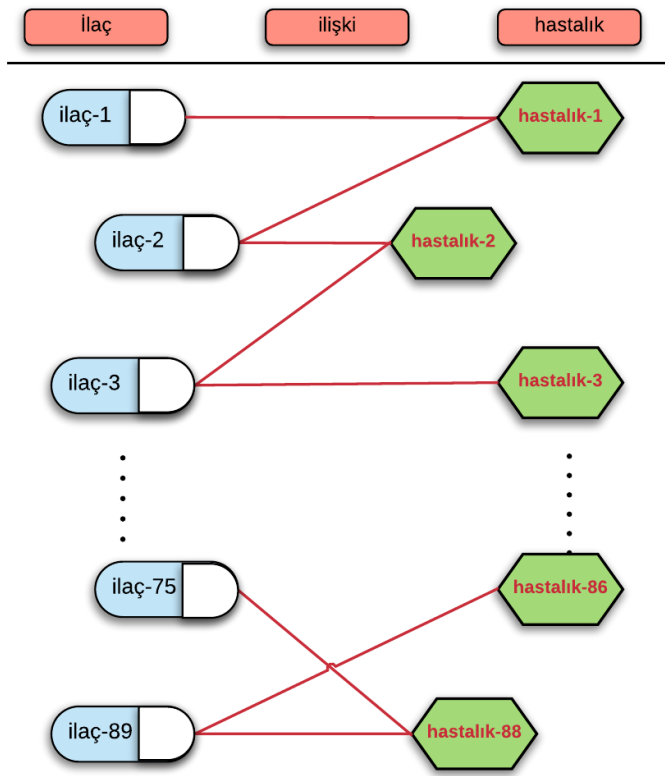
- **İlaç-ilaç etkileşim ağı:** İlaç-İlaç etkileşimleri, çoklu ilaç alımlarında bir ilacın, diğer bir ilacın farmakokinetiğini veya farmakodinamiğini değiştirerek etkisini düşürmesi veya artırması sonucu oluşur. Bazı istenen



Şekil 5.2: Homojen ilaç-ilaç etkileşim ağı

ilaç-ilaç etkileşimleri klinik yararları için araştırılırken, istenmeyen ilaç-ilaç etkileşimlerinin çoğunlukla zararlı etkileri olmaktadır. Bu etkileşimleri, bir ağ/çizge olarak modellediğimizde düğümler ilaçları, ayrıtlar ise etkileşimleri gösterir ve homojen bir ağ oluşturur (Şekil 5.2).

- **İlaç-hastalık ilişki ağı:** İlaç-hastalık ilişkisi, bir çok bağlamda kullanılmaktadır. Bir ilacın bir hastalığı iyileştirme ilişkisine endikasyon, ilacın bir hastalık için kullanılmaması durumuna kontra-endikasyon ve ilacın kullanılması durumunda yol açtığı etkiye yan etki denir. Bu çalışmada odaklanan ilişki türü, ilacın hastalığı iyileştirme (endikasyon) ilişkisidir. Şekil 5.3’de görüldüğü üzere ilaç-hastalık ilişkileri iki parçalı (bipartite) çizge yapısındadır ve heterojen bir ağ yapısına sahiptir.



Şekil 5.3: İki parçalı ilaç-hastalık etkileşim ağı

6 ÖNERİLEN YÖNTEMİN KULLANIMINA YÖNELİK DURUM ÇALIŞMALARI

Bu bölümde, önerilen yöntemin işlevselliğinin test edilmesi için geliştirilen iki durum çalışması ele alınacaktır. İlk olarak, ilaç-ilaç etkileşimlerinin tahmini için bir durum çalışması incelenecektir. Daha sonra ise, yeni ilaç endikasyonu tahmini için durum çalışması ele alınacaktır.

6.1 İlaç-İlaç etkileşimlerinin tahmini

İstenmeyen ilaç etkileşimleri (Adverse Drug Reaction- ADR) halk sağlığına yönelik önde gelen tehditlerden biridir. Lazarou ve arkadaşları (Lazarou et al., 1998) tarafından yapılan çalışmada Amerika Birleşik Devletleri'nde hastanelerde yatan hastaların % 6,7'si ciddi istenmeyen ilaç etkileşimlerine maruz kalmakta ve bu hastaların % 0,32 'inde bu etkileşimlerin ölümcül olduğu tahmin edilmektedir. A.B.D'de 2014'yılında 807.270 ciddi istenmeyen ilaç etkileşimi ve buna bağlı 123.927 can kaybı vakası raporlanmıştır ³.İstenmeyen ilaç etkileşimleri, hastaneye tekrar yatış, morbidite, mortalite ve sağlık hizmetleri kullanımı maliyetleri açısından sağlıkta önemli bir yük oluşturmaktadır. Bu istenmeyen ilaç etkileşimlerinin önemli bir kısmı bir ilacın diğer ilaçlar ile birlikte alınmasıyla ortaya çıkan ilaç-ilaç etkileşimlerinden (drug-drug interaction - DDI) kaynaklanmaktadır (Bates et al., 1995). Yaşlı hastalar ve kanser hastaları gibi bazı hasta gruplarının çoklu ilaç kullanma olasılıkları daha yüksektir ve bu da DDI riskini artırmaktadır (Riechelmann et al., 2007; Doubova Dubova et al., 2007). İlacın satışa sunulmasından sonra yapılan ilaç izleme ve öncesindeki klinik değerlendirme süreçleri doğru ve eksiz bir bilgi sağlamaz ve ilaç hastaya verilmeden olası ilaç etkileşimlerine karşı hastayı ve doktoru uyaramaz (Percha and Altman, 2013). Potansiyel ilaç-ilaç etkileşimlerini tahmin etmek, beklenmedik ilaç etkileşimlerini ve ilaç geliştirme maliyetlerini azaltmaya ve ilaç geliştirme süreçlerini optimize etmeye yardımcı olur. Bu sebeple ilaç-ilaç etkileşimlerini otomatik tespit etmek için yeni yöntemlere ihtiyaç vardır.

Son zamanlarda giderek artan sayıda biyolojik veri ve bilgi tabanı Anlamsal Web teknolojileri üzerine bina edilmekte ve bilgi getirimi, veri entegrasyonu ve federasyonu için bilgi çizgesi kullanılmaktadır (Katayama et al., 2014). Bu doğrultuda birçok veri tabanı, verilerini Anlamsal Web teknolojilerini kullanarak, bir çizge veri modeli olan Bağlı Açık Veri (LOD) şeklinde sunmaya başlamıştır

³FDA ADE Web sayfası: <https://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/ucm070461.htm>

(Smith et al., 2005; Jupp et al., 2014). Bilgi çizgeleri (Knowledge Graph - KG) verilerin tanımlanması için güçlü bir veri modeli sağlarken aynı zamanda anlamlı bilgilerin çıkarılması için altta yatan çizge yapısının kullanılmasını mümkün kılar. Bu çalışmada, LOD kullanılarak Bilgi Çizge Gömmesi (Knowledge Graph Embedding) yöntemleri ile ilaç öznitelik vektörü temsilleri çıkarılmış ve bunlar potansiyel ilaç-ilac etkileşimlerini tahmin etmek için kullanılmıştır. Bu çalışma aynı zamanda farklı gömme yöntemlerini ve farklı bağlı veri kaynaklarının DDI tahminine etkisini araştırmayı amaçlamıştır.

Bu çalışmanın makine öğrenmesi temelli önceki çalışmalardan farklı olarak ortaya koyduğu yenilikler maddeler halinde sıralanmıştır; i) Var olan yöntemlerin çoğu ilacın hedef, yan etki, kimyasal yapısı ve endikasyonu gibi özellikleri temel alan benzerlikleri kullanır. Her bir benzerlik, sınıflandırıcı için bir öznitelik olarak kullanılır fakat bu özelliklerin sayısı kısıtlıdır. Önerilen yaklaşımda bir ilaç, sınıflandırıcının tahmin gücünü artıracak büyüklükte olan bir yoğun (dense) öznitelik vektörü ile karakterize edilir. Bu vektörün ilaca ilişkin başka makine öğrenmesi işlerinde (i.e. ilaç-hedef, ilaç-advers etki) kullanılması mümkündür. ii) Bu çalışmada kullanılan yöntem, hakkında eksik veya yetersiz bilgi olan ilaçlar için de tahmin yapabilmektedir. Bağlı veri ağı kullanılması sayesinde ilacın ağı içinde bulunması, özniteliklerin makine öğrenmesinde kullanılmak üzere çıkarılması için yeterli olmaktadır. Bu sayede birçok ilaç ve böylelikle ilaç-ilac etkileşimi eğitim setinde kullanılabilir. Aksine benzerlik temelli yaklaşımlarda ise her ilaç çifti için farklı benzerliklerin hesaplanması, ilaçlara ilişkin yetersiz bilgi yüzünden mümkün olmayabilmektedir. iii) DDI tahminini daha doğru değerlendirmek için ilaç çiftleri üç test sınıfına ayrılarak test senaryoları üretilmiştir. Önceki çalışmalar geleneksel çapraz doğrulama yaparken, bilinen etkileşimleri hiç olmayan ilaçların durumlarını dikkate almamaktadırlar.

DDI tahmini için bilgi çizge gömme temelli ilaç vektörleri kullanarak üç farklı sınıflandırıcı; Lojistik Regresyon (LR), Naive Bayes (NB) ve Rastgele Orman (RO) eğitilmiştir. Bilgi çizgelerinden öznitelik vektörlerini çıkarmak için son teknoloji başarılı bilgi çizgesi gömme yöntemleri kullanılmıştır; RDF2Vec (Ristoski and Paulheim, 2016), TransE (Bordes et al., 2013) ve TransD (Ji et al., 2015). Sonuçlar ilaç vektör temsillerin tahmin performansının, var olan farmakolojik benzerlik temelli metotlar ile karşılaştırılabilir olduğunu göstermektedir. Önerilen yöntem Drugbank veri seti kullanılarak 10 kat çapraz doğrulama ile 0.93 AUC değerine ve 0.86 F-puanına ulaşmaktadır.

Çizelge 6.1: Gömme temsilini öğrenmek için kullanılacak her bir bilgi kaynağına ilişkin üçlü, düğüm ve bağ tipi sayısı

Veri seti	Üçlü sayısı	Düğüm sayısı	Bağ tipi sayısı
Drugbank	2,588,933	574,152	76
KEGG	12,347,534	4,035,716	260
PharmGKB	1,050,530	162,502	36

6.1.1 Veri seti

Drugbank v5.0, 2.551 ilaçtan oluşan 288.856 farklı ilaç-ilaç çiftini içermektedir. 2.251 adet ilacın 2.124'ü için öznitelikler çıkarılmış; hesaplanmış özellik vektörü olmayan ilaçlar filtrelenmiştir. Dolayısıyla, ilaç-ilaç etkileşimlerinin sayısı 253.449'ye düşmüştür.

6.1.2 Bilgi Çizgesi Oluşturma

LOD, verileri yayınlama, açıklama ve bağlama tekniğidir (Bizer et al., 2009). LOD, URI ve RDF gibi standartlar aracılığıyla tanımlar. RDF Web'de kaynakları tanımlamak ve değiş tokuş etmek için güçlü bir veri modeli sağlar.

İlaç özelliklerini çıkarmak için arka plan bilgisi olarak Bio2RDF (Callahan et al., 2013) olarak adlandırılan önceden bağlantılı bir açık biyomedikal veri kümesi kullanılmıştır. Bio2RDF, farklı web sitelerinde bulunan çok sayıda yaşam bilimleri veri tabanını entegre eden açık kaynaklı bir projedir. Bio2RDF, bilimsel araştırmacılar için çeşitli kaynaklardan veri bütünleştirme hizmeti sunmaktadır. Bio2RDF betikleri, SQL ve XML gibi biçimlendirilmiş verileri heterojen olarak RDF formatına dönüştürür. Bio2RDF içerisinde Drugbank, KEGG ve SIDER gibi veri tabanları dahil olmak üzere büyük biyolojik veri tabanlarından gelen verileri birbirine bağlayan büyük bir RDF çizgesi meydana getirir. Bu çalışmada, Bio2RDF projesi içerisindeki Drugbank (Wishart et al., 2007), KEGG (Kanehisa, 2000) ve PharmGKB (Klein et al., 2001) veri setleri, arka plan bilgi çizgesi olarak kullanılmıştır. Her bir veri setiyle ilgili üçlü, düğüm ve bağ tipi sayısı Çizelge 6.1'de verilmiştir.

6.1.3 Öznitelik Çıkarma

Bilgi çizgelerinden öznitelik vektörü çıkarmak amacıyla başarılı son teknoloji bilgi çizgesi gömme yöntemleri kullanılmıştır; RDF2Vec (Ristoski and Paulheim,

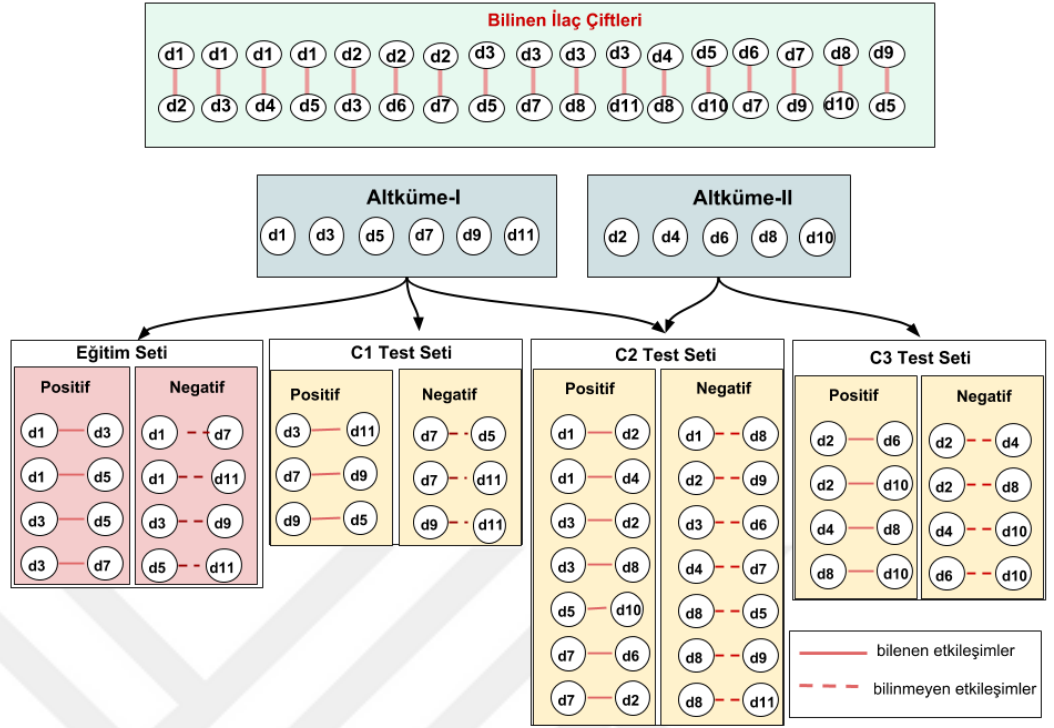
2016), TransE (Bordes et al., 2013) ve TransD (Ji et al., 2015).

6.1.4 Tahmin ve Değerlendirme

- **Yeni ilaçların tahmin performansını değerlendirme:** Önceki bazı çalışmalarda (örn., (Gottlieb et al., 2012; Zhang et al., 2015)) etkileşim bilgisi olmayan yeni ilaçlar için öngörülerde bulunmak mümkünken, bu ilaçlar için test senaryoları tasarlanmamıştır. Çift girdili (paired-input) yöntemlerinde, aynı ilaçların hem eğitim setinde hem de test setinde bulunması değerlendirmede kusurlar üretmektedir. Park ve ark. (Park and Marcotte, 2012), protein-protein etkileşimi tahmin yöntemlerini daha gerçekçi olarak değerlendirmek için test setlerini 3 sınıfa (C1, C2, C3) ayırmıştır. Bu sınıflar, DDI tahminlerini değerlendirmek için bilinen ilaçlar arasındaki etkileşimlerine (C1), yeni ilaçlar ile bilinen ilaçlar arasındaki etkileşimlerine (C2) ve yeni geliştirilen ilaçlar arasındaki etkileşimlerine (C3) denk gelmektedir.

Geleneksel çapraz doğrulamadan (cross-validation - CV) farklı olarak ilaç çiftlerinin ayrık olması yerine ilaç ayrık (ilaçlar birbiriyle örtüşmeyecek) şekilde eğitim ve test seti olarak ayrılmıştır. DDI tahminin gerçekçi bir değerlendirmesini yapmak için Park ve arkadaşlarının önerdiği şekilde girdiler üç temel test sınıfına ayrılmıştır. Bu sınıfları eğitim ve test olarak ayırmak için ilaç kümesi iki kümeye bölünür ve bu kümelerde C1, C2 ve C3 test sınıfları Şekil 6.1'te ki gibi oluşturulur. İlaçlar 2 alt kümeye (AltKüme-I ve AltKüme-II) bölünmüştür. Daha sonra bu ilaç kümelerinden biri (AltKüme-I) eğitim setini oluşturmak için kullanılır. Eğitim verisi bilinen ilaç çiftleri arasından her bir elemanı bu ilaç kümesinin içinde bulunan olarak seçilir. Test üç sınıfa ayrılır; C1 sınıf test kümesi, eğitim kümesini oluşturan ilaç kümesinden (AltKüme-I) çekilmiştir. C2 test sınıfındaki örnek çiftlerin bir elemanı AltKüme-I'den bir elemanı AltKüme-II'den çekilir. Son olarak C3 test sınıfı için örnekler AltKüme-II'den oluşturulmuştur.

- **Veri dengesi:** Daha önceki çalışmalarda ikili sınıflandırıcıya girilen negatif küme, bilinmeyen tüm etkileşimlerin tümü olarak veya etkileşimler arasından bir kısım olarak seçilir. Bilinmeyen tüm etkileşimleri negatif küme olarak belirlemek, bir veri dengesi sorunu oluşturur ve performans ölçütlerini etkiler (AUPR ve F puanı gibi). Bu problem, bilinmeyen etkileşimlerden rastgele bir alt örnekleme yapılarak (Cheng and Zhao, 2014) veya kümeleme (Hameed et al., 2017) oluyla negatif örnekler belirlenerek ele alınabilir. Bilinen ilaçlar pozitif küme olarak alınırken, negatif set bilinmeyen ilaçlar arasından pozitif kümenin sayısı kadar seçilir.



Şekil 6.1: İlaç çiftlerinin test sınıflarına göre eğitim ve test kümelerine ayrılması

- Değerlendirme Metrikleri:** Birçok çalışmada hesaplamalı ilaç-ilâç etkileşimleri tahmini için AUC skoru kullansa da, bazı çalışmalar bu skorun yeterince gerçekçi bir ölçüt olmadığı, bunun yerine AUPR ve F puanı gibi ölçümlerin kullanılması gerektiği vurgulamıştır (Hameed et al., 2017; Abdelaziz et al., 2017). Sınıflandırıcılarımızın performansını doğru bir şekilde ölçmek için doğruluk, duyarlılık, F puanı (F1) ve AUPR gibi değerlendirme ölçütleri de kullanılmıştır.
- Parametreler:** Çizge yürüşleri oluşturmak için kullanılan derinlikler 1, 2, 3, 4 ve varlık başına 250 yürüyüş gerçekleştirilmiştir. CBOW ve SG yapay sinir ağ mimarileri aşağıdaki parametreler kullanılarak eğitildi; pencere boyutu= 5, iterasyon sayısı= 5, negatif örnek sayısı= 25 ve gömme vektör boyutu= 100. Bir ilaç çifti için öznitelik oluştururken her iki ilacın vektörleri birleştirildiğinden sınıflandırıcılar 200 öznitelik kullanmışlardır. TransE ve TransD modelleri OpenKE aracı kullanarak eğitildiğinden bu araç tarafından varsayılan parametreler kullanılmıştır. Lojistik Regresyon (LR), Naive Bayes (NB) ve Rastgele Orman (RO) modellerin eğitimi için Scikit-learn makine öğrenme paketinden yararlanılmıştır. Sınıflandırıcıları oluşturmak için ise kullanılan parametreler şöyledir; LR için C=0.01, RF için ağaç sayısı= 200 ve NB için ise Gaussian çekirdeği kullanılmıştır.

Çizelge 6.2: Drugbank Bilgi Çizgesi kullanılarak Drugbank 5.0 veri seti üzerinde farklı gömme yöntemlerinin tahmin performans değerleri

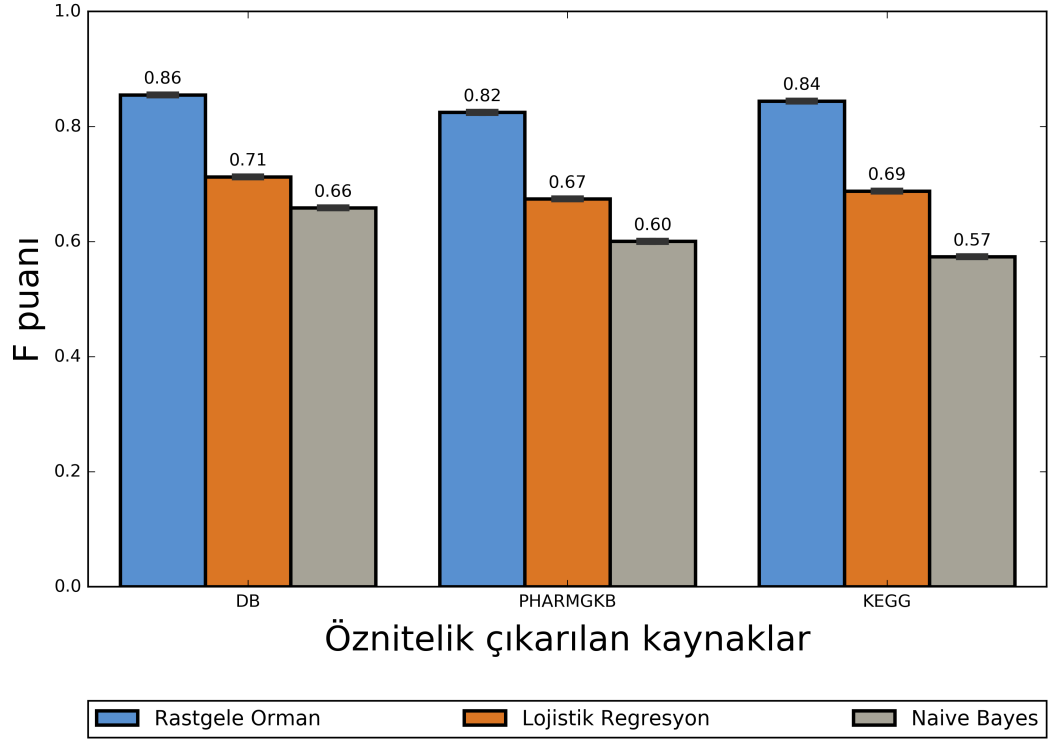
Gömme Metodu			AUC			F Puanı		
			NB	LR	RO	NB	LR	RO
RDF2Vec	Tekdüze	CBOW	0.703	0.787	0.927	0.640	0.715	0.853
		SG	0.739	0.799	0.932	0.683	0.730	0.860
	PageRank	CBOW	0.625	0.682	0.891	0.495	0.622	0.815
		SG	0.642	0.695	0.893	0.621	0.634	0.816
	PageRank Split	CBOW	0.618	0.676	0.893	0.548	0.619	0.817
		SG	0.636	0.706	0.894	0.635	0.644	0.817
TransE			0.734	0.763	0.912	0.686	0.700	0.835
TransD			0.714	0.738	0.914	0.676	0.679	0.836

6.1.5 Deneysel Sonuçlar ve Tartışma

İlk olarak, Drugbank KG üzerinde 10-kat geleneksel çapraz doğrulama (CV) kullanarak Lojistik Regresyon, Naive Bayes ve Rastgele Orman uygulayarak üç iyi bilinen bilgi çizge gömme yöntemi, RDF2Vec, TransE and TransD, denenmiştir. Deneylerin sonuçları Çizelge 6.2’de verilmiştir. Tekdüze (Uniform) ağırlık stratejisi kullanan RDF2Vec, diğer çizge yürüme oluşturma ve gömme yöntemlerinden daha iyi tahmin performansı göstermiştir. Skip-Gram Yapay Sinir Ağı kullanılarak RDF2Vec tekdüze ağırlıklandırma ile oluşturulan gömme vektörleri en iyi tahmin performansına sahiptir. Elde edilen en iyi sonuçlar, AUC değeri 0,932 ve en iyi F puan değeri 0,860, Rastgele Orman öğrenme algoritması ile bulunmuştur.

Ayrıca KEGG ve PharmGKB bilgi çizgeleri kullanarak çıkarılan öznetelik vektörlerinin tahmin performansına etkisi incelenmiştir. Şekil 6.2, farklı bilgi kaynakları kullanıldığında en iyi öznetelik vektörleri (RDF2Vec tekdüze ağırlık stratejisi) için F puanını göstermektedir. Drugbank için 0,860, PharmGKB için 0,825 ve KEGG için 0,845 F puan değerleri elde edilmiştir. PharmGKB ve KEGG veri kaynakları tek başlarında kullanıldığında Drugbank’tan daha kötü bir tahmin performans göstermiştir. Ek olarak, bu veri kaynakları bütünleştirilerek kullanıldığında ilaç vektörlerinin tahmin gücü önemli ölçüde artmamıştır (F puanı=0.860, Çizelge 6.3’te detaylar bulunmaktadır.)

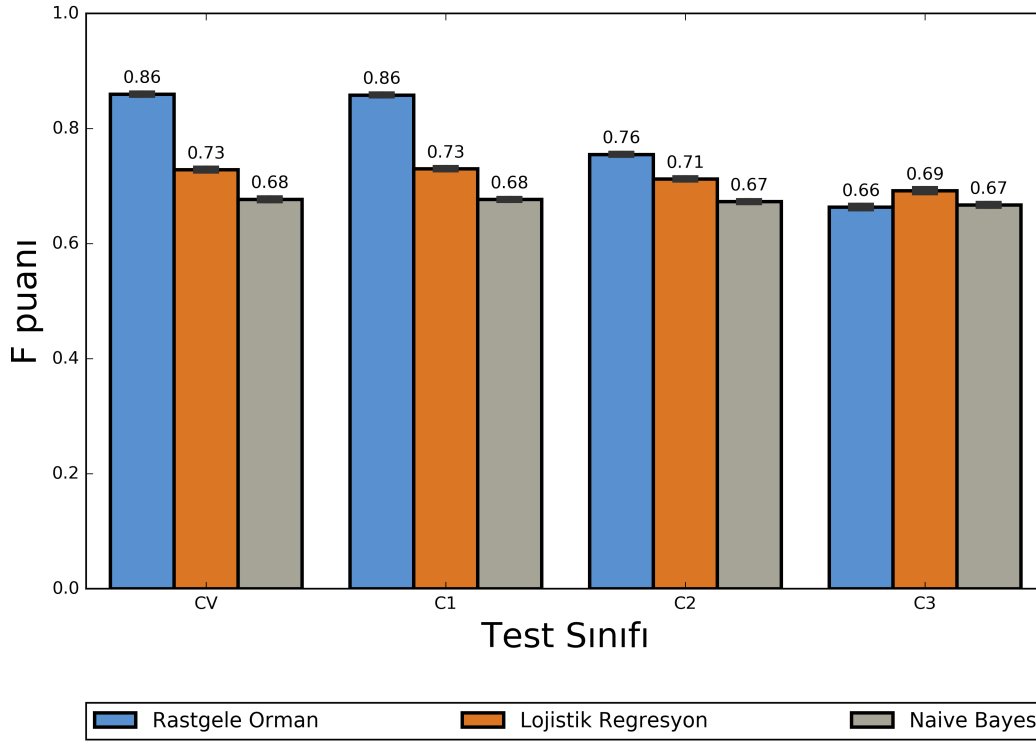
Daha sonra, önceden bahsedilen üç test sınıfını (C1, C2, C3) ve geleneksel çapraz doğrulama (CV) kullanarak RDF2Vec (tekdüze ağırlık ile) tahminleyici gücü karşılaştırılmıştır. Lojistik Regresyon, Naive Bayes ve Rastgele Orman ile 10 kat çapraz doğrulama kullanarak elde edilen sonuçların F-puan değerleri Şekil 6.3’te gösterilmiştir. Deneyler ile ilgili ayrıntılı sonuçlar ayrıca Çizelge 6.4’te verilmiştir. C1 test sınıfı geleneksel çapraz doğrulama (CV) ile benzer sonuçlar üretmektedir. C2 ve C3 sınıfları için ise F puan değerleri tüm modellerde beklendiği gibi düşme



Şekil 6.2: Geleneksel çapraz doğrulama için öznitelik veri kaynaklarının F puan karşılaştırılması

Çizelge 6.3: Birleştirilmiş (DB + KEGG + PHARMGKB) Bilgi Çizgesi kullanılarak Drugbank 5.0 veri seti üzerinde farklı gömme yöntemlerinin tahmin performans değerleri

Gömme Metodu			AUC			F Puanı		
			NB	LR	RO	NB	LR	RO
RDF2Vec	Tekdüze	CBOW	0.695	0.78	0.922	0.631	0.706	0.847
		SG	0.735	0.801	0.9340	0.677	0.729	0.86
	PageRank	CBOW	0.694	0.719	0.894	0.639	0.656	0.819
		SG	0.667	0.701	0.897	0.603	0.64	0.82
	PageRank Split	CBOW	0.676	0.706	0.894	0.624	0.647	0.818
		SG	0.68	0.699	0.898	0.641	0.638	0.821
TransE			0.703	0.732	0.909	0.670	0.675	0.831
TransD			0.700	0.734	0.908	0.671	0.679	0.830



Şekil 6.3: Modellerin test sınıflarına göre karşılaştırılması (DB + KEGG + PHARMGKB birleşimi, bilgi çizgesi olarak kullanılmıştır.)

Çizelge 6.4: Farklı çapraz doğrulama düzenekleri kullanılarak birleştirilmiş RDF bilgi çizgesinin performansı

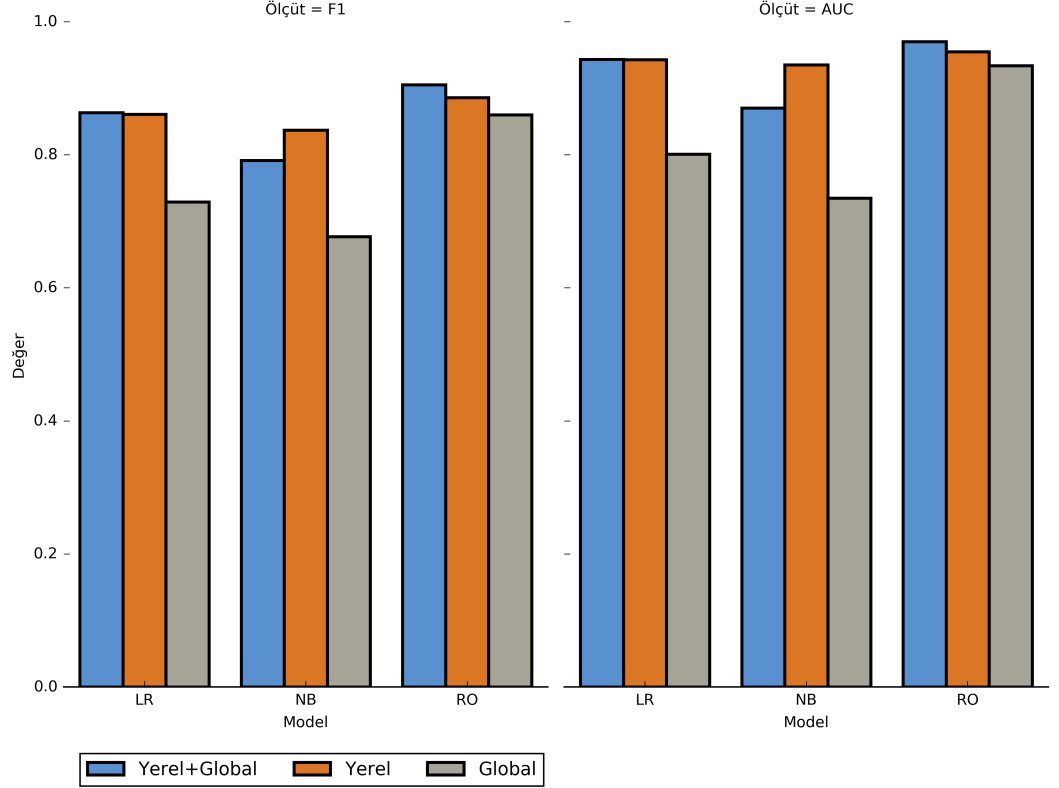
Sınıflandırıcı	Test Sınıfı	Doğruluk	AUC	AUPR	Kesinlik	Duyarlılık	F1
Lojistik Regresyon	CV	0.726	0.801	0.797	0.723	0.734	0.729
	C1	0.728	0.803	0.798	0.724	0.736	0.73
	C2	0.714	0.785	0.78	0.716	0.709	0.713
	C3	0.698	0.766	0.76	0.706	0.682	0.692
Naïve Bayes	CV	0.674	0.735	0.745	0.671	0.683	0.677
	C1	0.674	0.735	0.746	0.671	0.683	0.677
	C2	0.671	0.731	0.741	0.669	0.678	0.673
	C3	0.667	0.726	0.735	0.666	0.671	0.667
Rastgele Orman	CV	0.856	0.934	0.928	0.835	0.887	0.86
	C1	0.854	0.932	0.926	0.833	0.886	0.858
	C2	0.773	0.867	0.858	0.818	0.702	0.755
	C3	0.681	0.743	0.749	0.702	0.631	0.663

eğilimi göstermiştir. Rastgele Orman ile C1 ve C2 sınıfları için en iyi tahmin modeli *azami_derinlik* parametresi sınırsız olduğu durumda bulunurken, C3 sınıfı için en iyi tahmin modeli *azami_derinlik* parametresi 5 olarak atandığında bulunmuştur.

Ayrıca veri ağımız homojen bir yapıda olduğu için geleneksel bağ tahmin benzerlikleri sınıflandırıcımız için öznitelik olarak kullanılmıştır. Ortak Komşu Sayısı, Jaccard, Adamic ve PropFlow yakınlık ölçütleri eğitim için bir öznitelik vektörü olarak kullanılmıştır. Fakat ilaçlar eğitim ve test kümesinde ayrık olarak

Çizelge 6.5: Ağ yakınlık ölçütleri kullanılarak elde edilen tahmin sonuçları

Sınıflandırıcı	Doğruluk	AUC	AUPR	Kesinlik	Duyarlılık	F1
Lojistik Regresyon	0.866	0.943	0.945	0.899	0.826	0.861
Naïve Bayes	0.848	0.935	0.934	0.902	0.782	0.837
Rastgele Orman	0.885	0.955	0.953	0.881	0.891	0.886



Şekil 6.4: Modellerin öznel gruplarına göre karşılaştırılması (DB + KEGG + PHARMGKB birleşimi, bilgi çizgesi olarak kullanılmıştır.)

seçilmesinden dolayı C2 ve C3 test sınıfları için test verisindeki ilaç çiftlerinin benzerlik skorları hesaplanamamaktadır ve bu da bu çiftler için tahmin yapmayı mümkün kılmamaktadır. 10-kat klasik çapraz doğrulama yapılarak alınan sonuçlar Çizelge 6.5'teki verilmiştir. Homojen ağ yakınlık yöntemleri, klasik çapraz doğrulamada vektör yöntemlerine göre daha iyi sonuçlar elde etmesine rağmen etkileşim bilgisi bilinmeyen ilaç veya ilaç çiftleri için tahmin yapılmamaktadır.

Şekil 6.4'te global (RDF2Vec) ve yerel (ağ yakınlık) öznel grupların tahmin performanslarına bakıldığında genel olarak bütün sınıflandırıcılar ile yerel öznel grupların, global öznel gruplardan daha üstün olduğu görülmektedir. Ayrıca bu iki öznel grup, Rastgele Orman sınıflandırıcısı ile birleştirildiğinde en iyi tahmin sonucu elde edilmektedir (AUC=0,970, F puanı= 0,905).

RDF2Vec, TransE ve TransD kullanarak 10 kat geleneksel çapraz doğrulama

(CV) yaparak Drugbank KG üzerinde Lojistik Regresyon, Naive Bayes ve Rastgele Orman öğrenme algoritmaları uygulanmıştır. Tekdüze ağırlıkla RDF2Vec diğer gömme yöntemlerini geçmektedir. Ayrıca bilinen ilaç-ilaç etkileşimlerini öngörmek için Drugbank, PharmGKB ve KEGG gibi çeşitli ilaç bilgi çizgeleri üzerinde en başarılı yöntem (Tekdüze ağırlıklı RDF2Vec) uygulanmıştır. Bu üç veri kümesi tek başlarına ve beraber kullanıldığında tahmin performansının önemli bir biçimde artmadığı gözlenmiştir. Drugbank'ın, DDI tahmini için ilaçları karakterize etmede içerdiği bilgi bakımından diğer bilgi çizgelerini baskıladığı görülmüştür.

Aynı nesnenin (ilacın) hem eğitim setinde hem de test setinde bulunması, çift girdili (ilaç-ilaç etkileşimleri gibi) yöntemlerin değerlendirilmesinde hatalı sonuçlara ve yanlış yorumlara sebebiyet vermektedir. Bu durumlar C1, C2 ve C3 test sınıflarına ayrılarak ele alınmıştır. Daha önceki çalışmalar bu durumları dikkate almadığından performans değerlendirmesinde iyimser sonuçlara yol açmıştır. Elde edilen en iyi F-puan değerleri sırasıyla C1, C2 ve C3 için 0,858, 0,755 ve 0,692 olarak bulunmuştur. C2 ve C3 için F-puanı düşük gibi görünse de sonuçların anlamlı olduğu gözlenmiştir.

Metodun bilinen yöntemler ile karşılaştırılması:

DDI tahmini için önerilen çok sayıda yöntem olmasına rağmen, bu çalışmalarda farklı veri kümelerinin (bilinen DDI'lar) ve metodolojilerin kullanılmış olması karşılaştırma için yeterli temeli oluşturmamaktadır. Bu çalışmaların en dikkat edici olanlarından Tiresias çerçevesi (Abdelaziz et al., 2017), hem farmakolojik benzerlikleri hem de gömme özniteliklerinden elde edilen benzerlikleri kullanmıştır. Tiresias, Drugbank 4.0 veri kümesi üzerinde en iyi sonucu, tüm öznitelikleri kullanarak 0,851 F puanı ve 0,919 AUPR bulmuştur. Yalnız farmakolojik benzerlikler kullanarak ise 0.813 F puanı ve 0.887 AUPR (INDI (Gottlieb et al., 2012) ile eşit performansa sahip) değeri raporlanmıştır. Gömme vektörleri ile 0,867 F puanı ve 0,918 AUPR değeri elde edilmiştir. Önerilen gömme tabanlı yaklaşımın günümüz en gelişmiş yöntemler ile karşılaştırılabilir bir performans göstermektedir.

6.2 İlaç Endikasyon Tahmini

Genomik ve teknolojik gelişmelere rağmen ilaç keşfi ve geliştirmesi zaman alıcı ve maliyetli bir süreç olmaya devam etmektedir. Onaylanan yeni ilaç sayısı, ilaç ve sağlık endüstrisine önemli yatırımlara rağmen beklentilerin çok altında kalmıştır. Bu nedenle, halihazırda onaylanmış bilinen ilaçların kullanımının

kapsamını genişletilmesi, diğer ilgi çekici bir seçenektir ve ilaç geliştirme zamanını ve maliyetini azaltır. Bu ilaçların ABD Gıda ve İlaç İdaresi tarafından sıkı onay süreçlerden geçtiği göz önüne alındığında, güvenlik ve tolere edilebilirlik açısından asgari risk vardır. Onaylanmış ilaçların başka hastalıklar için kullanımı diğer bir adıyla ilaç yeniden konumlandırma, geliştirme sürelerini ve keşiften klinik onay aşamalarına kadar maliyetleri önemli ölçüde azaltabilir.

Makine öğrenmesi ilaç endikasyon tahminini sınıflandırma problemi olarak ele aldığından bilinen ilaç endikasyonlarını (pozitif küme) ve endikasyonu olmadığı bilinen ilaç-hastalık ikilisini (negatif küme) tanımlamayı gerektirir. Bir ilaç için bilinen bir endikasyonun olmaması, etkinlik eksikliği için güçlü bir sinyal olarak algılanabilir. Bunun altındaki temel varsayım rastgele ilaçların belirli bir endikasyon üzerinde hiçbir etkinliğinin olmayacağıdır. Onaylanmış ilaç endikasyonları bilinirken klinik çalışmalarda başarısız olmuş ilaç hastalık denemeleri çoğu zaman raporlanmaz. Brown ve arkadaşları (Brown and Patel, 2017), klinik denemelerin sonucunda etkinliği başarısız bulunan ilaç ilişkilerini de içeren bir veri tabanı sunmuştur. Bununla birlikte, başarısız ilaç denemelerini negatif örnek kaynağı olarak kullanmak sorunlu görünmektedir, zira pek çok ilaç kötü etkinliği nedeniyle değil, yan etkilerden dolayı başarısız olmuş olabilir.

Bu çalışmada önerilen yöntemin performansını karşılaştırmak için yaygın olarak kullanılan öznitelikler ile eğitilen sınıflandırıcı, dayanak yöntem (baseline method) olarak ele alınmıştır. Bu dayanak yöntem, yeni ilaç endikasyonları tahmini için sınıflandırıcıyı eğitmek üzere açık verinin sorgulanması ile çıkarılan ilaç ve hastalık özniteliklerini kullanmaktadır. Yöntemin tahmin performansı, farklı doğrulama düzenekleri için de karşılaştırılmıştır.

Ayrıca önerilen yöntem, daha önceki ilaç endikasyon yöntemleri ile karşılaştırılmıştır. Yöntemin ilaç ayrık çapraz doğrulama düzenğinde Gottlieb ve arkadaşlarının PREDICT yönetiminden ve Zhang ve arkadaşların k-en yakın komşuluk yönteminden daha iyi tahmin performansına sahip olduğu gözlenmiştir.

6.2.1 Dayanak Yöntem

Dayanak yöntem, açık kaynaklardan Bağlı Veri Teknolojilerinden yararlanarak elde edilen ilaçların hedefleri, kimyasal yapısı, yan etkileri ve hastalık fenotipik tanımı ile öznitelik vektörünü kullanmaktadır.

- **Açık Kaynaklı Veri Setlerinden Derlenen Veriler**

Bu çalışmada ilaçların potansiyel yeniden kullanımını öngörmek amacıyla açık kaynak verilerinden çıkarılan öznitelikler kullanılmıştır. Bilinmeyen potansiyel ilaç endikasyonlarını tahmin etmek için bilinen ilaç endikasyonları, ilaç ve hastalık öznitelikleri için yararlı diğer bilgiler kullanılmıştır.

Çoğu araştırmada ilaçların yeniden kullanımı için önceden elde edilmiş veri setlerini kullanılır. Bu çalışmada daha önceki çalışmalardan farklı olarak açık bağlı verinin sorgulanmasıyla elde edilmiş veriler, ilaç yeniden konulandırılması için kullanılmıştır.

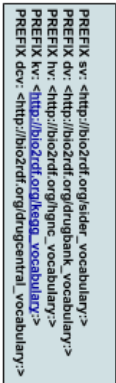
İlaç ve hastalıklarla ilgili verilere erişmek, bunları bağlamak ve entegre etmek için Bağlı Açık Veri teknolojileri ve standartları kullanılmıştır. Bio2RDF projesi (Callahan et al., 2013) bağlı veri standartlarını kullanarak Drugbank, KEGG ve SIDER dahil olmak üzere birçok büyük biyolojik veri tabanını birbirine bağlayan bir veri ağı oluşturmuştur. Bio2RDF'yi kullanarak, öğrenme modelimizin özniteliklerini üretmek için web üzerinde geniş kapsamlı kaynaklardan ilaç ve hastalık verileri elde edilmiştir. Bu bilgi ağı farklı kaynaklardan birçok farklı veri içerir, bu bilgi ağının sorgulanması bağlantıların takip edilmesiyle bulunabilir. Şekil 6.5'te Clozapine için Bio2RDF'teki farklı veri kaynaklarından elde edilen normalleştirilmiş ve entegre ilaç bilgi ağı örneklendirilmiştir. İlaç için Drugbank tanımlayıcıları, ilaç hedefleri için NCBI Gene tanımlayıcıları, hastalık ve yan etkiler için ise UMLS tanımlayıcıları kullanılarak çeşitli terminolojiler bütünleştirilmiştir.

816 ilacın DrugBank ve KEGG veri setinden hedeflerini, DrugBank veri kümesindeki kimyasal yapı bilgisini, SIDER'daki yan etki bilgisini ve BioPortal (Noy et al., 2009)'dan hastalık NDFRT ve MedDRA kavramlarını elde etmek için SPARQL sorguları yazılmıştır.

İlaç endikasyon altın standardı için Drug Central ve NDFRT veri tabanlarındaki ilaç-hastalık ilişkileri birleştirilmiştir ve oluşturulan altın standart kullanılmıştır. Bu altın standartta, 1594 ilaç ve 1611 hastalıktan oluşan toplam 8951 ilaç-hastalık ilişkisi vardır.

• Öznitelik çıkarımı

- **Kimyasal Yapı:** Moleküler seviyedeki ilaç yapısı, bağlanma aktivitesini açıklar. Kimyasal parmak izleri ilaçlar için en yaygın kullanılan yapısal profil işaretleyicidir. Parmak izleri, bazı kimyasal özelliklerin (örneğin bir C= N grubu, altı üyeli bir halka) varlığını (1) veya yokluğunu (0) gösteren bit vektörleridir. OpenBabel kütüphanesini, bir giriş kimyasal formülü (SMILES ID) alıp, 166 uzunluğundaki Moleküler Erişim Sistemi (MACCS) ikili yapısal özellik listelerini üretmek için



Şekil 6.5: Bio2RDF alt kümesinde Clozapine için bilgi ağının görselleştirilmesi

```

PREFIX dv: <http://bio2rdf.org/drugbank_vocabulary:>

SELECT distinct ?drugid ?smiles
{
  ?d <http://bio2rdf.org/openpredict_vocabulary:indication> ?i .
  ?d dv:calculated-properties ?cp .
  ?cp a dv:SMILES .
  ?cp dv:value ?smiles .
  BIND( STRAFTER(str(?d), "http://bio2rdf.org/drugbank:") AS ?drugid)
}

```

Şekil 6.6: İlaç kimyasal bilgilerini sorgulamak için yazılan SPARQL sorgusu

kullanılmıştır. İlaçların SMILES bilgilerini elde etmek için yazılan SPARQL sorgu Şekil 6.6'da verilmiştir.

- **İlaç Hedefleri:** Hedefler, ilacın kontrol ettiği biyolojik süreçlere ışık tutabilir. Burada DrugBank ve KEGG'den alınan varsayımsal hedeflerin toplam bit vektörleri kullanılmıştır. 1, ilacın bir hedefini ve 0, ilaç için bir hedef olmadığını temsil eder. İlaçların hedef bilgilerini indirmek için Bio2RDF'in KEGG ve Drugbank veri kümeleri üzerinde Şekil 6.7'de verilen birleşik bir sorgu yazılmıştır.
- **İlaç Yan Etkiler:** İlaçların ortaya çıkardığı yan etkilerin, fizyolojik bir rol olduğu düşünülmektedir. İlaç etkileri, bu verinin etiketlenmesindeki potansiyel hataya rağmen birçok çalışmada ilaç benzerliğini tanımlayabildiği belirtilmiştir (Campillos et al., 2008; Yang and Agarwal, 2011). İlaç yan etki bilgileri için SIDER (Kuhn et al., 2016) veritabanı kullanılmıştır. SIDER'daki yan etkiler metin madenciliği yapılarak elde edildiğinden yüksek sayıda yanlış pozitif içermeye meyillidir. İlaç yan etkilerini SIDER veri kümesinden edinmek için oluşturulan SPARQL sorgusu Şekil 6.8'de verilmiştir.
- **Hastalık Tanımı:** Bir ilaç, orijinal endikasyonlarından daha fazla sayıda hastalık için endike olabilir. Bu gibi durumları öğrenebilmek için hastalıkların ne kadar benzer olduğunu tanımlayan profiller çıkarmak gerekir. Hastalık profili çıkarmak için hastalığın bir ontolojide paylaştığı üst kavramlar öz nitelik olarak kullanılmıştır. BioPortal'daki hastalıkları modelleyen ontolojilerden NDFRT ve MedDRA üst kavramları bir hastalık tanımlayıcı kavramlar olarak ele alınmıştır. Hastalık tanımlarını sorgulamak için yazılan SPARQL sorgusu Şekil 6.9'da verilmiştir. Bir hastalık bu ontolojilerde yer alıyorsa, bu kavram ile ilişkili üst kavramlar öznitelik vektöründe 1 (kavram var), yoksa 0 (kavram yok) olarak temsil edilmiştir.

```

PREFIX iv: <http://bio2rdf.org/irefindex_vocabulary:>
PREFIX dv: <http://bio2rdf.org/drugbank_vocabulary:>
PREFIX hv: <http://bio2rdf.org/hgnc_vocabulary:>
PREFIX kv: <http://bio2rdf.org/kegg_vocabulary:>
SELECT distinct ?drugid ?geneid
WHERE
{
  ?drug <http://bio2rdf.org/openpredict_vocabulary:indication> ?i .
  {
    ?d a kv:Drug .
    ?d kv:target ?l .
    ?l kv:link ?t .
    BIND (URI( REPLACE(str(?t),"HSA","hsa")) AS ?target) .
    ?target a kv:Gene .
    ?target kv:x-ncbigene ?ncbi .
    #?target kv:x-uniprot ?ncbi .
    ?d kv:x-drugbank ?drug .
  }
  UNION
  {
    ?drug a dv:Drug .
    ?drug dv:target ?target .
    ?target dv:x-hgnc ?hgnc .
    ?hgnc hv:x-ncbigene ?ncbi .
    #?hgnc hv:uniprot ?ncbi .
  }
  BIND ( STRAFTER(str(?ncbi),"http://bio2rdf.org/ncbigene:") AS ?geneid)
  BIND( STRAFTER(str(?drug), "http://bio2rdf.org/drugbank:") AS ?drugid)
}

```

Şekil 6.7: İlaç hedef bilgileri sorgulamak için yazılan SPARQL sorgusu

```

PREFIX dv: <http://bio2rdf.org/drugbank_vocabulary:>
PREFIX sider: <http://bio2rdf.org/sider_vocabulary:>

SELECT distinct ?drugid ?se
{
  ?dia <http://bio2rdf.org/sider_vocabulary:drug> ?stitch_flat .
  ?dia <http://bio2rdf.org/sider_vocabulary:effect> ?se .
  {
    ?stitch_flat <http://bio2rdf.org/sider_vocabulary:x-pubchem.compound> ?pc .
  }
  UNION{
    ?stitch_flat <http://bio2rdf.org/sider_vocabulary:stitch-stereo> ?stitch_stereo .
    ?stitch_stereo <http://bio2rdf.org/sider_vocabulary:x-pubchem.compound> ?pc .
  }
  ?d dv:x-pubchemcompound ?pc .
  BIND(STRAFTER( str(?d), "http://bio2rdf.org/drugbank:") AS ?drugid)
}

```

Şekil 6.8: İlaç yan etkileri bilgilerini sorgulamak için yazılan SPARQL sorgusu

```

PREFIX sider: <http://bio2rdf.org/sider_vocabulary:>
SELECT distinct ?umlsid ?parent_id
WHERE {
  ?drug <http://bio2rdf.org/openpredict_vocabulary:indication> ?i .
  ?a <http://bioportal.bioontology.org/ontologies/umls/cui> ?i .
  ?a rdfs:subClassOf* ?parent .
  FILTER (regex(str(?parent), "^http://bio2rdf.org/ndfirt:")
    || regex(str(?parent), "^http://bio2rdf.org/meddra:") )
  BIND (STRAFTER(str(?i), "http://bio2rdf.org/umls:") AS ?umlsid)
  BIND (STRAFTER(str(?parent), "http://bio2rdf.org/") AS ?parent_id)
}

```

Şekil 6.9: Hastalık tanımlarını sorgulamak için yazılan SPARQL sorgusu

6.2.2 Eğitim ve Doğrulama

• Negatif Örneklerin Seçimi ve Oranın Dengelenmesi

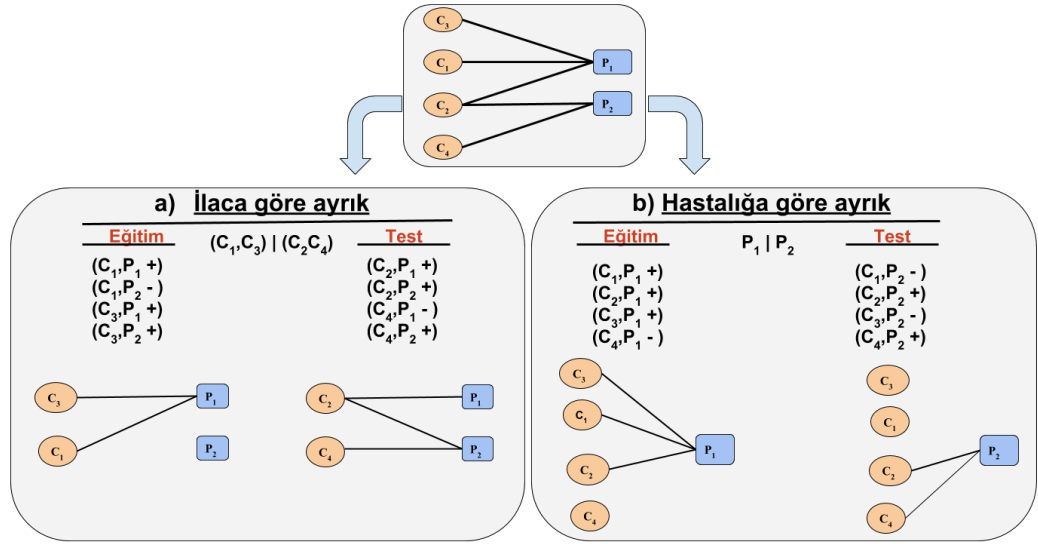
Daha önceki çalışmalarda (Gottlieb et al., 2011; Zhang et al., 2013) negatif örneklerin, pozitiflerin iki katı oranında rastgele seçilerek dengelenmesi önerilmiştir. Bu çalışmada benzer şekilde negatif set, pozitif setin iki katı oranında bilinmeyen ilaç-hastalık ilişkilerinden rastgele seçilmiştir. Bilinmeyen ilaç-hastalık ilişkileri bütün hastalıklar yerine az bir ilacın kullanıldığı hastalıklar içinden alınmıştır.

• Sınıflandırıcı Eğitimi

Sınıflandırıcıları oluşturmak için Python Scikit-learn makine öğrenmesi paketi kullanılmıştır. Lojistik Regresyon (LR), Naive Bayes (NB) ve Rastgele Orman (RO) yöntemleri ile sınıflandırıcılar oluşturulmuştur.

• Doğrulama

Var olan ilaç endikasyon tahmini çalışmaları genellikle test setinin içindeki ilaçların ayrıca eğitim setinde olacağını öngörmektedirler. Fakat araştırmacılar daha çok hiç bir endikasyonu bilinmeyen ilaçları keşfetmek ile ilgilenirler, bu yüzden bu şekilde kurulan doğrulama yeni ilaç endikasyonu tahmini ile ilgili yanıltıcı bilgi verebilir. Eğitim setinde olmayan bir ilaç veya hastalık için güvenilir tahminler yapabilmek için farklı çapraz doğrulama şemaları kullanılmıştır. İlaç endikasyonu tahmini ilgili daha önceki çalışmalar çoğunlukla ilaç-hastalık ikilisini birbiriyle örtüşmeyecek şekilde eğitim ve test seti olarak ayırmışlardır (Gottlieb et al., 2011; Zhang et al., 2013). Guney (Guney, 2016) eğitim ve test setindeki ilaçların ayrık olduğu, test setindeki ilaçların eğitim setinde olmadığı durumu incelemiştir. Bu çalışmada bu kavram genişletilerek, ilaç bazlı ayrık çapraz doğrulamanın yanında hastalık bazlı ayrık çapraz doğrulama tasarımı uygulanmıştır. Şekil 6.10'te



Şekil 6.10: Eğitim ve test kümesi ayrımının sentetik veri ile gösterimi

eğitim ve test ayrımı bir sentetik veri ile gösterilmiştir. Bu sentetik veride c_1, c_2, c_3, c_4 dört farklı birleşigi; p_1, p_2 ise iki farklı fenotipi göstermektedir. Bir birleşik ile bir fenotip arasında bilinen bir ilaç endikasyonu, çizgede ayrıt olarak gösterilmiştir. Bu örnekte iki-kat çapraz doğrulama uygulandığından, ilaç veya hastalık yönünden iki gruba ayrılmıştır. Şekil 6.10-a)'da ilaç ve ilişkileri, Şekil 6.10-b)'de ise hastalık ve ilişkileri gruplanarak eğitim ve test verisi olarak parçalanmıştır. Böylece test setindeki örneklerin, eğitim setindekilerle ilaç ve hastalık yönünden benzeşmeme durumunda tahmin performansı değişimi gözlenmiştir.

Aynı ilaç veya hastalıkların, eğitim ve test seti arasında paylaşmadığı düzenekleri inceleyerek 10-kat çapraz doğrulama uygulanmıştır. Ayrıca çapraz doğrulama 10 defa tekrarlanarak ortalama değer ölçülmüştür.

6.2.3 Deneysel Sonuçlar ve Tartışma

Tezde önerilen yöntemin kullandığı iki farklı öznitelik grubu, ayrı ayrı eğitildikten sonra bu öznitelik grupları birleştirilerek 10 kat klasik çapraz doğrulama uygulanmıştır. Her öznitelik grubu için farklı sınıflandırıcılar ile eğitilen modeller ve birleştirilen model, dayanak yöntem ile karşılaştırılmıştır. Çizelge 6.6'da elde edilen sonuçlara bakıldığında ağ yakınlıklarını içeren öznitelikler, dayanak yöntemini ve RDF2Vec yöntemini performans olarak baskılamıştır. Bu iki öznitelik grubu birleştirildiğinde tahmin performansının belirgin bir şekilde arttığı görülmektedir. Bu iki öznitelik grubu beraber toplu bir öğrenme ile eğitildiğinde, elde edilen en iyi AUC değeri 0.935 ve F-ölçütü ise 0.852'dir. Dayanak yöntemin

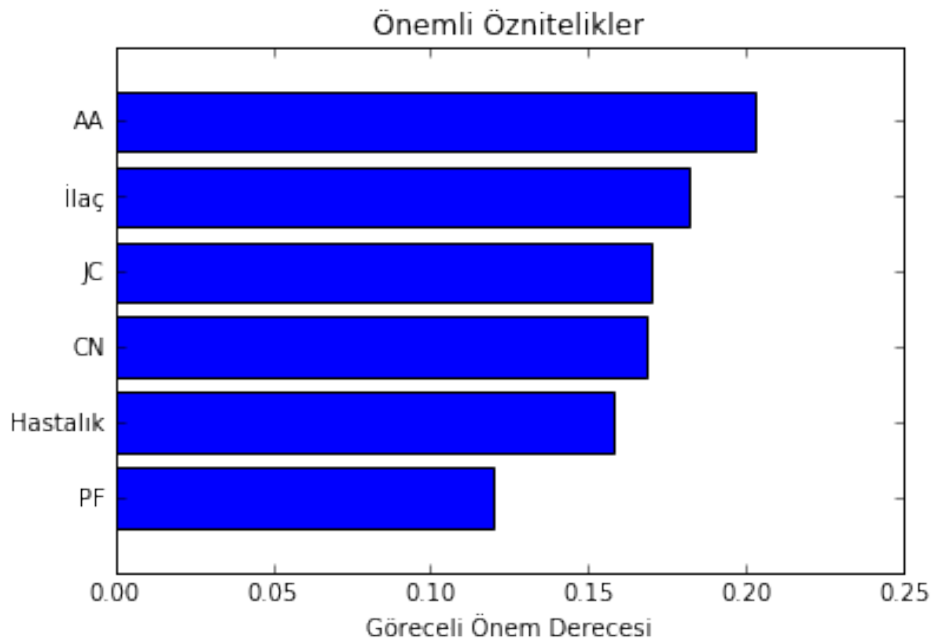
Çizelge 6.6: 10-kat çapraz doğrulama kullanarak yöntemlerin tahmin sonuçları

Yöntem		Doğruluk	AUC	AUPR	Kesinlik	Duyarlılık	F1
Dayanak Yöntem	LR	0.775	0.800	0.677	0.682	0.607	0.642
	NB	0.605	0.644	0.422	0.446	0.769	0.565
	RO	0.827	0.885	0.809	0.787	0.660	0.717
RDF2Vec	LR	0.738	0.752	0.631	0.662	0.440	0.528
	NB	0.655	0.661	0.446	0.479	0.377	0.422
	RO	0.799	0.846	0.770	0.798	0.531	0.638
Ağ Yakınlıkları	LR	0.878	0.865	0.852	0.923	0.693	0.792
	NB	0.881	0.896	0.848	0.945	0.682	0.792
	RO	0.884	0.877	0.836	0.924	0.711	0.803
Ağ Yakınlıkları + RDF2Vec	LR	0.900	0.935	0.921	0.940	0.766	0.844
	NB	0.817	0.872	0.793	0.739	0.741	0.740
	RO	0.900	0.928	0.912	0.892	0.815	0.852

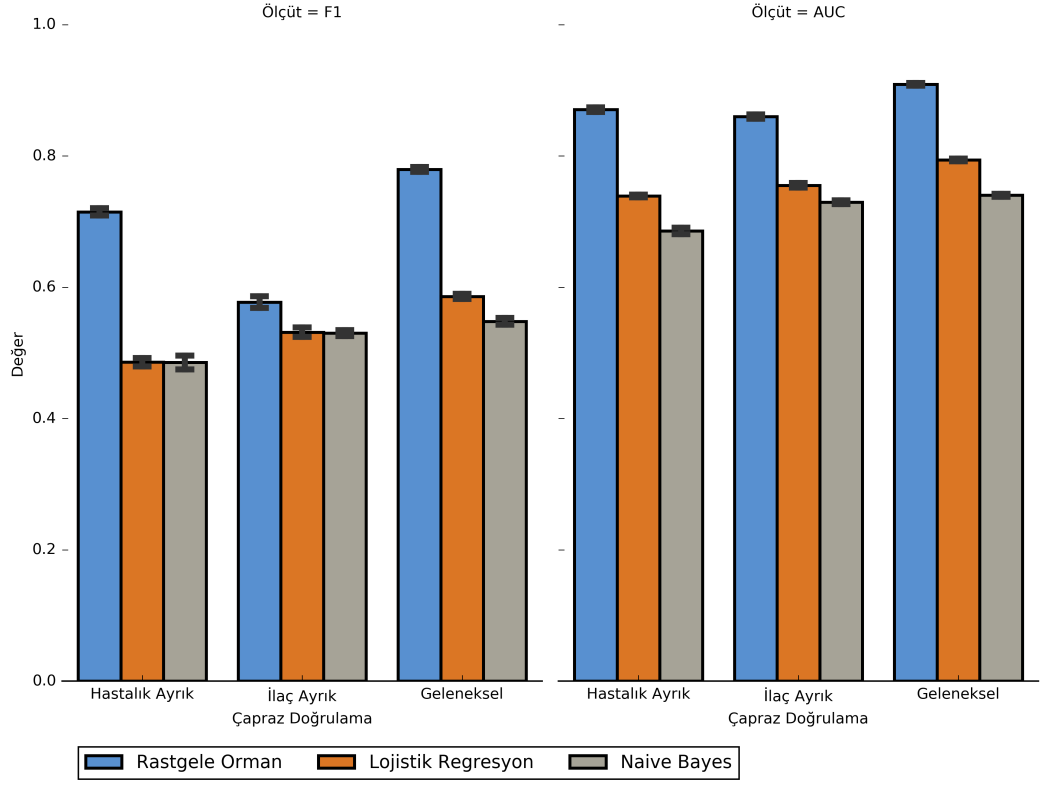
en iyi tahmin performansı Rastgele Orman sınıflandırıcı kullanılarak elde edilmiştir, bu yöntem en yüksek AUC=0.885 ve F-ölçütü= 0.717 değerlerine ulaşmıştır.

Rasgele Orman ile eğitilen birleşik modelin özniteliklerinin göreceli önem dereceleri incelendiğinde, en iyi tahmin gücüne sahip özniteliğin Adamic/Adar (AA) benzerliği olduğu görülmektedir (Şekil 6.11).

Bunun yanı sıra ilaç özniteliklerinin, ikinci en önemli öznitelik olduğu görülmektedir. İlacın kimyasal, hedef ve yan etki bilgileri gibi özniteliklerinin ilaç-hastalık ilişkisini tahmin etmede sıkça kullanılması, bu özniteliklerin önemli bir kestirici olduğunu doğrular. Ayrıca hastalığı karakterize eden hastalık öznitelik vektörünün tahmin için önemli bir işlevi olduğu görülmektedir.



Şekil 6.11: Özniteliklerin önem dereceleri



Şekil 6.12: Dayanak Yöntem için farklı çapraz doğrulama düzenekleri için elde edilen AUC ve F-puan değerleri

Dayanak yöntem bahsedilen farklı çapraz doğrulama düzenekleri için test edildiğinde, yöntemin AUC değeri, test kümesindeki ilaçların eğitim kümesi içinde olmaması durumunda (ilaç ayırık) 0.85'e, hastalıkların ayırık olduğu durumda (hastalık ayırık) ise 0.83'e kadar düşmektedir (Şekil 6.12).

Dayanak yöntem ile elde edilen veri kümesi ile oluşturulan ilaç ve hastalık öznitelikleri kullanıldığında, klasik 10-kat çapraz doğrulama yapıldığında lojistik regresyon ile eğitilen PREDICT ve SLAMS modellerinin AUC değerleri sırasıyla 0,84 ve 0,85'tir. İlaç ayırık çapraz doğrulama düzeneği göz önüne alındığında, SLAMS lojistik regresyon ile 0.66 AUC değerine düşmüştür. PREDICT ise ilaç ayırık çapraz doğrulama düzeneği altında lojistik regresyon ile ortalama 0.60 AUC değerine ulaşabilmiştir.

Yeni Endikasyon Tahminlerinin Analizi

Önerilen yöntemin tahmin gücünü değerlendirmek üzere Reboksetin ilacı için yapılan tahminler incelenmiştir. Reboksetin, presinaptik aralıkta noradrenalin geri alımını selektif olarak inhibe eden, noradrenalin taşıyıcısı için yüksek afiniteye sahip depresyon tedavisinde kullanılan etkili bir antidepresan ilaçtır.

Çizelge 6.7: Rastgele Orman sınıflandırıcısının Reboksetin ilacı için ürettiği ilk 10 endikasyon tahmini

#	UMLS	Hastalık	Olasılık	Kanıt
1	C0011581	Depressive disorder	0.995	https://www.drugbank.ca/drugs/DB00234
2	C0270549	Generalised anxiety disorder	0.990	https://www.drugbank.ca/drugs/DB00234
3	C0028768	Obsessive-Compulsive Disorder	0.820	(Fontenelle et al., 2005)
4	C0027404	Lethargy	0.815	Yok
5	C0270458	Psychotic depression	0.772	https://www.drugbank.ca/drugs/DB00234
6	C0011615	Allergic dermatitis	0.772	Yok
7	C0015967	Body temperature increased	0.770	Yok
8	C0030319	Panic disorder	0.760	(Versiani et al., 2002)
9	C1263846	Attention deficit hyperactivity disorder	0.750	(Tehrani-Doost et al., 2008; Ratner et al., 2005)
10	C0038436	Post traumatic stress disorder	0.590	(Lemke, 2002)

Güncel literatür ışığında Reboksetinin depresyon tedavisi yanı sıra uyku bozuklukları (Larrosa et al., 2001; Schmidt et al., 2016), yeme ile ilişkili bozukluklar (Hay and Claudino, 2015; Silveira et al., 2005), dikkat eksikliği ve hiperaktivite bozukluğu (ADHD) (Tehrani-Doost et al., 2008; Ratner et al., 2005), panik atak (Versiani et al., 2002), parkinson hastalarında depresyon tedavisinde (Lemke, 2002) etkili ve güvenli bir seçenek olarak önerilmektedir.

Birleştirilmiş altın standardında Reboksetin için yalnız bir (Major Depressive Disorder) endikasyon bulunmaktadır. Bu ilacın olası endikasyonları için sınıflandırıcı tarafından yapılan olasılıkları yüksek ilk 10 tahmin Çizelge 6.7’de verilmiştir. Rastgele Orman sınıflandırıcısı, Reboksetin için P değeri 0.5’ten büyük sırlamada ilk 10 olası endikasyonlara bakıldığında güncel literatür ile desteklendiği anlaşılmaktadır. Bu tahminler içinden Letarji (Lethargy) (P=0.815), Alerjik dermatit (Allergic dermatitis) (P=0.772) ve Vücut ısısı artışı (Body temperature increased) (P=0.770) gibi endikasyonların literatür ile desteklenme ihtiyacı vardır.

7 SONUÇLAR

Bu bölümde tezde yapılan çalışma özetlenerek literatüre yapılan katkılara ve tez çalışmasında elde edilen sonuçlara değinilecektir. Ayrıca önerilen yöntemin kişiye özgü ilaç önermesinde kullanımı hakkında öneriler ve ileriki çalışmalar ele alınacaktır.

7.1 Özet ve Katkılar

Biyomedikal alandaki bir çok verinin erişebilir olması ve bu verilerin bütünleştirme, verinin bütüncül bir bakış açısıyla analizini ve madenciliğini mümkün kılmaktadır. Fakat verinin büyüklüğü yanında heterojenliği bunu zorlaştırmaktadır. Bu tezde, farklı şematik yapıda ve (aynı varlık için) farklı isimlendirmeye sahip veri kümelerinden meydana gelen heterojen bilgiyi birleştirmek ve temsil etmek için çizge veri modeli kullanılmıştır. Bu heterojen bilgi ağlarında eksik bağlantıları keşfetmek için sosyal ağ analizinde kullanılan yöntemlerin ve vektör tabanlı yöntemlerin gücünden faydalanarak melez bir bağ tahmini yaklaşımı sunulmuştur. Tezin katkıları/sonuçları maddeler halinde sıralanmıştır:

1. Sosyal ağ analizinde kullanılan geleneksel bağ tahmini yöntemlerinin heterojen ağlar için uygunluğu irdelenmiş olup literatürdeki diğer benzer yaklaşımlar incelenmiştir.
2. Bağ tahmini (çift girdili) yöntemlerinin değerlendirilmesinde göz ardı edilen durumlar için test senaryoları geliştirilmiştir. Bu test senaryoları için önerilen yöntemlerin başarısı diğer yaklaşımlar ile karşılaştırılmıştır.
3. Önerilen yaklaşım için ilaç keşfi ve halk sağlığı açısından oldukça önem arz eden iki farklı kullanım durumu ele alınmış ve biyomedikal veriler üzerinde başarısı, var olan yaklaşımlar ile karşılaştırılmıştır.
4. İlk kullanım durumu ilaç-ilac etkileşimi tahminidir. İlaç-ilac etkileşimi, birden fazla ilacın aynı anda birlikte verilmesinden kaynaklanır ve bu etkileşimler hasta ölümlerine ya da ilacın geri çekilmesine neden olduğundan önemli bir araştırma potansiyeli barındırmaktadır. Bu çalışma için elde edilen sonuçlar şöyledir:

- Bildiğimiz kadarıyla bu çalışmada şu ana kadarki en büyük bilinen DDI veri seti (DDI sayısı=253.449), makine öğrenmesi modelleri

için girdi olarak kullanılmıştır. Kullanılan metodoloji çok sayıda ilaç için öznitelik çıkarılmasını sağlamıştır. Önceki çalışmalarda kullanılan bilinen DDI sayısı çok azdır (40-50 bin).

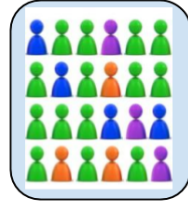
- Bu çalışmada potansiyel DDI'ları tahmin etmek için bilgi çizge vektör gömme tabanlı öznitelik vektörleri kullanılmıştır. Farklı gömme yöntemleri ve bağlı veri kaynaklarının etkisi araştırılmıştır. Bilgi gömme temsillerinin güçlü tahminleyiciler olduğu ve son teknoloji yöntemlerle karşılaştırılabilir olduğu gösterilmiştir.
- Yeni geliştirilen ilaçların tahmin performanslarını doğru bir şekilde değerlendirmek için değerlendirme senaryoları tasarlanmış ve yeni geliştirilen ilaçlar için tasarlanan bu düzeneklerin beklendiği gibi daha düşük F puanları ürettiği ancak ilaç etkileşimlerini tahmin etmede anlamlı olduğu gözlemlenmiştir.
- Yöntemin bir kısıtı, gömme özelliklerinin bir kara kutu (yapay sinir ağı) tarafından oluşturulmasından kaynaklı, ilaçların etkileşimlerinin mekanistiği hakkında bilgi sağlayamamasıdır.
- Elektronik Sağlık Kayıtları (ESK) ile entegrasyonu sayesinde yaklaşımın öngördüğü tahminler, potansiyel tehlikeli ilaçların birlikte reçetelendirmesinin önüne geçeceğinden sağlık sistemi için değerli olacaktır. Ayrıca bu tahminler dikkate alınarak büyük ölçekli klinik denemelerin tasarlanması için de yararlı olacaktır.

5. İkinci durum ise yeni ilaç-hastalık ilişkisi tahminidir. Yeni ilaç keşfi yüksek bütçeli yatırımlar gerektirmektedir. Yeni ilacın piyasaya çıkması için ulusal ve uluslararası öngörülen birçok testi geçerek onay alması gerekmektedir. Bu süreçler genellikle uzun ve pahalı süreçlerdir. Bu ilişkilerin laboratuvar ve klinik çalışmalara başlamadan çok iyi değerlendirilmesi gerekmektedir. Çalışma sonucu şöyledir:

- 816 ilaç ve 1393 hastalık içeren çok geniş bir veri kümesi ve özellikleri, Bağlı Açık Veri teknolojileri ile sorgulanarak elde edilmiş ve ilaç-hastalık ilişkisi kestirimi için kullanılmıştır. Önerilen yöntem bir kaç kaynaktan birleştirilmiş ilaç endikasyon altın standardı ile test edilmiştir.
- Gerçekçi tahmin yapmak için ilaç bilgisinin eğitim kümesinde örtüşmemesi için Guney (Guney, 2016) tarafından önerilen doğrulama yaklaşımı, hastalık için de genişletilmiştir. Yöntem, farklı sınıflandırıcılar eğitilerek belirtilen doğrulama düzenekleri için dayanak yöntem ve bilinen yöntemler ile karşılaştırılmıştır.

Geleneksel “Her Soruna Uygun Tek Çözüm “ Yöntemi

Aynı tanıyı almış tüm hastalara aynı tedavi



Bireye Özgü Tedavi Yöntemi

Hastanın benzersiz genetik profiline göre tedavi stratejisi



Şekil 7.1: Bireye Özgü Tedaviye Genel Bakış

7.2 İleriki Çalışmalar

Son yıllarda hastalık odaklı tedavi yaklaşımı yerine hasta odaklı tedavi yaklaşımı ön plana çıkmaktadır. Bireye özgü tedavi olarak nitelendirilen bu yaklaşım hastanın genetik, klinik ve diğer özelliklerini birlikte ele alıp kişiye özel bir tedavi sunmayı amaçlamaktadır. Bu yaklaşım hastanın hastalığını ve tedavi sürecini etkileyen bütün faktörlerin beraberce analizini gerektirir. Bireye özgü tedavi için geleneksel olarak kullanılan istatistiksel yöntemler yetersiz kalmakta ve bütüncül bir yaklaşıma ihtiyaç duyulmaktadır (Barabasi and Oltvai, 2004). Bu kapsamda hastalar için tedavi seçeneklerinin bireye özgü değerlendirilmesi amacıyla farklı kaynak ve ortamlardaki bilgilerin birleştirilmesini ve bütünleştirilen verinin analizini mümkün kılacak bir karar destek sistemi altyapısı geliştirilmesi için önerilen metodoloji kullanılabilir.

Bahsedilen sistemin geliştirilmesi için anlamsal web ve bağlı veri teknolojileri kullanılarak verinin bütünleştirilmesi sağlanabilir ve ortaya çıkan biyomedikal veri ağının analizi için ise anlamsal çizge madenciliği yöntemleri kullanılabilir.

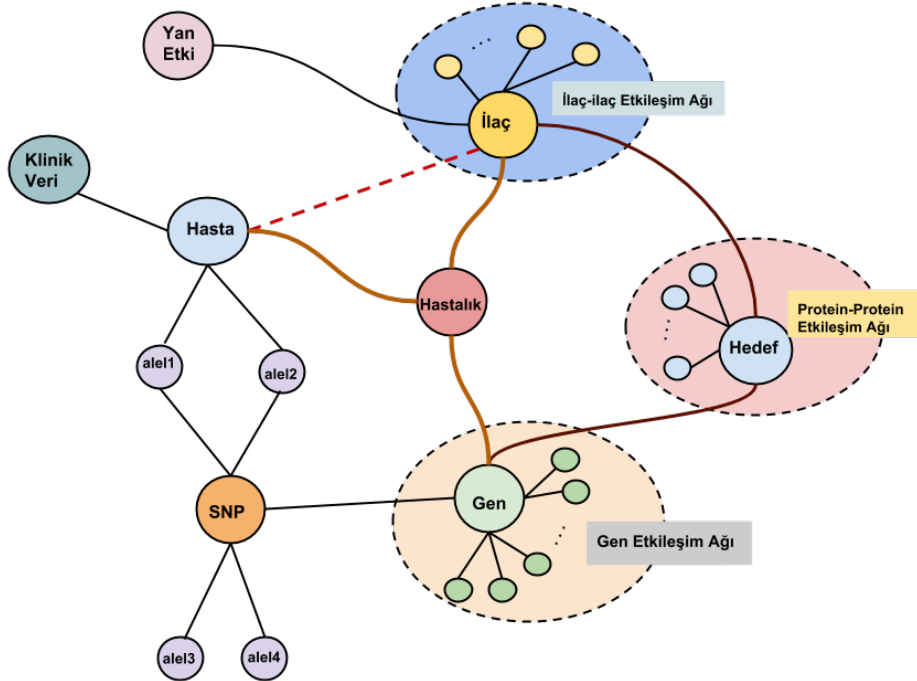
Önerilen yaklaşım, hasta geçmişine ilişkin klinik bilgileri (semptom, farmakokinetik, farmakodinamik, sonuçlar, yan etkiler) ve biyomedikal bilgileri (genetik, yolak, enzim, kimyasallar) doğrudan ilişkilendirebilecek ve hastalığın sebebini ve potansiyel tedavilerini ortaya çıkarmada önemli bakış açıları kazandıracaktır.

Bireye özgü tedavi anlayışının yerleşmesinden önce belli bir kanser türüne ve evresine sahip çoğu hastaya aynı tedavi önerilmekteydi. Genetik bilginin

artmasıyla, insanlardaki genetik farklılıkların, tümörün tedaviye verdiği farklı yanıtlar ile ilişkilendirilebileceği anlaşılmıştır. Günümüzde hastalara standart tedavi planı uygulanması yerine bireye özgü tedavi planları önerilmesi mümkün olabilmektedir (Şekil 7.1). Farmakogenomik araştırmalar kapsamında artık tümörler için belli bir gen veya protein için test yaptıktan sonra, kişinin genetiğinin ve vücut işleyişinin ilaca verdiği yanıtı nasıl etkilediğine bakılabilmektedir.

Farklı kaynaklardan elde edilen hasta bilgilerini (genotipik çeşitlilikler ile ilaç hedeflerini etkileyen olası somatik farklılıkları) anlamsal ağ teknolojilerini kullanarak birbiri ile birleştirip, önerilen yöntem ile hastalar için en iyi tedavi veya ilaç ilişkisi bulunabilir. Bireye özgü tedaviye yönelik karar destek sistemleri geliştirmek için aşağıdaki adımların gerçekleştirilmesi önerilmektedir:

1. Kişiyi özgü tedavi için gen ifadenme profili (gene expression profile), genotip profili (tek nükleotid çeşitlilikler (SNP)), klinik veri (demografik, aile geçmişi, laboratuvar sonuçları, tedavi uygulamaları ve çevresel bilgi) gibi kişiyi özel bilgilerin yanı sıra açık veri kaynaklarından kimyasalların formülasyonu, farmakogenomik bilgiler, klinik denemeler ve satışı yapılan ilaçlar hakkındaki bilimsel veriler gibi verilerin birlikte kullanılması



Şekil 7.2: Bağ tahmini yöntemlerinin bireye özgü ilaç önerilmesinde kullanılmasına yönelik bir örnek ağ

2. Hastaların genetik ve klinik bilgilerinin yanı sıra farklı biyolojik kaynaklardaki bilgilerin (gen, yolak, protein, ilaç, SNP vb.) yeni nesil bütünleştirme teknolojisi olarak ortaya çıkan Anlamsal Web teknolojileri kullanılarak bütünleştirilmesi
3. Oluşturulacak biyomedikal veri ağı ile verinin analizi için anlamsal çizge madenciliği yöntemleri kullanılması

Bağ tahmini yöntemlerinin bireye özgü ilaç önerilmesinde nasıl kullanılacağı Şekil 7.2'deki bir örnek ile gösterilmiştir. Şekilde hasta, ilaç, hedef, gen, SNP, hastalıktan oluşan veri ağının içinden bağlantısı bulunmayan hasta ile ilaç arasında bir ilişkinin oluşma ihtimali, bağ tahmini yöntemleri kullanılarak hesaplanacaktır. Bu ilacın diğer ilaçlar içinden hastaya önerilebilmesinin ne kadar anlamlı olduğu araştırılacaktır.

KAYNAKLAR DİZİNİ

- Abdelaziz, I., Fokoue, A., Hassanzadeh, O., Zhang, P., and Sadoghi, M.:** 2017, Large-scale structural and textual similarity-based mining of knowledge graph to predict drug–drug interactions, *Web Semantics: Science, Services and Agents on the World Wide Web* **44**, 104
- Adamic, L. A. and Adar, E.:** 2003, Friends and neighbors on the web, *Soc. Networks* **25**(3), 211
- Agrawal, R., Srikant, R., et al.:** 1994, Fast algorithms for mining association rules, in *Proc. 20th int. conf. very large data bases, VLDB*, Vol. 1215, 487–499 pp
- Alshahrani, M., Khan, M. A., Maddouri, O., Kinjo, A. R., Queralt-Rosinach, N., and Hoehndorf, R.:** 2017, Neuro-symbolic representation learning on biological knowledge graphs, *Bioinformatics* **33**(17), 2723
- Antoniou, G., van Harmelen, F., and Hoekstra, R.:** 2012, *A Semantic Web Primer*, MIT Press
- Balmin, A., Hristidis, V., and Papakonstantinou, Y.:** 2004, Objectrank: Authority-based keyword search in databases, in *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, VLDB Endowment, 564–575 pp
- Barabasi, A.-L. and Oltvai, Z. N.:** 2004, Network biology: understanding the cell’s functional organization, *Nature reviews genetics* **5**(2), 101
- Bates, D. W., Cullen, D. J., and Laird, N.:** 1995, Incidence of adverse drug events and potential adverse drug events, *Int. J. Trauma Nurs.* **1**(4), 117
- Berners-Lee, T., Hendler, J., and Lassila, O.:** 2001, The semantic web, *Sci. Am.* **284**(5), 34
- Bizer, C., Heath, T., and Berners-Lee, T.:** 2009, Linked data - the story so far, *Int. J. Semant. Web Inf. Syst.* **5**(3), 1
- Bizer, C. and Schultz, A.:** 2010, The r2r framework: Publishing and discovering mappings on the web., *COLD* 665
- Bobadilla, J., Ortega, F., Hernando, A., and Gutiérrez, A.:** 2013, Recommender systems survey, *Knowledge-Based Systems* **46**, 109
- Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., and Yakhnenko, O.:** 2013, Translating embeddings for modeling multi-relational data, in *Advances in neural information processing systems*, 2787–2795 pp
- Brown, A. S. and Patel, C. J.:** 2017, A standard database for drug repositioning,

Scientific Data **4**, 170029

- Callahan, A., Cruz-Toledo, J., Ansell, P., and Dumontier, M.:** 2013, Bio2RDF release 2: Improved coverage, interoperability and provenance of life science linked data, in P. Cimiano, O. Corcho, V. Presutti, L. Hollink, and S. Rudolph (eds.), *The Semantic Web: Semantics and Big Data*, Lecture Notes in Computer Science, Springer Berlin Heidelberg, 200–212 pp
- Cami, A., Manzi, S., Arnold, A., and Reis, B. Y.:** 2013, Pharmacointeraction network models predict unknown drug-drug interactions, *PLoS One* **8(4)**, e61468
- Campillos, M., Kuhn, M., Gavin, A.-C., Jensen, L. J., and Bork, P.:** 2008, Drug target identification using side-effect similarity, *Science* **321(5886)**, 263
- Cao, B., Kong, X., and Yu, P. S.:** 2014, Collective prediction of multiple types of links in heterogeneous information networks, in *2014 IEEE International Conference on Data Mining*
- Carroll, J. J., Bizer, C., Hayes, P., and Stickler, P.:** 2005, Named graphs, provenance and trust, in *Proceedings of the 14th international conference on World Wide Web*, ACM, 613–622 pp
- Celebi, R., Mostafapour, V., Yasar, E., Gumus, O., and Dikenelli, O.:** 2015, Prediction of Drug-Drug interactions using pharmacological similarities of drugs, in *2015 26th International Workshop on Database and Expert Systems Applications (DEXA)*
- Chen, H., Ding, L., Wu, Z., Yu, T., Dhanapalan, L., and Chen, J. Y.:** 2009, Semantic web for integrated network analysis in biomedicine, *Brief. Bioinform.* **10(2)**, 177
- Chen, H., Li, X., and Huang, Z.:** 2005, Link prediction approach to collaborative filtering, in *Digital Libraries, 2005. JCDL'05. Proceedings of the 5th ACM/IEEE-CS Joint Conference on*, IEEE, 141–142 pp
- Cheng, F. and Zhao, Z.:** 2014, Machine learning-based prediction of drug-drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties, *J. Am. Med. Inform. Assoc.* **21(e2)**, e278
- Cheng, W., Kasneci, G., Graepel, T., Stern, D., and Herbrich, R.:** 2011, Automated feature generation from structured knowledge, in *Proceedings of the 20th ACM international conference on Information and knowledge management - CIKM '11*
- Chiang, A. P. and Butte, A. J.:** 2009, Systematic evaluation of drug-disease relationships to identify leads for novel drug uses, *Clin. Pharmacol. Ther.* **86(5)**,

507

- Clark, D. E.:** 2004, Practical introduction to record linkage for injury research, *Inj. Prev.* **10(3)**, 186
- Davis, D., Lichtenwalter, R., and Chawla, N. V.:** 2011, Multi-relational link prediction in heterogeneous information networks, in *2011 International Conference on Advances in Social Networks Analysis and Mining*
- Davis, D., Lichtenwalter, R., and Chawla, N. V.:** 2012, Supervised methods for multi-relational link prediction, *Social Network Analysis and Mining* **3(2)**, 127
- de Vries, G. K. D.:** 2013, A fast approximation of the Weisfeiler-Lehman graph kernel for RDF data, in *Lecture Notes in Computer Science*, 606–621 pp
- de Vries, G. K. D. and de Rooij, S.:** 2015, Substructure counting graph kernels for machine learning from RDF data, *Web Semantics: Science, Services and Agents on the World Wide Web* **35**, 71
- Dehaspe, L., Toivonen, H., and King, R. D.:** 1998, Finding frequent substructures in chemical compounds., in *KDD*, Vol. 98, 1998 p.
- Ding, C. H., He, X., Zha, H., Gu, M., and Simon, H. D.:** 2001, A min-max cut algorithm for graph partitioning and data clustering, in *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, IEEE, 107–114 pp
- Ding, L., Peristeras, V., and Hausenblas, M.:** 2012, Linked open government data [guest editors' introduction], *IEEE Intell. Syst.* **27(3)**, 11
- Domingos, P. and Pazzani, M.:** 1997, On the optimality of the simple bayesian classifier under zero-one loss, *Machine learning* **29(2-3)**, 103
- Doubova Dubova, S. V., Reyes-Morales, H., Torres-Arreola, L. d. P., and Suárez-Ortega, M.:** 2007, Potential drug-drug and drug-disease interactions in prescriptions for ambulatory patients over 50 years of age in family medicine clinics in mexico city, *BMC Health Serv. Res.* **7**, 147
- Enright, A. J., Van Dongen, S., and Ouzounis, C. A.:** 2002, An efficient algorithm for large-scale detection of protein families, *Nucleic acids research* **30(7)**, 1575
- Ferdousi, R., Safdari, R., and Omid, Y.:** 2017, Computational prediction of drug-drug interactions based on drugs functional similarities, *J. Biomed. Inform.* **70**, 54
- Fontenelle, L. F., Mendlowicz, M. V., Miguel, E. C., and Versiani, M.:** 2005, Citalopram plus reboxetine in treatment-resistant obsessive-compulsive disorder, *The World Journal of Biological Psychiatry* **6(1)**, 57
- Gibson, D., Kleinberg, J., and Raghavan, P.:** 1998, Inferring web communities

from link topology, in *Proceedings of the ninth ACM conference on Hypertext and hypermedia: links, objects, time and space—structure in hypermedia systems: links, objects, time and space—structure in hypermedia systems*, ACM, 225–234 pp

Gottlieb, A., Stein, G. Y., Oron, Y., Ruppín, E., and Sharan, R.: 2012, INDI: a computational framework for inferring drug interactions and their associated recommendations, *Mol. Syst. Biol.* 8

Gottlieb, A., Stein, G. Y., Ruppín, E., and Sharan, R.: 2011, PREDICT: a method for inferring novel drug indications with application to personalized medicine, *Mol. Syst. Biol.* 7, 496

Griffith, O. L. and Griffith, M.: 2004, OMIM (online mendelian inheritance in man), in *Dictionary of Bioinformatics and Computational Biology*

Guney, E.: 2016, REPRODUCIBLE DRUG REPURPOSING: WHEN SIMILARITY DOES NOT SUFFICE, *Pac. Symp. Biocomput.* 22, 132

Haarslev, V. and Möller, R.: 2001, Racer system description, in *International Joint Conference on Automated Reasoning*, Springer, 701–705 pp

Hameed, P. N., Verspoor, K., Kusljic, S., and Halgamuge, S.: 2017, Positive-Unlabeled learning for inferring drug interactions based on heterogeneous attributes, *BMC Bioinformatics* 18(1), 140

Hancock, J. M.: 2004, SPARQL (SPARQL protocol and RDF query language), in *Dictionary of Bioinformatics and Computational Biology*

Harris, S., Seaborne, A., and Prud'hommeaux, E.: 2013, Sparql 1.1 query language, *W3C recommendation* 21(10)

Hay, P. J. and Claudino, A. M.: 2015, Bulimia nervosa: online interventions, *BMJ clinical evidence* 2015

Heath, T. and Bizer, C.: 2011, Linked data: Evolving the web into a global data space, *Synthesis Lectures on the Semantic Web: Theory and Technology* 1(1), 1

Ho, T. K.: 1995, Random decision forests, in *Document analysis and recognition, 1995., proceedings of the third international conference on*, Vol. 1, IEEE, 278–282 pp

Horrocks, I.: 1999, Fact and ifact., *Description logics* 22, 133

Horrocks, I., Patel-Schneider, P. F., Boley, H., Tabet, S., Grosz, B., Dean, M., et al.: 2004, Swrl: A semantic web rule language combining owl and ruleml, *W3C Member submission* 21, 79

Horrocks, I., Patel-Schneider, P. F., McGuinness, D. L., and Welty, C. A., OWL:

a Description-Logic-Based ontology language for the semantic web, in *The Description Logic Handbook*, 458–486 pp

Hu, G. and Agarwal, P.: 2009, Human disease-drug network based on genomic expression profiles, *PLoS One* **4(8)**, e6536

Inokuchi, A., Washio, T., and Motoda, H.: 2000, An apriori-based algorithm for mining frequent substructures from graph data, in *European Conference on Principles of Data Mining and Knowledge Discovery*, Springer, 13–23 pp

Isele, R. and Bizer, C.: 2012, Learning expressive linkage rules using genetic programming, *Proceedings VLDB Endowment* **5(11)**, 1638

Jaccard, P.: 1912, THE DISTRIBUTION OF THE FLORA IN THE ALPINE ZONE.1, *New Phytol.* **11(2)**, 37

Ji, G., He, S., Xu, L., Liu, K., and Zhao, J.: 2015, Knowledge graph embedding via dynamic mapping matrix, in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Vol. 1, 687–696 pp

Johnson, R. A., Yang, Y., Aguiar, E., Rider, A., and Chawla, N. V.: 2013, ALIVE: A multi-relational link prediction environment for the healthcare domain, in *Lecture Notes in Computer Science*, 36–46 pp

Jupp, S., Malone, J., Bolleman, J., Brandizi, M., Davies, M., Garcia, L., Gaulton, A., Gehant, S., Laibe, C., Redaschi, N., Wimalaratne, S. M., Martin, M., Le Novère, N., Parkinson, H., Birney, E., and Jenkinson, A. M.: 2014, The EBI RDF platform: linked open data for the life sciences, *Bioinformatics* **30(9)**, 1338

Kanehisa, M.: 2000, KEGG: Kyoto encyclopedia of genes and genomes, *Nucleic Acids Res.* **28(1)**, 27

Katayama, T., Wilkinson, M. D., Aoki-Kinoshita, K. F., Kawashima, S., Yamamoto, Y., Yamaguchi, A., Okamoto, S., Kawano, S., Kim, J.-D., Wang, Y., Wu, H., Kano, Y., Ono, H., Bono, H., Kocbek, S., Aerts, J., Akune, Y., Antezana, E., Arakawa, K., Aranda, B., Baran, J., Bolleman, J., Bonnal, R. J., Buttigieg, P. L., Campbell, M. P., Chen, Y.-A., Chiba, H., Cock, P. J., Cohen, K. B., Constantin, A., Duck, G., Dumontier, M., Fujisawa, T., Fujiwara, T., Goto, N., Hoehndorf, R., Igarashi, Y., Itaya, H., Ito, M., Iwasaki, W., Kalaš, M., Katoda, T., Kim, T., Kokubu, A., Komiyama, Y., Kotera, M., Laibe, C., Lapp, H., Lütteke, T., Marshall, M. S., Mori, T., Mori, H., Morita, M., Murakami, K., Nakao, M., Narimatsu, H., Nishide, H., Nishimura, Y., Nystrom-Persson, J., Ogishima, S., Okamura, Y., Okuda,

- S., Oshita, K., Packer, N. H., Prins, P., Ranzinger, R., Rocca-Serra, P., Sansone, S., Sawaki, H., Shin, S.-H., Splendiani, A., Strozzi, F., Tadaka, S., Toukach, P., Uchiyama, I., Umezaki, M., Vos, R., Whetzel, P. L., Yamada, I., Yamasaki, C., Yamashita, R., York, W. S., Zmasek, C. M., Kawamoto, S., and Takagi, T.: 2014, BioHackathon series in 2011 and 2012: penetration of ontology and linked data in life science domains, *J. Biomed. Semantics* **5**(1), 5
- Katz, L.: 1953, A new status index derived from sociometric analysis, *Psychometrika* **18**(1), 39
- Kempe, D., Kleinberg, J., and Tardos, É.: 2003, Maximizing the spread of influence through a social network, in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '03*
- Kifer, M.: 2008, Rule interchange format: The framework, in *International Conference on Web Reasoning and Rule Systems*, Springer, 1–11 pp
- Klein, T. E., Chang, J. T., Cho, M. K., Easton, K. L., Fergerson, R., Hewett, M., Lin, Z., Liu, Y., Liu, S., Oliver, D. E., Rubin, D. L., Shafa, F., Stuart, J. M., and Altman, R. B.: 2001, Integrating genotype and phenotype information: an overview of the PharmGKB project. pharmacogenetics research network and knowledge base, *Pharmacogenomics J.* **1**(3), 167
- Kleinberg, J. M.: 1999, Authoritative sources in a hyperlinked environment, *Journal of the ACM (JACM)* **46**(5), 604
- Kuhn, M., Letunic, I., Jensen, L. J., and Bork, P.: 2016, The SIDER database of drugs and side effects, *Nucleic Acids Res.* **44**(D1), D1075
- Kuramochi, M. and Karypis, G.: 2001, Frequent subgraph discovery, in *Data Mining, 2001. ICDM 2001, Proceedings IEEE international conference on*, IEEE, 313–320 pp
- Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., Lerner, J., Brunet, J.-P., Subramanian, A., Ross, K. N., Reich, M., Hieronymus, H., Wei, G., Armstrong, S. A., Haggarty, S. J., Clemons, P. A., Wei, R., Carr, S. A., Lander, E. S., and Golub, T. R.: 2006, The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease, *Science* **313**(5795), 1929
- Lao, N. and Cohen, W. W.: 2010, Relational retrieval using a combination of path-constrained random walks, *Mach. Learn.* **81**(1), 53
- Larrosa, O., de la Llave, Y., Barrio, S., Granizo, J. J., and Garcia-Borreguero, D.: 2001, Stimulant and anticonvulsant effects of reboxetine in patients with narcolepsy: a pilot study, *Sleep* **24**(3), 282

- Lazarou, J., Pomeranz, B. H., and Corey, P. N.:** 1998, Incidence of adverse drug reactions in hospitalized patients, *JAMA* **279(15)**, 1200
- Lemke, M. R.:** 2002, Effect of reboxetine on depression in parkinson's disease patients., *The Journal of clinical psychiatry* **63(4)**, 300
- Li, P., Huang, C., Fu, Y., Wang, J., Wu, Z., Ru, J., Zheng, C., Guo, Z., Chen, X., Zhou, W., Zhang, W., Li, Y., Chen, J., Lu, A., and Wang, Y.:** 2015, Large-scale exploration and analysis of drug combinations, *Bioinformatics* **31(12)**, 2007
- Liben-Nowell, D. and Kleinberg, J.:** 2003, The link prediction problem for social networks, in *Proceedings of the twelfth international conference on Information and knowledge management - CIKM '03*
- Liben-Nowell, D. and Kleinberg, J.:** 2007, The link-prediction problem for social networks, *J. Am. Soc. Inf. Sci. Technol.* **58(7)**, 1019
- Lichtenwalter, R. N., Lussier, J. T., and Chawla, N. V.:** 2010, New perspectives and methods in link prediction, in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '10*, New York, NY, USA, ACM, 243–252 pp
- Lin, H., Liu, Y., Wang, W., Yue, Y., and Lin, Z.:** 2017, Learning entity and relation embeddings for knowledge resolution, *Procedia Comput. Sci.* **108**, 345
- Lösch, U., Bloehdorn, S., and Rettinger, A.:** 2012, Graph kernels for RDF data, in *Lecture Notes in Computer Science*, 134–148 pp
- Lü, L., Linyuan, L., and Zhou, T.:** 2011, Link prediction in complex networks: A survey, *Physica A: Statistical Mechanics and its Applications* **390(6)**, 1150
- Lü, L. and Zhou, T.:** 2011, Link prediction in complex networks: A survey, *Physica A: statistical mechanics and its applications* **390(6)**, 1150
- Luo, H., Wang, J., Li, M., Luo, J., Peng, X., Wu, F.-X., and Pan, Y.:** 2016, Drug repositioning based on comprehensive similarity measures and bi-random walk algorithm, *Bioinformatics* **32(17)**, 2664
- Luo, H., Zhang, P., Huang, H., Huang, J., Kao, E., Shi, L., He, L., and Yang, L.:** 2014, DDI-CPI, a server that predicts drug–drug interactions through implementing the chemical–protein interactome, *Nucleic Acids Res.* **42(W1)**, W46
- Macropol, K., Can, T., and Singh, A. K.:** 2009, Rrw: repeated random walks on genome-scale protein networks for local cluster discovery, *BMC bioinformatics* **10(1)**, 283
- Manola, F., Miller, E., McBride, B., et al.:** 2004, Rdf primer, W3C

recommendation **10(1-107)**, 6

- Martínez, V., Navarro, C., Cano, C., Fajardo, W., and Blanco, A.:** 2015, Drugnet: Network-based drug–disease prioritization by integrating heterogeneous data, *Artificial intelligence in medicine* **63(1)**, 41
- McBride, B.:** 2004, The resource description framework (RDF) and its vocabulary description language RDFS, in *Handbook on Ontologies*, 51–65 pp
- McGuinness, D. L., Van Harmelen, F., et al.:** 2004, Owl web ontology language overview, *W3C recommendation* **10(10)**, 2004
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J.:** 2013, Distributed representations of words and phrases and their compositionality, in *Advances in neural information processing systems*, 3111–3119 pp
- Newman, M. E.:** 2001, Clustering and preferential attachment in growing networks, *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **64(2 Pt 2)**, 025102
- Newman, M. E.:** 2004, Detecting community structure in networks, *The European Physical Journal B* **38(2)**, 321
- Newman, M. E. and Park, J.:** 2003, Why social networks are different from other types of networks, *Physical Review E* **68(3)**, 036122
- Ngomo, A.-C. N. and Auer, S.:** 2011, Limes-a time-efficient approach for large-scale link discovery on the web of data., in *IJCAI*, 2312–2317 pp
- Nie, Z., Zhang, Y., Wen, J.-R., and Ma, W.-Y.:** 2005, Object-level ranking: bringing order to web objects, in *Proceedings of the 14th international conference on World Wide Web*, ACM, 567–574 pp
- Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D. L., Storey, M.-A., Chute, C. G., and Musen, M. A.:** 2009, BioPortal: ontologies and integrated data resources at the click of a mouse, *Nucleic Acids Res.* **37(Web Server issue)**, W170
- P-Star, P-STAR, PGRN Systems Biology Group, and PharmGKB:** 2011, SYSTEMS PHARMACOGENOMICS–BRIDGING THE GAP, in *Biocomputing 2012*
- Page, L., Brin, S., Motwani, R., and Winograd, T.:** 1999, *The PageRank citation ranking: Bringing order to the web.*, Technical report, Stanford InfoLab
- Pan, Y., Wang, J., and Li, M.:** 2013, *Algorithmic and artificial intelligence methods for protein bioinformatics*, Vol. 22, John Wiley & Sons
- Park, Y. and Marcotte, E. M.:** 2012, Flaws in evaluation schemes for pair-input computational predictions, *Nat. Methods* **9(12)**, 1134

- Paulheim, H. and Fümkrantz, J.:** 2012, Unsupervised generation of data mining features from linked open data, in *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics - WIMS '12*
- Percha, B. and Altman, R. B.:** 2013, Informatics confronts drug–drug interactions, *Trends Pharmacol. Sci.* **34(3)**, 178
- Perozzi, B., Al-Rfou, R., and Skiena, S.:** 2014, DeepWalk, in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14*
- Polikar, R.:** 2006, Ensemble based systems in decision making, *IEEE Circuits and Systems Magazine* **6(3)**, 21
- Ratner, S., Laor, N., Bronstein, Y., Weizman, A., and Toren, P.:** 2005, Six-week open-label reboxetine treatment in children and adolescents with attention-deficit/hyperactivity disorder, *Journal of the American Academy of Child & Adolescent Psychiatry* **44(5)**, 428
- Riechelmann, R. P., Tannock, I. F., Wang, L., Saad, E. D., Taback, N. A., and Krzyzanowska, M. K.:** 2007, Potential drug interactions and duplicate prescriptions among cancer patients, *J. Natl. Cancer Inst.* **99(8)**, 592
- Ristoski, P., Bizer, C., and Paulheim, H.:** 2015, Mining the web of linked data with RapidMiner, *Web Semantics: Science, Services and Agents on the World Wide Web* **35**, 142
- Ristoski, P. and Paulheim, H.:** 2016, RDF2Vec: RDF graph embeddings for data mining, in *Lecture Notes in Computer Science*, 498–514 pp
- Sauermann, L., Cyganiak, R., and Völkel, M.:** 2007, Cool uris for the semantic web, *W3C recommendation*
- Schmidt, C., Leibiger, J., and Fendt, M.:** 2016, The norepinephrine reuptake inhibitor reboxetine is more potent in treating murine narcoleptic episodes than the serotonin reuptake inhibitor escitalopram, *Behavioural brain research* **308**, 205
- Schultz, A., Matteini, A., Isele, R., Bizer, C., and Becker, C.:** 2011, Ldif-linked data integration framework, in *Proceedings of the Second International Conference on Consuming Linked Data-Volume 782*, CEUR-WS.org, 125–130 pp
- Selbas, R., Sencan, A., and Ecir, U.:** 2011, Data mining method for energy system applications, in *Knowledge-Oriented Applications in Data Mining*
- Silveira, R. O., Zanatto, V., Appolinario, J., and Kapczinski, F.:** 2005, An open trial of reboxetine in obese patients with binge eating disorder, *Eating and*

Weight Disorders-Studies on Anorexia, Bulimia and Obesity **10(4)**, e93

- Sirin, E., Parsia, B., Grau, B. C., Kalyanpur, A., and Katz, Y.:** 2007, Pellet: A practical owl-dl reasoner, *Web Semantics: science, services and agents on the World Wide Web* **5(2)**, 51
- Smith, B., Ceusters, W., Klagges, B., Köhler, J., Kumar, A., Lomax, J., Mungall, C., Neuhaus, F., Rector, A. L., and Rosse, C.:** 2005, Relations in biomedical ontologies, *Genome Biol.* **6(5)**, R46
- Sun, Y., Barber, R., Gupta, M., Aggarwal, C. C., and Han, J.:** 2011, Co-author relationship prediction in heterogeneous bibliographic networks, in *2011 International Conference on Advances in Social Networks Analysis and Mining*
- Tehrani-Doost, M., Moallemi, S., and Shahrivar, Z.:** 2008, An open-label trial of reboxetine in children and adolescents with attention-deficit/hyperactivity disorder, *Journal of child and adolescent psychopharmacology* **18(2)**, 179
- Versiani, M., Cassano, G., Perugi, G., Benedetti, A., Mastalli, L., Nardi, A., and Savino, M.:** 2002, Reboxetine, a selective norepinephrine reuptake inhibitor, is an effective and well-tolerated treatment for panic disorder., *The Journal of clinical psychiatry*
- Vilar, S., Uriarte, E., Santana, L., Lorberbaum, T., Hripcsak, G., Friedman, C., and Tatonetti, N. P.:** 2014, Similarity-based modeling in large-scale prediction of drug-drug interactions, *Nat. Protoc.* **9(9)**, 2147
- Volz, J., Bizer, C., Gaedke, M., and Kobilarov, G.:** 2009, Silk-a link discovery framework for the web of data., *LDOW* 538
- Wang, W., Yang, S., and Li, J.:** 2013, Drug target predictions based on heterogeneous graph inference, in *Biocomputing 2013*, World Scientific, 53–64 pp
- Wang, X., Mohanty, N., and McCallum, A.:** 2005, Group and topic discovery from relations and text, in *Proceedings of the 3rd international workshop on Link discovery*, ACM, 28–35 pp
- Wilcke, X., Bloem, P., and de Boer, V.:** 2017, The knowledge graph as the default data model for learning on heterogeneous knowledge, *Data Science* 1–19 pp
- Wishart, D. S., Knox, C., Guo, A. C., Cheng, D., Shrivastava, S., Tzur, D., Gautam, B., and Hassanali, M.:** 2007, Drugbank: a knowledgebase for drugs, drug actions and drug targets, *Nucleic acids research* **36(suppl_1)**, D901
- Wu, C., Gudivada, R. C., Aronow, B. J., and Jegga, A. G.:** 2013, Computational drug repositioning through heterogeneous network clustering, *BMC systems biology* **7(5)**, S6

- Yan, X. and Han, J.:** 2002, gspan: Graph-based substructure pattern mining, in *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*, IEEE, 721–724 pp
- Yang, L. and Agarwal, P.:** 2011, Systematic drug repositioning based on clinical side-effects, *PLoS One* **6**(12), e28025
- Yang, Y., Chawla, N., Sun, Y., and Hani, J.:** 2012, Predicting links in multi-relational and heterogeneous networks, in *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, IEEE, 755–764 pp
- Yu, X., Gu, Q., Zhou, M., and Han, J.:** 2012, Citation prediction in heterogeneous bibliographic networks, in *Proceedings of the 2012 SIAM International Conference on Data Mining*, 1119–1130 pp
- Zhang, J., Yu, P. S., and Zhou, Z.-H.:** 2014, Meta-path based multi-network collective link prediction, in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14*
- Zhang, P., Agarwal, P., and Obradovic, Z.:** 2013, Computational drug repositioning by ranking and integrating multiple data sources, in *Machine Learning and Knowledge Discovery in Databases*, Springer, Berlin, Heidelberg, 579–594 pp
- Zhang, P., Wang, F., Hu, J., and Sorrentino, R.:** 2015, Label propagation prediction of Drug-Drug interactions based on clinical side effects, *Sci. Rep.* **5**, 12339
- Zhang, W., Chen, Y., Liu, F., Luo, F., Tian, G., and Li, X.:** 2017, Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data, *BMC Bioinformatics* **18**(1), 18

ÖZGEÇMİŞ

Adı Soyadı: Remzi ÇELEBİ
Doğum Tarihi: 1982
Doğum Yeri: Batman
Uyruğu: T.C.

EĞİTİM

Yüksek Lisans: Orta Doğu Teknik Üniversitesi
Bilgisayar Mühendisliği Bölümü, 2009

Lisans: Atılım Üniversitesi
Bilgisayar Mühendisliği Bölümü, 2006

İLGİ ALANLARI

Veri Bilimi, Anlamsal Web, Biyoenformatik, Yapay Zeka

TÜRKÇE-İNGİLİZCE TERİMLER SÖZLÜĞÜ

açık kaynak	<i>open source</i>
anlamsal ağ	<i>semantic network</i>
Anlamsal Web	<i>Semantic Web</i>
arada bulunma	<i>betweenness</i>
arka plan bilgisi	<i>background knowledge</i>
bağ tahmini	<i>link prediction</i>
bağlanma	<i>docking</i>
Bağlı Açık Veri	<i>Linked Open Data</i>
Bağlı Veri	<i>Linked Data</i>
bilgi çizgesi	<i>knowledge graph</i>
bilgi çizgesi tamamlama	<i>knowledge graph completion</i>
bilgi erişimi	<i>information retrieval</i>
bilgi tabanı	<i>knowledge base</i>
bilişsel özellikler	<i>cognitive traits</i>
budama	<i>pruning</i>
bütünleştirme	<i>integration</i>
bütüncül	<i>holistic</i>
büyük boyutlu	<i>high dimensional</i>
çekirdek	<i>kernel</i>
çıkarsama	<i>reasoning</i>
çıkarsayıcı	<i>reasoner</i>
çift girdili	<i>paired input</i>
çizge	<i>graph</i>
çizgecik	<i>graphlet</i>
çoklu-ilişkisel	<i>multi-relational</i>
dağıtık gömme temsili	<i>distributed embedding representation</i>
denetimli	<i>supervised</i>
denetlimsiz	<i>unsupervised</i>
derece merkezliği	<i>degree centrality</i>
derecelendirme	<i>ranking</i>
devamlı	<i>continuous</i>
doğruluk	<i>accuracy</i>
durdurma kuralı	<i>stopping condition</i>
duyarlılık	<i>recall</i>
düğüm	<i>node</i>
eğitim kümesi	<i>training set</i>
en kısa patika	<i>shortest path</i>
endikasyon	<i>indication</i>

etiket	<i>label</i>
etki maksimizasyonu	<i>influence maximization</i>
etki yayılım	<i>influence propagation</i>
etkileşim	<i>interaction</i>
F-puanı	<i>F-score</i>
genişleme	<i>expansion</i>
gerçek	<i>fact</i>
hafıza tabanlı	<i>memory based</i>
ifade	<i>statement</i>
iki modlu	<i>two mode</i>
iki parçalı	<i>bipartite</i>
kapsama	<i>subsumption</i>
kavram şeması patikası	<i>concept schema path</i>
kesinlik	<i>precision</i>
kesişim ağacı	<i>intersection tree</i>
kesişim çizgesi	<i>intersection graph</i>
kestirici	<i>predictor</i>
komşuluk	<i>neighborhood</i>
kontra-endikasyon	<i>contra-indication</i>
topluluk öğrenme	<i>ensemble learning</i>
kök düğüm	<i>root node</i>
model tabanlı	<i>model based</i>
moleküler parmak izi	<i>molecular fingerprint</i>
nesne	<i>object</i>
ontoloji	<i>ontology</i>
ölçüt	<i>metric</i>
özgüllük	<i>specificity</i>
özne	<i>subject</i>
öznitelik	<i>feature</i>
özvektör merkezliği	<i>eigenvector centrality</i>
patika	<i>path</i>
popülerite	<i>popularity</i>
rastgele	<i>random</i>
rastgele yürüyüş	<i>random walk</i>
gürbüz	<i>robust</i>
seyrek	<i>sparse</i>
SPARQL uç notkası	<i>SPARQL endpoint</i>
spektral çizge bölümleme	<i>spectral graph portioning</i>
şema	<i>schema</i>
şişme	<i>inflation</i>

tavsiye sistemleri	<i>recommendation systems</i>
tek sıcak kodlama	<i>one-hot encoding</i>
test kümesi	<i>test set</i>
uyarlama	<i>adaptation</i>
üçlü	<i>triple</i>
üçlü deposu	<i>triple store</i>
üçlü sayısı	<i>triad census</i>
üst veri	<i>meta-data</i>
üst patika	<i>meta-path</i>
varlık	<i>entity</i>
veri patikası	<i>data path</i>
yakınlık	<i>proximity</i>
yan etki	<i>side effect</i>
yapay sinir ağı	<i>artificial neural network</i>
yerel	<i>local</i>
yolak	<i>pathway</i>
yüklem	<i>predicate</i>
yürüyüş	<i>walk</i>
zıplama	<i>hop</i>