

**T.C.
FIRAT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

**SOSYAL AĞ MODELLEME KULLANILARAK TIBBİ VERİ AĞINDA
BAĞLANTI TAHMİNİ YÖNTEMLERİ**

DOKTORA TEZİ

**Buket KAYA
(102113201)**

Anabilim Dalı: Elektrik-Elektronik Mühendisliği

Programı: Devreler ve Sistemler

Danışman: Prof. Dr. Mustafa POYRAZ

Tezin Enstitüye Verildiği Tarih: 29 Aralık 2014

OCAK-2015

**T.C
FIRAT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

**SOSYAL AĞ MODELLEME KULLANILARAK TIBBİ VERİ AĞINDA
BAĞLANTI TAHMİNİ YÖNTEMLERİ**

Buket KAYA

(102113201)

Tezin Verildiği Tarih: 29 Aralık 2014

Tezin Savunulduğu Tarih: 22 Ocak 2015

Tez Danışmanı : Prof.Dr. Mustafa POYRAZ (F.Ü.)

Diğer Jüri Üyeleri : Prof.Dr. M. Kemal KIYMIK (K.S.Ü.)

Prof.Dr. Yakup DEMİR (F.Ü.)

Prof.Dr. İbrahim TÜRKOĞLU (F.Ü.)

Doç.Dr. Arif GÜLTEN (F.Ü.)

OCAK-2015

ÖNSÖZ

Bu tez çalışması, sosyal ağ analizi ve modellemesiyle tıbbi bakım hizmetlerinin birleştirildiği ilk gayretlerden biridir. Uzun ve yorucu bir çalışmanın ürünü olarak ortaya çıkan bu doktora tezinde birçok kişinin katkısı bulunmaktadır. Bu nedenle öncelikle tez çalışmam süresince sağladığı destek ve sabrı dolayısıyla değerli danışman hocam sayın Prof.Dr. Mustafa POYRAZ’a sonsuz teşekkürlerimi sunarım.

Ayrıca tezimin değişik kademelerinde yönlendirmeleriyle ılık tutan Yrd.Doç.Dr. Haluk EREN’e ve Prof.Dr. İbrahim Türkoğlu’na teşekkür ederim. Girişimsel olmayan etik kurul raporu olarak elde ettiğimiz hastaların verilerini bize sağladığı için F.Ü Tıp Fakültesi Hastanesine, Fırat Üniversitesi Tıp Fakültesi öğretim üyesi Doç. Dr. Dilara KAMAN’a ve Yrd. Doç. Dr. Ayhan AKBAL’a şükranlarımı sunarım.

Son olarak, beni her zaman destekleyen ve teşvik eden, başarılarımla gururlanan, kısaca her zaman benimle olan eşime ve aileme sonsuz teşekkür ederim.

Buket KAYA

ELAZIĞ – 2015

İÇİNDEKİLER

	<u>Sayfa No</u>
ÖNSÖZ	II
İÇİNDEKİLER.....	III
ÖZET	VI
SUMMARY	VII
ŞEKİLLER LİSTESİ	VIII
TABLolar LİSTESİ	IX
KISALTMALAR LİSTESİ	XI
SEMBOLLER LİSTESİ	XI
1. GİRİŞ.....	1
1.1. Tezin Amacı ve Katkıları	2
1.2. Tezin Yapısı	3
2. SOSYAL AĞLAR.....	5
2.1. Sosyal Ağ Analizi.....	6
2.2. Sosyal Ağ Analizinde Kullanılan Temel Kavramlar.....	9
2.2.1. Derece Merkeziliği	9
2.2.2. Yakınlık Merkeziliği.....	9
2.2.3. Arasındalık Merkeziliği.....	10
2.2.4. Kenar Arasındalık Merkeziliği.....	10
2.2.5. Özvektör Merkeziliği	12
2.2.6. Diğer Kavramlar.....	13
3. BAĞLANTI TAHMİNİ	15
3.1. Bağlantı Tahmini Yöntemi.....	15
3.2. Bağlantı Tahmini ile İlgili Yapılan Çalışmalar	18
3.2.1. Çizge Teorili Yaklaşımlar	18
3.2.2. Olasılıksal Yaklaşımlar	21
3.2.3. Benzerlik Tabanlı Yaklaşımlar.....	24
3.2.3.1. Komşuluk Tabanlı Ölçütler	24
3.2.3.2. Yol Tabanlı Ölçütler.....	26
3.2.4. Diğer Yaklaşımlar	28

4.	BAĞLANTI TAHMİNİNDE MAKİNE ÖĞRENMESİ YÖNTEMLERİ.....	32
4.1.	Denetimsiz Yöntemler.....	32
4.2.	Denetimli Yöntemler.....	33
4.3.	Öğrenme Modellerini Değerlendirme Ölçütleri.....	43
4.3.1.	Sınıflandırma Hatası.....	43
4.3.2.	Karışıklık Matrisi	43
5.	GELİŞEN DURUMLU HASTALIK BELİRTİLERİ AĞINDA	
	BAĞLANTI TAHMİNİ	47
5.1.	Hastalık Belirti Ağının Oluşturulması.....	47
5.2.	Hastalık Belirti Ağları İçin Önerilen Denetimli Bağlantı Tahmini.....	48
5.2.1.	Hastalık Belirti Ağının Gelişen Yapısı.....	49
5.2.2.	Gelişen Durumlar	50
5.2.2.1.	Tutarlı Durum	50
5.2.2.2.	Güçlendiren Durum.....	50
5.2.2.3.	Zayıflatan Durum	51
5.2.3.	Denetimli Bağlantı Tahmini İçin Tahmin Edici Niteliklerin ve Sınıf Seviyesinin Belirlenmesi	51
5.2.4.	Gelişen Durumlu Hastalık Belirti Ağında Denetimli Bağlantı Tahmini Algoritması.....	54
5.2.4.2.	Değerlendirme Sonuçları.....	55
5.2.5.1.	Hastalık Belirti Ağının Versiyonları	55
5.2.5.2.	Değerlendirme Sonuçları.....	56
5.3.	Hastalık Belirti Ağları için Önerilen Denetimsiz Bağlantı Tahmini Algoritması.....	62
5.3.1.	Değerlendirme Sonuçları.....	64
5.3.2.	Son Çerçeveleri Daha Önemli Kılarak Elde Edilen Sonuçlar	70
6.	GELİŞEN DURUMLU HASTALIK AĞINDA YAŞ SERİLERİ	
	TABANLI BAĞLANTI TAHMİNİ.....	75
6.1.	Hastalık Ağının Oluşturulması.....	75
6.2.	Yaş Serileri Tabanlı Bağlantı Tahmini Algoritması.....	76
6.2.1.	Uygulama Sonuçları	77
6.2.1.1.	Hastalık Ağı.....	77
6.2.1.2.	Değerlendirme Sonuçları.....	79

6.3.	Yaş Serileri Tabanlı Denetimli Bağlantı Tahmini Sonuçları	90
7.	SONUÇLAR VE ÖNERİLER.....	93
	KAYNAKLAR.....	95
	ÖZGEÇMİŞ	102

ÖZET

Son yıllarda sağlık hizmetlerine sosyal ağ analizi ve yöntemlerinin uygulanmasına yönelik artan bir ilginin olduğuna şahit olunmaktadır. Bağlantı tahmini, sosyal ağ analizi ile ele alınan önemli bir problemdir. Birçok farklı alan için bağlantı tahmini, anlaşılması doğrudan zor veya pahalı muhtemel ilişkileri tanımlamak veya gelecek davranışı tahmin etmek için kullanılır. Bu alanlardan biri, tıbbi bakım araştırma alanıyla ilgili olanıdır. Tıbbi bakım alanının hastalık ve riskinin başlangıcını tanımlamada duyarlı olmaktan ziyade, daha ileriye yönelik olmaya ihtiyacı vardır. Şu anda doktorlar, hastanın sağlık aşamasının daha ileriye yönelik tespiti için, laboratuvar sonuçlarını kullanırlar. Fakat böyle bir tespit, sadece bir kaç tıbbi parametreye (hastalık belirtisi) odaklanma veya doktorun tecrübe, hafıza veya zamanıyla bağlantılı olma dezavantajlarını içerir. Bundan dolayı, mevcut tıbbi gelişme ileriye dönük değildir ve bir hastalığı en erken safhada işaretleriyle ele alma veya yok etmede yeterli değildir. Yukarıda bahsedilen problemlere bir çare olarak, bu tezde bireylerin hastalığa yakalanma riskinin önceden belirlenmesi ve hastalığın en erken işaretlerinde doğru hareketler yapılması için sosyal ağ ile modellemeden faydalanılan bazı yöntemler geliştirilmiştir. Bu maksatla öncelikle, hastalığın anormal parametreleri arasındaki ilişkileri gösteren ağırlıklandırılmış bir tıbbi veri ağı oluşturulacaktır. Daha sonra farklı tür laboratuvar elde edilen sonuçlar toplanarak, herhangi bir hastalığa neden olabilecek bütün parametreler arasındaki ilişkileri tanımlamak için oluşturan tıbbi veri ağının gelişimine dayalı denetimli ve denetimsiz bağlantı tahmin yöntemleri önerilmiştir. En son olarak, 210.000'den daha fazla hastanın laboratuvar sonuçlarıyla oluşturulan tıbbi veri ağı üzerinden önerilen yöntemler test edilmiştir. Tezin sonraki bölümünde ise bu defa hastalıkları doğrudan ele alan bir hastalık ağı oluşturulmuştur. Yaşa göre gruplandırılmış ağ serilerinden, belirli yaş aralıklarında kişilerin hastalığa yakalanma riski yine bağlantı tahmini yöntemiyle elde edilmeye çalışılmıştır. Uygulama sonuçları, denetimli ve denetimsiz bağlantı tahmini adı altında sunduğumuz her iki yöntemin de doğruluğunu ve gelecek hastalık riskleri yakalamadaki başarısını gösterir.

Anahtar Kelimeler: Sosyal Ağ Analizi, Bağlantı Tahmini, Hastalık Belirti Ağı, Hastalık Ağı, Denetimli ve Denetimsiz Öğrenme

SUMMARY

Link Prediction Methods in Medical Data Network using Social Network Modelling

In the last few years we are witnessing to an increasing interest in the application of social network analysis and methods to health care information and management systems. Link prediction is an important task treated by social network analysis. For many different areas, link prediction can be used to expect future behavior or to recognize likely relationships that are hard or expensive to understand directly. One of these areas is related to medical care research area. Medical care area needs to become more proactive than reactive in recognizing the onset of disease and risk. Currently, physicians use laboratory results to further determine the patient's stage of health. However, there is some disadvantages such as generally to focus on only a few medical parameters (symptoms) and to linked by a particular doctor's experience, memory, and time. Therefore, current medical care is not proactive and is not enough treating or eliminating a disease at the earliest signs. As a remedy to the above mentioned problems, in this thesis, we propose a predictor to determine the risk of individuals to develop disease, and to undertake the correct actions at the earliest signs of illness. To this purpose, we first construct a weighted medical data network which indicates the relationships between abnormal parameters of disease. Then, we propose supervised and unsupervised link prediction methods based on the evolution of the constructed medical data network in order to identify the relations between all the parameters which can cause any disease, gathering the results obtained at several laboratories. Finally, we test the proposed method on the medical data network constructed with laboratory results of patients more than 210,000. In the next section of the thesis, a disease network considering the relations between diseases is proposed. Then, we present two link prediction methods based on supervised and unsupervised strategies to identify the connections between diseases, building the evolving structure of medical data network with respect to patients' ages. Experiments on a real network demonstrate that the proposed approach can reveal new abnormal parameter and disease correlations accurately and perform well at capturing future disease risks.

Keywords: Social Network Analysis, Link Prediction, Symptom Network, Disease Network, Supervised and Unsupervised Learning.

ŞEKİLLER LİSTESİ

	<u>Sayfa No</u>
Şekil 2.1.	Sosyal ağ örnekleri 6
Şekil 2.2.	Düğüm, kenarlar ve kenarların düğüm kümeleri 8
Şekil 2.3.	Yönlü ve yönsüz çizgeler 8
Şekil 2.4.	Kite ağı 9
Şekil 2.5.	Kenar arasındalık merkeziliğini gösteren örnek bir ağ 10
Şekil 2.6.	Kenar arasındalık değerin hesaplandığı örnek iki ağ 12
Şekil 2.7.	Özvektör, derece, yakınlık ve arasındalık merkeziliği gösterilen örnek bir ağ 13
Şekil 3.1.	Sosyal ağın zamansal değişimi 17
Şekil 3.2.	Bağlantı tahmini 17
Şekil 4.1.	MRLP hesaplamasına bir örnek 38
Şekil 4.2.	Kesinlik ve anma arasındaki ilişki 46
Şekil 5.1.	Hastalık belirti ağının bir parçası 48
Şekil 5.2.	Beş çerçeveli bir hastalık belirti ağının gelişen yapısı 49
Şekil 5.3.	Hastalık belirti ağında farklı tahmin edicilerin kesinlik değerleri 65
Şekil 5.4.	Hastalık belirti ağında farklı tahmin edicilerin anma değerleri 65
Şekil 5.5.	Hastalık belirti ağında farklı tahmin edicilerin F-ölçütü değerleri 66
Şekil 5.6.	Daha küçük aralıklı çerçevelerde farklı tahmin edicilerin kesinlik değerleri 68
Şekil 5.7.	Daha küçük aralıklı çerçevelerde farklı tahmin edicilerin anma değerleri 68
Şekil 5.8.	Daha küçük aralıklı çerçevelerde farklı tahmin edicilerin F- ölçütü değerleri 69
Şekil 5.9.	Önem çerçeveli hastalık belirti ağında altı farklı tahmin edicinin kesinlik değerleri 72
Şekil 5.10.	Önem çerçeveli hastalık belirti ağında altı farklı tahmin edicinin anma değerleri 72
Şekil 5.11.	Önem çerçeveli hastalık belirti ağında altı farklı tahmin edicinin F-ölçütü değerleri 73
Şekil 6.1.	ICD -10 kodlu hastalık ağının bir parçası 76

TABLÖLAR LİSTESİ

	<u>Sayfa No</u>
Tablo 4.1. Özelliklerin listesi	36
Tablo 4.2. Karışıklık matrisi	43
Tablo 5.1. Tam kan sayımı testindeki ilk 10 anormal parametre ve bunların görülme sıklığı	57
Tablo 5.2. Klinik biyokimya testindeki ilk 10 anormal parametre ve bunların görülme sıklığı	58
Tablo 5.3. İdrar testindeki ilk 10 anormal parametre ve bunların görülme sıklığı	58
Tablo 5.4. Hormon testindeki ilk 10 anormal parametre ve bunların görülme sıklığı.....	59
Tablo 5.5. Hastalık belirti ağında karşılaştırılan yöntemlerin doğruluk değerleri (%).....	59
Tablo 5.6. Hastalık belirti ağında karşılaştırılan yöntemlerin kesinlik değerleri (%).....	60
Tablo 5.7. Hastalık belirti ağında karşılaştırılan yöntemlerin anma değerleri (%)	60
Tablo 5.8. Daha yaşlı hastalı hastalık belirti ağında karşılaştırılan yöntemlerin doğruluk değerleri	61
Tablo 5.9. Daha yaşlı hastalı hastalık belirti ağında karşılaştırılan yöntemlerin kesinlik değerleri(%).....	62
Tablo 5.10. Daha yaşlı hastalı hastalık belirti ağında karşılaştırılan yöntemlerin anma değerleri (%).....	62
Tablo 5.11. . [40,50) Geçerlilik kümesinde tahmin edilen en yüksek skorlu 10 bağlantı türü 67	
Tablo 5.12. [50,55) Geçerlilik kümesinde tahmin edilen en yüksek skorlu 10 bağlantı türü	70
Tablo 5.13. Farklı önem fonksiyonlarına göre altı farklı yöntemin kesinlik değerleri (%)	73
Tablo 6.1. Uygulamalarda kullanılan yaş serileri	78
Tablo 6.2. Birinci uygulamanın geçerlilik kümesinde en sık görülen 20 hastalık	79
Tablo 6.3. Küçük boyutlu çerçevelerde altı farklı yöntemin kesinlik değerleri(%).....	80

Tablo 6.4.	Küçük boyutlu çerçevelerde altı farklı yöntemin anma değerleri (%).....	81
Tablo 6.5.	Küçük boyutlu çerçevelerde altı farklı yöntemin F-ölçütü değerleri (%).....	81
Tablo 6.6.	Uygulamanın birinci kümesinde önerilen yöntem tarafından bulunan en yüksek 10 ilişki	82
Tablo 6.7.	Daha büyük boyutlu çerçevelerde altı farklı yöntemin kesinlik değerleri (%)	83
Tablo 6.8.	Daha büyük boyutlu çerçevelerde altı farklı yöntemin anma değerleri (%).....	83
Tablo 6.9.	Daha büyük boyutlu çerçevelerde altı farklı yöntemin F-ölçütü değerleri (%)	83
Tablo 6.10.	Uygulamanın ikinci kümesinde önerilen yöntem tarafından bulunan en yüksek 10 ilişki	84
Tablo 6.11.	Daha genç hastalarda altı farklı yöntemin kesinlik değerleri (%).....	85
Tablo 6.12.	Daha genç hastalarda altı farklı yöntemin anma değerleri (%).....	85
Tablo 6.13.	Daha genç hastalarda altı farklı yöntemin F- ölçütü değerleri (%).....	85
Tablo 6.14.	Uygulamanın üçüncü kümesinde önerilen yöntem tarafından bulunan en yüksek 10 ilişki	86
Tablo 6.15.	Daha yaşlı hastalarda altı farklı yöntemin kesinlik değerleri (%).....	87
Tablo 6.16.	Daha yaşlı hastalarda altı farklı yöntemin anma değerleri (%).....	87
Tablo 6.17.	Daha yaşlı hastalarda altı farklı yöntemin F- ölçütü değerleri (%).....	87
Tablo 6.18.	Uygulamanın dördüncü kümesinde önerilen yöntem tarafından bulunan en yüksek 10 ilişki.....	88
Tablo 6.19.	Farklı önem fonksiyonları için önerilen yöntemin kesinlik değerleri (%).....	89
Tablo 6.20.	1 nolu yaş serisinde karşılaştırılan yöntemlerin doğruluk (D) ve kesinlik (K) değerleri (%)	91
Tablo 6. 21.	2 nolu yaş serisinde karşılaştırılan yöntemlerin doğruluk (D) ve kesinlik (K) değerleri (%)	91
Tablo 6.22.	3 nolu yaş serisinde karşılaştırılan yöntemlerin doğruluk (D) ve kesinlik (K) değerleri	92
Tablo 6.23.	4 nolu yaş serisinde karşılaştırılan yöntemlerin doğruluk (D) ve kesinlik (K) değerleri	92

KISALTMALAR LİSTESİ

HPLP	: Yüksek Performanslı Bağlantı Tahmini
MRLP	: Çok İlişkili Bağlantı Tahmini
SVM	: Destek Vektör Makinası
KNN	: K-En Yakın Komşuluk
VWS	: Dikey Rüzgar Hızı
HWS	: Yatay Rüzgar Hızı
SST	: Sıcaklık
PW	: Yağış
RH	: Bağıl Nem
GH	: Yükseklik
MRLP	: Çok İlişkili Bağlantı Tahmini Yöntemi
HPLP	: Yüksek Performanslı Bağlantı Tahmini Yöntemi
RBF	: Çapsal Tabanlı Fonksiyon
DP	: Doğru Pozitif
YN	: Yanlış Negatif
YP	: Yanlış Pozitif
DN	: Doğru Negatif
OK	: Ortak Komşular
JK	: Jaccard Katsayısı
TB	: Tercihli Bağlılık
AA	: Adamik/Adar Katsayısı
KP	: Kaynak Paylaştırma İndeksi
WHO	: Dünya Sağlık Örgütü
ICD	: Hastalıkların Uluslararası Sınıflandırılması

SEMBOLLER LİSTESİ

V	: Sonlu düğüm kümesi
E	: Sonlu kenar kümesi
C	: Kümelenme Katsayısı
$p(i)$	: Yeni bir düğümün i düğümüne bağlanma olasılığı
$C_c(i)$	: i düğümünün yakınlık merkeziliği
l_G	: i ve j düğümleri arasındaki en kısa yol
w	: Ağırlık vektörü
$D_k(x, y)$	: k . çerçevedeki x ve y düğümleri arasındaki zayıflatıcı durum skoru
$S_k(x, y)$	: k . çerçevedeki x ve y düğümleri arasındaki güçlendirici durum skoru
$C_k(x, y)$	: k . çerçevedeki x ve y düğümleri arasındaki tutarlı durum skoru
α	: İkincil durumların x ve y düğümlerini ne kadar güçlü etkilediğini gösteren kontrol faktörü
$\Gamma(x)$	: x düğümünün komşu dizi seti
$\Gamma(y)$	: y düğümünün komşu dizi seti
r	: Durum değişim oranı
$I(k)$	: k . çerçevenin önem değeri
I	: Birim matris
$H(x, y)$	: x düğümünden y düğümüne erişmesi için beklenen adım sayısı

1. GİRİŞ

Sosyal ağlar, varlıklar ve onlar arasındaki ilişkilerden oluşan sosyal yapılardır. Bu yapıda, yeni bağlantılar ve varlıklar zamanla görülür veya kaybolur. Bu durum, onları büyük ölçüde dinamik ve karmaşık sistemler yapar. Sosyal ağ analizi, sosyal ağların çalışması için yöntem ve stratejilerle ilgili araştırmanın geniş bir alanını içerir [1]. Bağlantı tahmini, sosyal ağ analizi ile ele alınan önemli bir problemdir. Bu iş ağda zaten mevcut etkileşimleri ve iki varlık arasındaki bir bağıntının varlığını tahmin etme problemiyle ilgilidir. Bu anlamda bağlantı tahmini bir sosyal ağdaki gelişimin nedenlerini açıklar.

Günümüze kadar bağlantı tahmini ile ilgili birçok yaklaşım önerilmiştir. Bu yöntemlerin çoğunluğu, düğümler arası yakınlığı gösteren ölçülere dayalıdır. Literatürde önerilen bu ölçüler anlamsal ve topolojik (yapısal) ölçümler olmak üzere iki sınıfa ayrılmıştır [2]. Anlamsal ölçülerde düğümlerin içeriği, yakınlığı değerlendirmek için göz önüne alınır. Örneğin, başlıklardan keşfedilen konular arasındaki benzerlik, ortak yazar ağında yazarlar arasındaki gelecek bağıntıları tahmin etmek için kullanılabilir. Anlamsal olanlardan farklı olarak, topolojik ölçüler yakınlık değerlerini elde etmek için, ağ yapısı kullanır. Topolojik ölçüler daha genel olduklarından, düğümün içeriği ile ilgili detaylı bilgiye gereksinim duymadıklarından daha yaygın olarak kabul görürler. Dahası bu içerik, göz önüne alınan sosyal ağda her zaman için mevcut olmayabilir. Topolojik ölçüler, komşuluk tabanlı ve yol tabanlı ölçüler olarak kategorize edilirler [3]. Komşuluk tabanlı ölçüler, düğümün anlık komşularını göz önüne alır. Bu ölçüye göre, eğer iki düğümün komşuları fazla bir çakışma gösterirlerse, onlar büyük bir olasılıkla bir bağlantı oluştururlar [2]. Düğümler arasında kullanılan ölçüler, ortak komşular [4], tercihi ilişkilendirme [5], Adamic-Adar indeksi [6] ve Jaccard indeksi [7] olarak yazılabilir. Yol tabanlı ölçüler, ilgili düğümler arasındaki yolları göz önüne alan bir yakınlık değeri üretir. Bu iddiaya göre, ilgili düğümler arasında kısa yollar varsa, iki düğümün bağlantı oluşturacağı muhtemeldir [8]. Bağlantı tahmini probleminin üstesinden gelmek için, birçok çalışma önerilmiştir [9-11]. Buna karşın gelecek zamandaki yeni ilişkileri tahmin etmek için, yapılan çalışmaların çoğu, mevcut zamanda ağdaki bağlantılı olmayan, düğüm çiftlerine yakınlık ölçülerinin uygulamaları üzerinedir.

Düğüm çiftleri arasındaki benzerliği gösteren yakınlık ölçüleri denetimsiz [9,10,12] veya denetimli [11,13,14] bağlantı tahmininden yararlanılabilir. Denetimsiz yöntemlerde

bir yakınlık ölçüsü seçilir ve ağdaki düğüm çiftleri sıralama için kullanılır. İyi skora sahip olan düğüm çiftlerinin bağlantılı olacağı tahmin edilir. Denetimli yaklaşımda bağlantı tahmini yöntemi, bir sınıflandırma işi olarak ele alınır ve ağın nitelikleri sınıflandırma yöntemi tarafından tahmin edici nitelikler olarak kullanılır.

Bir sınıflandırıcı, bağlantının gelecekte oluşup oluşmayacağını bulmak için ikili sınıflandırmayı gerçekleştiren bu nitelikleri kullanır. Yukarıda ifade edilen çalışmaların tamamında, yakınlık değerleri ağın gelişimini göz önüne almaksızın mevcut ağ durumunun bütün ağ verisini kullanarak hesaplar. Bundan dolayı, bilginin saklı ve muhtemelen değerli kaynağı bağlantı tahmini için yeterince göz önüne alınmaz.

Ardışıl zaman aralıklarında, düğüm çiftleri arasında meydana gelen özel durumlar vardır. Bu durumlar, düğümler arasındaki bağlantının oluşturulması ve silinmesi ile ilgilidirler. Örneğin, belirli zaman aralığında bağlı değilken, sonraki aralıkta iki düğüm arasında yeni bir bağlantı oluşabilir. Bu durum, ilgili sosyal ağın gelişimini temsil eder. Son zamanlarda bağlantı tahmini için, zaman aralıklarının kullanımı ile ilgili bazı çalışmalar yapılmıştır. Tylanda ve diğ. [15], ilgili düğümler arasında en son özel durumların yaşını gösteren link ağırlığı içeren bir ağ modeli önermiştir. Brigmann ve diğ. [16], ağın gelişimini açıklayan birliktelik kurallarının keşfi için geçici bilgili bir ağ önermiştir. Bağlantı tahmini için önerilen diğer yaklaşımlar işi, zaman serileri tahmin problemi olarak ele alır [17-19]. Bu çalışmalarda zaman seri modellerinde, yakınlık değerlerinin tahmininin doğruluğunu geliştirmek için kullanılmıştır.

1.1. Tezin Amacı ve Katkıları

Birçok farklı alan için bağlantı tahmini, anlaşılması doğrudan zor veya pahalı muhtemel ilişkileri tanımlamak veya gelecek davranışı tahmin etmek için kullanılır [20-22]. Bu alanlardan biri, tıbbi bakım araştırma alanıyla ilgili olanıdır. Tıbbi bakım alanının hastalık ve riskinin başlangıcını tanımlamada duyarlı olmaktan ziyade, daha ileriye yönelik olmaya ihtiyacı vardır [23,24]. Şu anda doktorlar, hastanın sağlık aşamasının daha ileriye yönelik tespiti için, laboratuvar sonuçlarını kullanırlar. Fakat böyle bir tespit, sadece bir kaç tıbbi parametreye (hastalık belirtisi) odaklanma veya doktorun tecrübe, hafıza veya zamanıyla bağlantılı olma dezavantajlarını içerir. Bundan dolayı, mevcut tıbbi gelişme ileriye dönük değildir ve bir hastalığı en erken safhada işaretleriyle ele alma veya yok etmede yeterli değildir. Yukarıda bahsedilen problemlere bir çare olarak, bu tezde

bireylerin hastalığa yakalanma riskini belirlemek ve hastalığın en erken işaretlerinde doğru hareketler yapmak için bir tahmin edici önerilecektir. Bu maksatla öncelikle, hastalığın anormal parametreleri arasındaki ilişkileri gösteren ağırlıklandırılmış bir tıbbi veri ağı oluşturulacaktır. Bu ağdaki her bir düğüm, hastaların anormal parametrelerini temsil eder. Kenarlar aynı hastalarda görülen bu parametreleri bağlar. Daha sonra farklı tür laboratuvar elde edilen sonuçlar toplanarak, herhangi bir hastalığa neden olabilecek bütün parametreler arasındaki ilişkileri tanımlamak için oluşturan tıbbi veri ağının gelişimine dayalı bir bağlantı tahmin yöntemi önerilmiştir. En son olarak, 210.000'den daha fazla hastanın laboratuvar sonuçlarıyla oluşturulan tıbbi veri ağı üzerinden önerilen yöntem test edilmiştir. Tezin sonraki bölümünde ise bu defa hastalıkları doğrudan ele alan bir hastalık ağı oluşturulmuştur. Bu ağda her bir düğüm, öncekinden farklı olarak farklı zamanlarda bile olsa aynı hastaların yakalandığı hastalıkları temsil eder. Yaşa göre gruplandırılmış ağ serilerinden, belirli yaş aralıklarında kişilerin hastalığa yakalanma riski yine bağlantı tahmini yöntemiyle elde edilmeye çalışılmıştır. Uygulama sonuçları, denetimli ve denetimsiz bağlantı tahmini adı altında sunduğumuz her iki yöntemin de doğruluğunu ve gelecek hastalık riskleri yakalamadaki başarısını gösterir.

1.2.Tezin Yapısı

Tezin bundan sonraki bölümleri aşağıdaki gibi düzenlenmiştir:

- Bölüm 2'de, sosyal ağlar ve sosyal ağ analizi üzerine kısa bilgiler verilmiş ve bu alanda yapılan çalışmalar sistematik olarak sıralanmıştır. Ayrıca, sosyal ağ analizinde kullanılan temel kavramlar tanıtılmıştır.
- Bölüm 3'de, öncelikle bağlantı tahmini probleminin tanımına yer verilmiştir. Daha sonra bağlantı tahmini için yapılan çalışmalar; çizge teorili yaklaşımlar, olasılıksal yaklaşımlar, benzerlik tabanlı yaklaşımlar ve diğer yaklaşımlar olmak üzere dört sınıfta kategorize edilerek bu alandaki çalışmalar hakkında bilgi verilmiştir.
- Bölüm 4'de, geçmişte bağlantı tahmini için yapılan makine öğrenmesi çalışmaları özetlenmiştir. Ayrıca, denetimli ve denetimsiz öğrenme yöntemlerinin anahtar özellikleri anlatılarak bağlantı tahmini için nasıl kullanılacakları belirlenmiştir. Bölümün sonunda ise tez boyunca geliştirilecek öğrenme yöntemlerinin performans değerlendirme ölçütlerine yer verilmiştir.

- Bölüm 5’de, öncelikle hastalık belirti ağı geliştirilmiş ve daha sonra bu ağ üzerinde önerilen denetimli ve denetimsiz bağlantı tahmini yöntemleri anlatılmıştır. Her iki yöntem için yapılan uygulama sonuçları da yine bu bölümde verilmektedir.
- Bölüm 6’da, hastalık ağı geliştirilmiş ve bu ağ üzerinde önerilen yaş serileri tabanlı denetimsiz ve denetimli bağlantı tahmini yöntemleri tanıtılmıştır. Bölüm ayrıca bu yöntemler için yapılan uygulama sonuçlarını da içermektedir.
- Bölüm 7’de, bu tezde elde edilen sonuçlar tartışılmış ve gelecek araştırmalar için bazı öneriler sunulmuştur.

2. SOSYAL AĞLAR

Son yıllarda, teknolojinin ve internetin gösterdiği hızlı gelişim toplum yapısını önemli ölçüde etkilemiştir. Özellikle internet, kısa bir geçmişe sahip olmasına rağmen dünya üzerinde yaygın bir kullanım alanı edinerek ayrıcalıklı bir konuma erişmiş bir kitle iletişim aracıdır. İnternetin gelişimi, beraberinde sosyal ağ kavramını ortaya çıkarmıştır. Sosyal ağ kavramı, bireyler arasında kişisel ve profesyonel ilişkilerin oluşmasını sağlamıştır. Bunun dışında insanların işbirliğini, çalışmasını ve öğrenme sürecini yeniden şekillendirmiştir. Kullanıcılar, sosyal ağlar üzerinden yeni iletişim ortamlarının sunduğu hemen hemen tüm özelliklerden yararlanmakta, sürekli bilgi paylaşımı sağlamaktadır. Bilgi değişiminin bu kadar yaygın hale gelmesi, her geçen gün yeni sosyal ağlar doğmasına yol açmaktadır. Günlük hayatın her alanında, birçok farklı sosyal ağ yapısı ile karşılaşılmaktadır. Örneğin; maddelerin atomları arasındaki bağlardan oluşan yapılar, insan vücudundaki sistemleri oluşturan ağ yapıları, canlılar ve çevre arasındaki bağlardan oluşan yapılar, toplumsal ve kültürel ağ yapıları, havayolu ağları ve internet ağları günlük yaşamda karşılaşılan ağlardan bazılarıdır.

Ağ kavramsal olarak “birçok nokta ile bunlar arasındaki bağlantılarla gösterilebilen bir dizgiye ilişkin yapı” olarak tanımlanabilir [25]. Sosyal ağ ise, düğümler olarak adlandırılan bireylerden veya elemanlardan oluşan ve bu düğümleri kendi aralarındaki çeşitli ilişki ve etkileşimlere bağlı olarak birbirine bağlayan bir yapı olarak ifade edilir [26].

Sosyal ağların sayısının her geçen gün artması insanların etkileşimini, işbirliğini, çalışmasını ve hatta öğrenme sürecini dahi değiştirmiştir. Bu değişiklikler, sosyal ağları kullananlara birçok olanak sağlamaktadır. Öğrenme ve öğretim deneyimlerini zenginleştirmek için yükseköğretimdeki öğrenciler, akademisyenler ve kurumlar sosyal ağların olanaklarını kullanmaktadır. Sosyal ağlar iletişim becerilerini geliştirir, katılımı ve sosyal bağlılığı genişletir, akran desteğini güçlendirir ve işbirliğine dayalı öğrenmenin gerçekleşmesini sağlar [27]. Bunun dışında, akademisyenlere akademik çalışmalarını yürütmek ve sonuçlarını paylaşmak için yeni imkanlar sunmaktadır. Sosyal ağlar sayesinde araştırmacılar, kendi çalışmalarını paylaşmak, yeni çıkan yayınları takip etmek, diğer araştırmacılarla işbirliği içinde çalışmalar yapmak gibi olanaklara sahip olabilmektedirler. Bu amaçla Facebook, Twitter gibi genel amaçlı sosyal ağların yanı sıra akademik çalışmaların paylaşıldığı sosyal ağlar da kullanılmaktadır. Bu ağlara Academia.edu,

Epernicus Network, ResearchGate, Methodspace, Sciencestage gibi siteler örnek olarak verilebilir (Şekil 2.1.).



Şekil 2.1. Sosyal ağ örnekleri

2.1.Sosyal Ağ Analizi

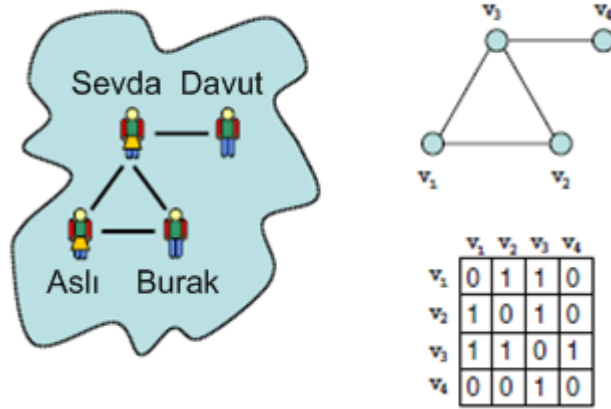
Sosyal ağlar, barındırdıkları sosyal aktörlerin ilişkilerine ait faydalı bilgiler içerirler. Bu yapı ve ilişkilerin analiz edilmesi ile yakınlıklar, benzeşmeler, eğilimler ve etkileşimler gibi verilere ulaşılarak ağdaki ilişkiler hakkında çeşitli yorumlara veya tahminlere varılabilir. Özellikle internet teknolojisinde yaşanan teknik ve kültürel gelişmeler, sosyal ağlara ilişkin verilerin somutlaştırılması ve ölçülebilmesi için sosyal ağ analizi konusunu gündeme getirmiştir. Sosyal ağ analizi bireyler, nesneler ya da kurumlar arasındaki ilişkilerin sayısallaştırarak sunulmasını sağlayan bir yöntemdir. Ayrıca sosyal ağın karakteristik yapısının gözlemlenmesi ve ilişkili öğelerin analiz edilmesini sağlar.

Sosyal ağ analizi sosyoloji, antropoloji, sosyal psikoloji, iletişim, ekonomi, matematik gibi birçok alanda sıklıkla kullanılan bir yöntemdir. Disiplinler arası bir çalışma alanı olan sosyal ağ analizi ile toplumların yapısı incelenmekte, ağ yapısı betimlenmeye çalışılmakta, toplumlar arasında kolayca gözlemlenemeyen ilişkiler görselleştirilerek var olan bağlantılar modellenmektedir. Sosyal ağ analizine dayalı yaklaşımda, toplum içindeki bağlantı ve ilişkilere odaklanılmaktadır.

Temel olarak sosyal ağların analizi aktör, ilişki ve bağ olarak isimlendirilen 3 eleman ile yapılmaktadır. Aktörler, bir sosyal ağın temel elemanıdır ve düğüm ile gösterilir. Tüm düğümlerin ilişkileri, bir diyagram ile gösterilir. Bu diyagramda çizgi ve düğümler, sırasıyla aktör ve ilişkileri gösterir. İlişkiler ise içerik, yön, yoğunluk ve aktif/pasif ilişkiler olmak üzere 4 gruba ayrılır. İçerik, iki aktör arasındaki ilişkiyi gösterir. Yön ise yönlü ve yönsüz olarak sınıflandırılır. Yoğunluk ise zamanla ifade edilir. Buna örnek olarak, bir okulda okuyan öğrenciler arasında iki yıldır süren ilişki verilebilir. Aktif ve pasif ilişkiler ise ilişkinin türünü gösterir. Sosyal ağın analizi için önemli olan bağlar ise güçlü ve zayıf olmak üzere ikiye ayrılır. Güçlü ilişkiler yakın, özel ve aktif ilişkilerdir. Zayıf ilişkiler ise az sayıda ilişki ya da iletişim bilgisi içerirler [28].

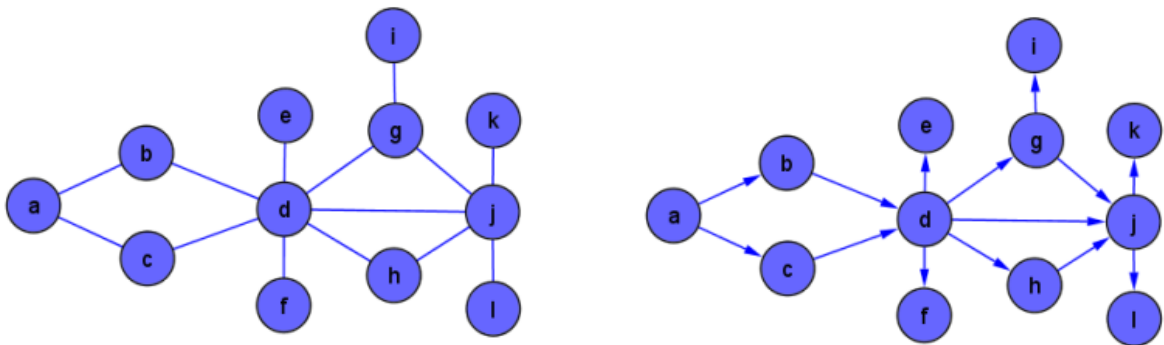
Sosyal ağlar, bilgisayar ortamında çizge olarak temsil edilebilirler. Genel çizge gösterimi $G = (V, E)$ şeklindedir. Bu gösterimdeki V sonlu düğüm kümesini ve E sonlu kenar kümesini belirtir. Her bir kenar, kendisi ile ilişkilendirmiş bir ya da iki adet düğüm içerir ve bu düğümleri birbirine bağlar. Sosyal ağlardaki kullanıcılar (insan, grup, kurum vb.) çizgede düğüm olarak temsil edilirken, sosyal ilişkiler (arkadaşlık, profesyonel iş ilişkisi vb.) çizgede kenar olarak temsil edilirler. Bazı çizgelerde, iki düğüm arasında birden fazla kenar bulunabilir. Eğer bir $E = \{U, V\}$ kenarı varsa U ve V düğümlerinin komşu olduğu söylenir. Böylesi bir durumda U ve V , E 'nin uç noktaları olarak adlandırılır ve E 'nin U ve V 'yi bağladığı söylenir. Şekil 2.2'de gösterildiği gibi V kümesindeki her v düğümü bir nokta (ya da küçük çember) ile temsil edilir ve her $E = \{v_1, v_2\}$ kenarı, v_1 ve v_2 uç noktalarını bağlayan bir çizgi ile gösterilir.

Çizge yapısının temsil edilmesi için genel olarak iki farklı yöntem tercih edilebilir. Bunlardan ilki, komşuluk listesi adı verilen, ağdaki her bir düğüm için komşu ya da bitişik olduğu düğümlerin listesinin dizi şeklinde gerçekleştirilmesidir. Bir düğümün komşuları liste dolaşarak tespit edilir. İkinci yöntem ise sık kullanılan komşuluk matrisi yöntemidir. Bu yöntemde çizgedeki n tane düğüm için oluşturulan $n*n$ boyutundaki matris üzerinde birbirine komşu olmayan düğümlerin girdileri 0, komşu düğümlerin girdileri 1 olarak atanır. Matris işlenirken de her bir girdinin değerine bakılarak, karşılık geldiği satır ve sütundaki düğümler arasında kenar (bağlantı) olup olmadığına karar verilebilir.



Şekil 2.2. Düğümler, kenarlar ve kenarların düğüm kümeleri

Aynı uç noktalarını bağlayan çoklu kenarla veya uç noktaları tek ve aynı düğüm olan döngüler barındıran çizgelere çoklu çizgeler denir. Çizgelerin resmi tanımları ne çoklu kenarlara ne de döngülere izin verir. Bazen, çizge terimine yüklenen anlam çoklu çizgeleri de içerir ve çoklu kenar ve döngü içermeyen çizgeler için basit çizge terimi kullanılır. Bir kenar bir düğümü kendisine de bağlayabilir. Bu tür döngüler çizge içerisinde yer alabilir. Şekil 2.3' de gösterildiği gibi kenarlar yönlü de olabilir. Yönlü çizgeler adı verilen bu çizge yapılarında, kenarların başlangıç ve bitiş noktaları bellidir. Yönlü ve yönsüz kenarların birlikte bulunduğu karışık çizgelerde mevcuttur. Bu tür çizge özellikleri daha da sıralanabilir. Günlük hayatta pek çok farklı alanda kullanılan çizge modelleri, sosyal ağları temsil etmek için kullanılabilir [29].

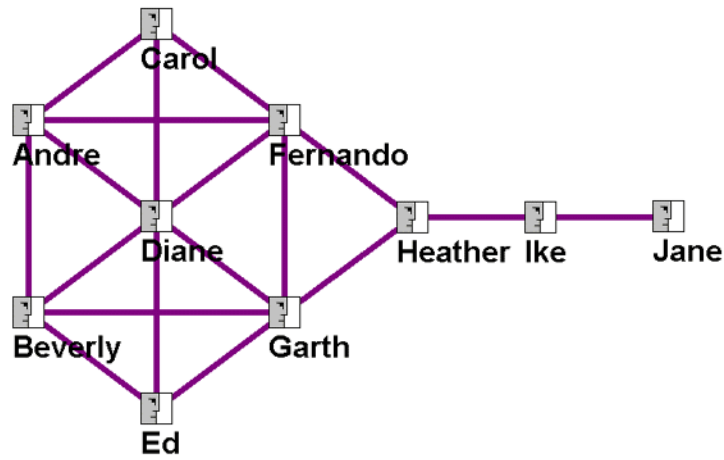


Şekil 2.3. Yönlü ve yönsüz çizgeler

2.2.Sosyal Ağ Analizinde Kullanılan Temel Kavramlar

2.2.1. Derece Merkeziliği

Derece merkeziliği bir sosyal ağda hangi aktörlerin güçlü, etkili ve kritik olduğunu anlamak için kullanılan bir ölçüt birimidir. Derece merkeziliği yüksek olan bireyler popüler aktif bireylerdir. Bir düğümün derecesi o düğümün bağlantılı olduğu düğüm sayısıdır. Şekil 2.4’de derece merkeziliği en yüksek olan düğüm *Diane* düğümüdür.



Şekil 2.4. Kite ağı [30].

2.2.2. Yakınlık Merkeziliği

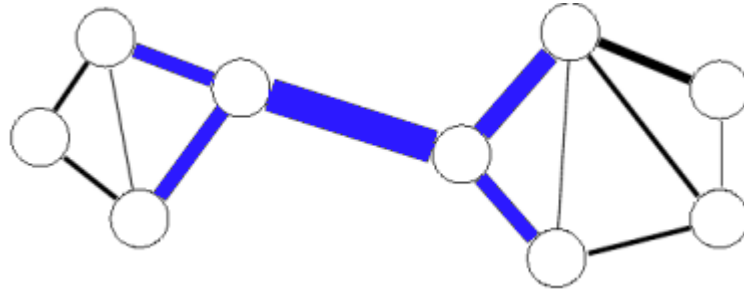
Yol uzunluğu, bir düğümün diğer tüm düğümlere olan en kısa yol uzunluklarının ortalaması alınarak hesaplanır. Bu ölçüt birimi, düğümün diğer tüm düğümlere yakınlık değerini gösterir. Tüm düğümlere uzaklığı en kısa olan düğüm, ağın yakınlık derecesi en yüksek olan düğümüdür. Bu düğümler, ağdaki bilgi akışını en iyi gözlemleyebilen düğümlerdir. Şekil 2.4’te Fernando ve Garth düğümleri yakınlık derecesi en yüksek olan düğümlerdir.

2.2.3. Arasındalık Merkeziliği

Freeman [31] tarafından bulunmuştur. Ağdaki grupların kesişim noktasındaki düğümleri belirlemek için kullanılır. Ağda, her bir düğümün tüm düğümlere olan en kısa yol uzunlukları hesaplanır. Bu hesaplamalar sonucunda, her bir düğümden kaç defa geçildiği bilgisi arasındalık değerini verir. En yüksek değere sahip olan düğüm, arasındalık değeri en iyi olan düğümdür. Şekil 2.4’de Heather düğümü arasındalık değeri en yüksek olan düğümdür.

2.2.4. Kenar Arasındalık Merkeziliği

Girvan ve Newman [32], Freeman’nın [31] düğümler için önerdiği arasındalık kavramından esinlenerek bu kavramı kenarlar için geliştirmiştir. Arasındalık merkeziliğine benzerdir. Arasındalık merkeziliğinde düğümler esas alınırken, kenar arasındalık merkeziliğinde kenarlar esas alınır. Her bir düğüm için diğer tüm düğümlere olan uzaklıklar hesaplanır. Bu hesaplama sona, kenarlardan kaç defa geçildiği bilgisi kenar arasındalık merkeziliği değerini verir. Şekil 2.5’de, kenar arasındalık merkeziliğini gösteren örnek bir ağ gösterilmiştir. Bu ağda kenarların kalınlıkları, kenar arasındalık değerini simgeler. Topluluk yapılarının olduğu ağlarda toplulukları birbirlerine bağlayan kenarlar, en kısa yollar üzerinde en çok geçilen kenarlardır. Bu ölçüt birimi, sosyal ağlarda toplulukları keşfetmek için kullanılmıştır. En yüksek kenar arasındalık değerine sahip kenarlar, toplulukları birbirlerine bağlayan kenarlardır.



Şekil 2.5. Kenar arasındalık merkeziliğini gösteren örnek bir ağ [33].

Kenar arasındalık değerinin hesaplanmasında, başlangıç düğümünden itibaren sıg öncelikli arama (breadh-first search) yapılır. Algoritmanın adımları şöyledir [34] :

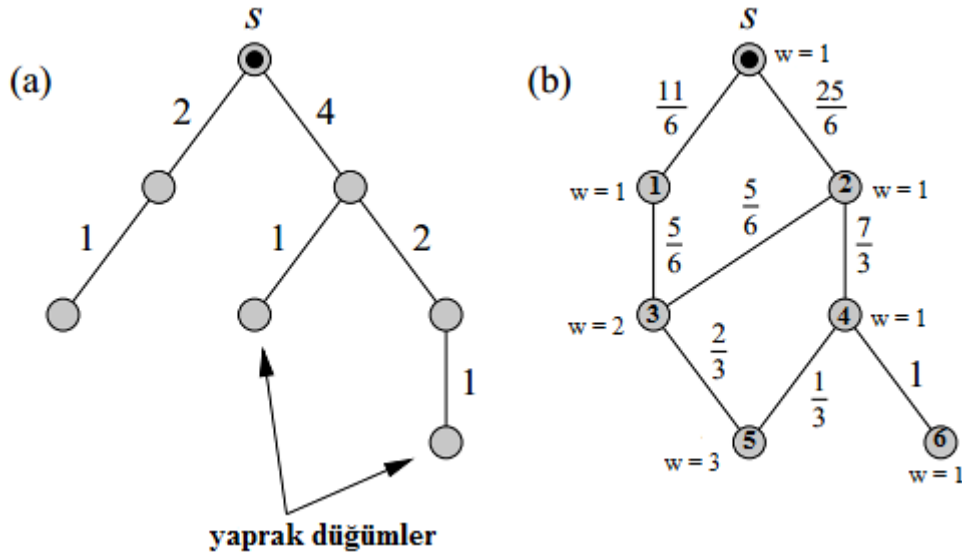
- 1) Başlangıç düğümü olan s düğümünün uzaklık değerine $d_s = 0$ ağırlığına $w_s = 1$ verilir.
- 2) s düğümünün komşusu olan her i düğümünün uzaklığı $d_i = d_s + 1 = 1$, ağırlığı ise $w_i = w_s = 1$ olarak verilir.
- 3) Bu i düğümlerinin komşuları olan her j düğümü için aşağıdaki üç adım uygulanır:
 - a. Eğer j düğümü için bir uzaklık atanmamışsa uzaklık için $d_j = d_i + 1$ ağırlık için $w_j = w_i$ atanır.
 - b. Eğer j için bir uzaklık atanmış ve $d_j = d_i + 1$ ise j düğümünün ağırlığı $w_j \leftarrow w_j + w_i$ olarak belirlenir.
 - c. Eğer j düğümünün bir uzaklığı varsa ve bu da $d_j < d_i + 1$ koşulunu sağlıyorsa herhangi bir şey yapılmaz
- 4) Ağda uzaklığı atanmamış hiçbir düğüm kalmayana kadar 3. adım tekrarlanır.

Bu adımlar sonucunda bulunan her bir i düğümünün ağırlığı, s düğümünden i düğümüne kaç tane farklı yol olduğunu gösterir. Bu aşamadan sonra kenar ağırlık değerini hesaplamak için şu adımlar uygulanır:

- 1) Ağdaki her bir yaprak düğüm (t) bulunur. Yaprak düğümler bir önceki aşamada ağırlıklar hesaplanırken ağın en altında kalan düğümlerdir. s düğümünden başka düğümlere gidebilmek için t düğümlerinden geçilen hiçbir yol yoktur.
- 2) t düğümünün komşusu olan tüm i düğümlerinin kenar değeri t den i 'ye doğru w_i / w_t 'dir.
- 3) t düğümlerinin kenar değerleri hesaplandıktan sonra, yukarıda kalan kenarların değerleri o kenara komşu olan aşağıda kalan kenarların değerleri toplamına 1 ilave edilerek hesaplanır.
- 4) Başlangıç düğümü olan s düğümüne ulaşınca kadar 3. adım tekrarlanır.

Şekil 2.6.'da kenar arasındalık değerinin hesaplanması için Newman ve Girvan'ın [34] kullandığı örnek iki ağ verilmiştir. Şekil 2.6 a'da başlangıç düğümünden diğer tüm

düğümlere olan en kısa yollar sadece bir tanedir. Alternatif kısa bir yol hiçbir düğüm için yoktur. Böyle bir ağda kenar arasıdalık değerini kolay bir şekilde hesaplanıyor. Şekil 2.6 b’de ise bazı düğümler için farklı kısa yolların olduğu bir ağ vardır. Yukarıda verilen algoritma adımlarına göre, önce ağırlıklar hesaplanmıştır. Daha sonra yaprak düğümlerden yukarıya doğru kenarların arasıdalık değerleri bulunmuştur. 5 ve 6 numaralı düğümler yaprak düğümlerdir. Yukarıya doğru ilk kenarların değerleri komşu düğümünün ağırlığının yaprak düğümlerinin ağırlığına oranı ile bulunur. 4-5 kenarının değeri $1/3$, 4-6 kenarının değeri ise 1 ’dir. 2-4 kenarı hesaplanırken 4-5 ve 4-6 kenar değerleri toplanır ve 1 ilave edilir o da bize $7/3$ değerini verir.



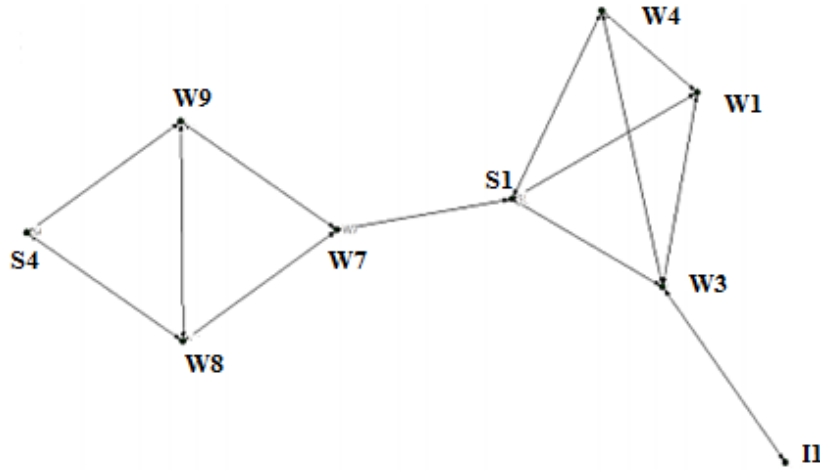
Şekil 2.6. Kenar arasıdalık değerinin hesaplandığı örnek iki ağ [34].

2.2.5. Özvektör Merkeziliği

Bir düğümün bir ağdaki etki derecesini ölçmek için kullanılır. Derece merkeziliği yüksek olan düğümler, aktif düğümler olarak tanımlanmıştır. Düğümün özvektör derecesi bunu daha nitelikli bir şekilde ölçer. Bir düğümün ağ içinde etki derecesini gösterir. Şekil 2.7’de örnek bir ağ için özvektör, derece, yakınlık ve arasıdalık merkeziliği değerleri gösterilmiştir. En yüksek derece merkeziliğine sahip düğümler W3 ve S1 iken, en yüksek özvektör merkeziliğine sahip düğüm S1’dir.

Necmi Gürsakal Sosyal Ağ Analizi kitabında özvektör merkeziliği için şöyle yazmıştır [35].

“... özvektör merkeziliği, bağlantıların sayısına olduğu kadar, kalitesine de bağlıdır. Eğer bir düğümün az sayıda yüksek kaliteli bağlantısı varsa, bu düğümün çok sayıda ortalama sahip bir düğümü özvektör merkeziliği anlamında geçebilir. Arama motoru Google’ın web sayfalarını sıralarken kullandığı PageRank, bu bağlamda çalışmaktadır ”.



Özvektör merkeziliği	Derece merkeziliği	Yakınlık merkeziliği	Arasındalık merkeziliği
S1 (.498)	W3, S1	S1	S1
W3 (.472)	W9, W8, W7, W1, W4	W7	W7
W1, W4 (.438)	S4	W3	W3
W7 (.254)	I1	W1, W4	W8, W9
W8, W9 (.159)		W8, W9	W1, I1, W4, S4 (0)
I1 (.147)		I1	
S4 (.099)		S4	

Şekil 2.7. Özvektör, derece, yakınlık ve arasındalık merkeziliği gösterilen örnek bir ağ [36].

2.2.6. Diğer Kavramlar

Köprü: Bağlantının kaldırılması diğer uç noktalara (düğümlere) ulaşılmamasına sebep oluyorsa bu bağlantıya köprü diyebiliriz. Ağ grupları arasındaki tek bağlantıdır. Kaldırılması bu gruplar arasındaki bağlantıyı keser.

Yol Uzunluđu: Bir düğümün diğeri düğümle arasındaki mesafeyi ölçmektir. Eğer bu iki düğüm arasındaki mesafe kısaysa bu düğümler arasındaki düğüm/kenar sayısı daha az demektir. Birbirine yakın düğümlerin ilişki kurma ihtimali daha fazladır.

Kümelenme katsayısı: Kümelenme katsayısı bir aktörün, komşuları ile arasındaki bağlantının ne derece güçlü olduğu hakkında bilgi verir. Bir aktörün komşularının kendi aralarındaki, gerçek bağlantıların olası tüm bağlantılara oranıyla hesaplanır.

Bağlılık: Hangi düğümlerin birbirleriyle güçlü bağlar oluşturduğunu ölçer. Güçlü bağlantılar oluşturmuş düğümlerin benzerliği daha fazladır.

Yoğunluk: Ağda var olan tüm bağlantıların, olası tüm bağlantılara oranı yoğunluk ölçütünü verir. Tüm aktörlerin birbirine bağlı olduğu ağlarda yoğunluk değeri 1 olacaktır. Yoğunluk değerinin yüksek olması ağdaki aktörlerin birbirlerine daha güçlü bağlı olduklarını ve birbirlerinden soyutlanmış durumda olmadıklarını gösterir.

3. BAĞLANTI TAHMİNİ

Gerçek hayatta kişiler arasındaki etkileşimin miktarı arttığı için kullanışlı sosyal veriler sosyal ağ analizi için temel oluşturmaktadır. Sosyal ağlar, farklı sosyal ilişkiler ile bağlanabilen kişilerden oluşan yapılardır. Sosyal ağlarda bağlantılar ve varlıklar zaman boyunca kaybolma ya da oluşma eğilimindedir. Sosyal ağ analizinde en bilinen problem bağlantı tahminidir. Bağlantı tahmini bir sosyal ağda oluşabilecek bağlantıları tahmin eden veya gizli bağlantıları tespit eden, sosyal ağ analizinin en bilinen görevlerinden biridir. Yakın gelecekte bağlantıların görünümünü tahmin etmek için verilen bir zamanda ağın yapısını kullanır. Bu yapı bibliyografik alan, moleküler biyoloji, suç soruşturmaları ve tavsiye sistemleri gibi birçok farklı alanda sıklıkla kullanılmaktadır.

3.1.Bağlantı Tahmini Problemi

Sosyal ağlarda bağlantı tahmini problemi, ağın gelecekteki yapısıyla ilgili önsezilerde bulunulmasıdır. Bağlantı tahmini yöntemleri sosyal ağdaki düğümlerin ve ilişkilerin özellikleri kullanılarak geliştirilmiştir. Ağdaki veriler düğümlerle, ilişkiler ise bağlantılarla tanımlanarak çizge yapısında ifade edilmektedir.

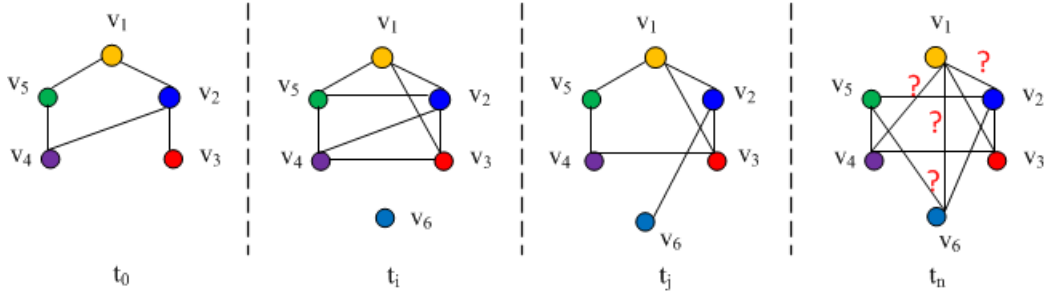
Sosyal ağlarda her düğüm vektörel yapıya sahip verilerdir. Veriler çeşitli ön işleme adımlarından geçtikten sonra tablolara dönüştürülür. Oluşturulan tablolarda satırlar değerleri, sütunlar ise düğümlerin özelliklerini belirtmektedir. Her bir düğümün niteliklerinin bilindiği ağlarda daha oluşmamış bağlantılar tahmin edilebilmektedir. Düğümler arası ilişkiler ve düğümlerin özellikleri kullanılarak düğümler arasındaki bağlantılar tahmin edilebilmektedir. Ağda ileride yeni ilişkiler oluşma ve ağa yeni düğümlerin katılma ihtimallerinin yanı sıra ağdan düğümlerin çıkma veya ilişkilerin eksilme ihtimalleri de bulunmaktadır. Ağlar dinamik yapıya sahip olduğu için bu tahminlerin yapılması oldukça zordur. Var olan ağ yapısının resmi çekilmektedir. Yakın gelecekte bu ağa yeni bireylerin dâhil olup olmayacağının, yeni bağlantıların oluşup oluşmayacağı veya var olan bağlantıların ileriki zamanlarda da devam edip etmeyeceğinin hesaplanması gerekmektedir. Ağdaki bilgilerin nasıl tanımlanacağı da önemli bir sorundur.

Var olan bilgilerin nasıl kullanılacağına, doğru hesaplamalar (tahminler) yapmada ne kadar etkili olacağı iyi düşünülmelidir.

Bağlantı tahmininde ağın topolojik özelliklerinin yanında ağdaki düğümlere ve ilişki ait özellik türlerine de ihtiyaç duyulmaktadır. Karmaşık olmayan seyrek ağlarda bu bağlantı tahmini yapılırken bu bilgiler fayda sağlamaktadır. Bu şekilde daha verimli sonuçların alınacağı tahminler yapılmaktadır. Bazı durumlarda ise veri setleri dengesiz durumlar ortaya çıkabilmektedir. Düğüm sayıları ve bağlantı sayıları arasında tutarsızlıklar oluşabilmektedir. Bu durum dengeli verilerde kullanılan birçok modelin etkinliğini engellemektedir. Geniş sayıdaki düğümler ve/veya kenarlar çizgede ölçeklenirse, hesaplama yapan modellerin verimli olması şarttır. Yani oluşturulan modelin büyük sistemlerde kullanılmasının zorluğu ön plana çıkmaktadır [37].

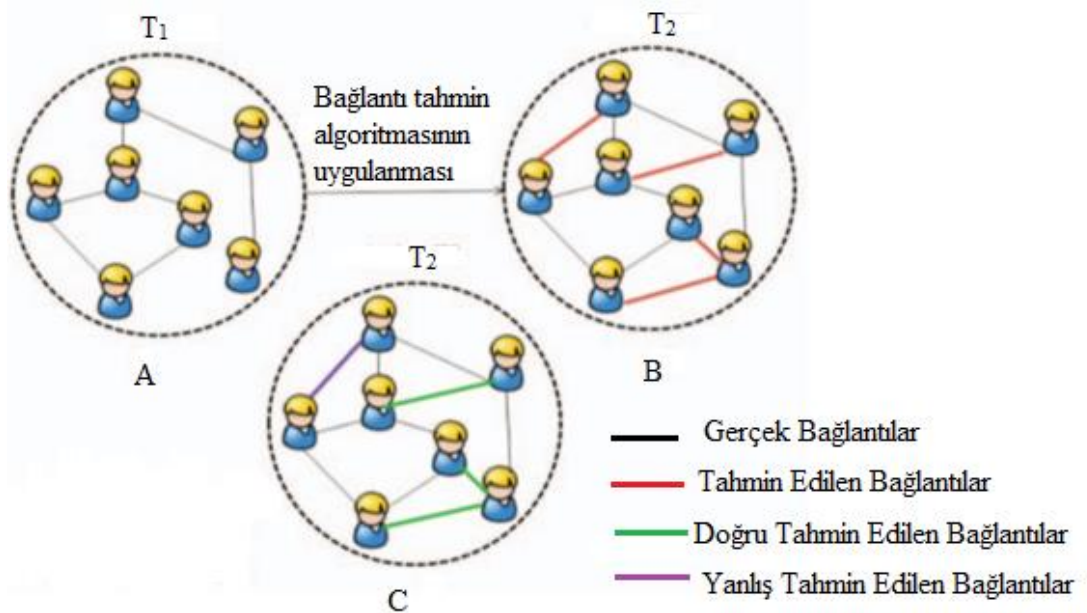
Bağlantı tahmini yapılırken bağlantının oluşumu, türü ve ağırlığı keşfedilip buna göre tahminin doğruluğu analiz edilmelidir. Bağlantı tahminiyle gelecekte sosyal ağda iki düğüm arasında bir ilişki kurulup kurulmayacağının tahmini yapılmaya çalışılmaktadır. Bu konuda daha çok çalışma yapılmasının sebebi ise bağlantının ağırlığı ve bağlantının önemi problemlerine kadar uzaması ve bunları da bir anlamda içinde barındırmasıdır. Bağlantının ağırlığı bağlantının ölçümünü, etkisini hesaplamaya çalışmaktadır. Bağlantının önemi benzer düğüm çiftleri arasında birden fazla bağlantı olup olmadığına odaklanmaktadır. Bağlantı türü ise iki nesne arasındaki ilişkinin türünü, rolünü anlamaya dayanmaktadır [38].

Bağlantı tahmini probleminde, düğümler V (veri örnekleri) olarak düşünülürse $V = \{v_i\}_{i=1}^n$, E ise bu veriler üzerinde var olan ilişkileri göstermektedir. Buna göre sosyal ağ $G = (V, E)$ olarak tanımlanabilmektedir. Buradan (v_i, v_j) düğüm çiftleri olacaktır. Bu düğüm çiftleri arasında, $e_{ij} \notin E$ oluşmamış bağlantının tahmini yapılmaya çalışılmaktadır. Düğüm çiftleri arasındaki bağlantının önemini gösteren ölçüt de $skor(x, y)$ fonksiyonu olarak tanımlanabilmektedir. Daha sonra bu fonksiyon değerleri büyükten küçüğe sıralanmaktadır. Bu şekildeki hesaplamalar ve ölçümlerle iki düğüm arasındaki benzerlik veya yakınlık bulunabilmektedir. İki düğüm arasındaki en kısa yol veya bağlantı durumlarına göre ilerde bağlantı oluşma ihtimali gibi hesaplamalar yapılabilmektedir [39]. Şekil 3.1’de oluşan ağın zamansal değişimi yapılan ölçümlerle karşılaştırılarak tahminin doğruluğu ölçülebilmektedir.



Şekil 3.1. Sosyal ağın zamansal değişimi

Bağlantı tahmini, t_1 anında anlık görüntüsü verilen bir sosyal ağın, t_1 anından verilen herhangi bir t_2 anına kadar ağa eklenebilecek düğümlerin tahmin edilmesi olarak tanımlanabilir [40]. Şekil 3.2 bir sosyal ağda yapılan bağlantı tahminini göstermektedir. A kümesi T₁ anında ağda bulunan bağlantıları siyah çizgiler ile göstermektedir. B kümesinde bağlantı tahmini sonucu T₂ anında oluşabilecek bağlantılar kırmızı çizgi ile gösterilmektedir. C kümesinde bulunan bağlantılardan yeşil ile gösterilenleri doğru tahmin edilmişken mor ile gösterilenleri yanlış tahmin edilmiştir. Yapılan tahminlerin doğruluğu kullanılan algoritmaya bağlı olarak değişiklik göstermektedir [41].



Şekil 3.2. Bağlantı tahmini [41].

3.2.Bağlantı Tahmini ile İlgili Yapılan Çalışmalar

Sosyal ağ verisi için bağlantı tahmini sosyal ağ analizi, bilgi kazanımı, tavsiyeci sistemler, pazarlama ve biyoinformatiği de içeren farklı uygulama alanlarında temel bir veri madenciliği işidir. Bu bölümde sosyal ağlar üzerine odaklanan bağlantı tahmini araştırmasının bazı temel çalışmaları incelenecektir. Çizge teorili yaklaşımlar, Olasılıksal yaklaşımlar ve Benzerlik tabanlı yaklaşımlar gibi farklı perspektif ve yöntemler üzerinde durularak bağlantı tahmini algoritmalarıyla ilgili en son ilerlemeler özetlenecektir. Bu yöntem ve algoritmalar gelecek muhtemel bağlantıları anlamada kullanılabilecek sosyal ağların gelişim mekanizmaları hakkında fikir sahibi olmamıza da yardımcı olur. Bölümün sonunda, birçok tahmin yöntemleriyle ağların dinamik ve geçici davranışlarını ele alma yetersizlikleri ana hatlarıyla belirlenecektir.

3.2.1. Çizge Teorili Yaklaşımlar

Çizge teorisi veya ağ teorisi çizgelerin veya ağların yapısını modellemek için kullanılan matematiksel bir yaklaşımdır. Matematiksel literatürde ağ, bağlantılarla birleştirilmiş düğümler topluluğudur. Matematiksel modeller mevcut ağın gelecek formunu tahmin etmek için bağlantı tahmini araştırmasında yaygın bir şekilde kullanılır. Çizge teorisi, çizge topolojisi olarak adlandırılan ağların yapısal örüntüleri üzerinde inşa edilmişlerdir. Bir ağın Kümelenme katsayısı, En kısa yollar, Ortalama yol uzunluğu, Arasındalık merkeziliği, Yakınlık merkeziliği derece dağılımı gibi topolojik özellikleri ağın gelişim modellerinin temellerini çıkarmada kullanılabilir. Yukarıda verilen bazı temel topolojik özelliklerinin tanımları aşağıda verilmiştir.

- Kümelenme Katsayısı C aşağıdaki gibi tanımlanır:

$$C = \frac{3 * \text{ağdaki üçgenlerin sayısı}}{\text{Düğümlerdeki bağlantılı üçlülerin sayısı}} \quad (3.1)$$

- En kısa yollar çizge teorisinde temel bir kavramdır. Verilen iki düğümü bağlayan kenarların en kısa yoludur. İki düğüm arasında yegane bir jeodezik uzaklık olmayabilir. Bir düğüm çifti iki veya daha fazla en kısa yola sahip olabilir.

Ortalama yol uzunluğu, kümelenme katsayısı ve onun derece dağılımıyla birlikte ağ topolojisinin en güçlü üç ölçüsünden biridir. n düğümlü ağırlıksız bir çizge göz önüne alındığında $d(i, j)$, i ve j düğümleri arasındaki en kısa yolu gösterebilir. Eğer $i \neq j$ ise o zaman ortalama yol uzunluğu l_G aşağıdaki gibi tanımlanır:

$$l_G = \frac{1}{n(n-1)} \sum_{i,j} d(i, j) \quad (3.2)$$

- Arasındalık merkeziliği bir düğüm boyunca en kısa yol geçişlerinin sayısını ölçer.
- Yakınlık merkeziliği bir düğümün ağdaki diğer düğümlere ne kadar yakın olduğunun bir ölçüsüdür. G bir çizge ve $i, j \in G$ düğümler olsun. Düğümler arasındaki en kısa yol $d_s(i, j)$ ise i 'nin yakınlık merkeziliği C_c aşağıdaki gibi tanımlanır:

$$C_c = \frac{1}{\sum_{j \in G0} d_s(i, j)} \quad (3.3)$$

- Derece dağılımı bütün ağ üzerinden bu derecelerin olasılık dağılımıdır.

Yukarıdaki topolojik özelliklerden başka daha önce tanımlanmış değişik yapısal özellikler de vardır. Onların çoğu temel topolojik özelliklerin farklı türleri veya geliştirilmiş halleridir. Newman ve diğ. [42] yaygın bir şekilde bilimsel iş birliği ağlarını kullanarak sosyal ağ gelişimi üzerine çalışmışlardır. Modern sosyal ağların gelişiminde kullanılan güçlü çizge üretim modelleri yapmak için rasgele çizge modellerine kümelenme katsayısı gibi topolojik örüntüleri dahil eden yeni modeller önermişlerdir [43]. Bu araştırmanın sonuçları, çoğu bilimsel işbirliği ağlarının gelişimi yukarıda ifade edilen temel çizge teorilerine uyduğunu gösterir [44]. Dahası birlikte yazarlıklar kullanarak ağırlıklı çizgelerin oluşturulması için bir yöntem geliştirmişlerdir. Bu yeni yöntemler ağ gelişimi için daha karmaşık modellerin oluşturulmasına neden olur.

Son zamanlarda internet ve online sosyal ağlar üzerine araştırmanın bir sonucu olarak araştırmacılar arasında ağ yapısındaki merak uyandıktan sonra, araştırmanın başka bir dalı, ağları analitik ve nümerik olarak modellemek için yöntemler ve ağların istatistiksel özelliklerini araştırır. Bu çalışmalardan oluşan önemli ve temel bir sonuç düğümlerle diğer

düğümler arasındaki bağlantıların sayısıdır. Bu sayı derece olarak ta adlandırılır. Birçok ağda düğüm derece dağılımının son derece çarpık olduğu belirlenmiştir [45]. Normalden daha fazla bağlantıya sahip bazı düğümler olabilir. Deneysel çalışmalar derece dağılımına dayalı rasgele çizge modellerinin sayısını önermiştir. Erdős ve Renyi'nin modeli onlar arasında tartışmasız en ünlü olanıdır. Bu rasgele çizge modelinin tanımı basittir. Düğümlerin n tane sayısı alınır ve onlar arasında bağlantılar veya kenarlar yerleştirilir. Böylece her bir i, j düğüm çifti bağımsız p olasılığıyla bir bağlantıya sahip olur. n düğüm sayılı rasgele bir çizgede bir i düğümü göz önüne alındığında, bu düğüm çizgedeki diğer $n - 1$ düğümün her biriyle eşit p olasılığıyla bağlıdır. Böylece i 'nin k derecesine sahip olma olasılığı aşağıdaki binomial dağılımla verilir:

$$p_k = \binom{n-1}{k} p^k (1-p)^{n-1-k} \quad (3.4)$$

Buna karşın gerçek dünya probleminin bir modeli olarak ciddi kusurlara sahiptir. Belki de en önemli olanı derece dağılımıdır. Genelde doğru olmayan bağımsız kabul nedeniyle çoğu gerçek dünya problemlerindekinden oldukça farklıdır. Tercihli bağlılığa dayalı iyi bilinen Barabási-Albert modeli [45] rasgele ölçekten bağımsız ağları üretmek için kullanılan bir algoritmadır. Ölçekten bağımsız ağlar, internet, WWW, atıf ağları ve bazı sosyal ağları da içeren doğal ve insan yapımı sistemlerde yaygın bir şekilde gözlenir. Barabási-Albert modelinde düğüm derecesiyle ağırlıklandırılmış bir olasılık dağılımı kullanarak düğümlere yeni bağlantılar ilişkilendirilir. Yeni düğümler ağa her bir zamanda eklenir. Her bir yeni düğüm, var olan düğümlerin zaten sahip olduğu bağlantıların sayısı ile orantılı bir olasılıkla var olan düğümlere bağlanır. Matematiksel olarak yeni bir düğümün i düğümüne bağlanma olasılığı;

$$p_i = \frac{k_i}{\sum_j k_j} \quad (3.5)$$

dir. Burada k_i , i 'ninci düğümün derecesi, toplam bütün önceden var olan j düğümlerin derecesi üzerinden yapılır (Yani payda ağdaki kenarların mevcut sayısıdır). Sadece birkaç bağlantılı düğümler veya daha düşük derece yeni bir bağlantı için hedef olarak seçilmesi muhtemel değilken, daha yüksek derece hızlı bir şekilde daha fazla bağlantıyı toplama

eğilimindedir. Yeni düğümler zaten ağırlıklı bağlantılı veya daha yüksek dereceli düğümlere kendilerini bağlama önceliğine sahiptir.

Clauset ve diğ. [46] tarafından önerilen diğer bir model güç yasası dağılımına dayalıdır. Bu model ölçekten bağımsız ağları modellemek için kullanılır. Güç yasası bir tip olasılık dağılımıdır. Yani, eğer bir olayın oluşma sıklığı o olayın bir niteliğinin gücüyle (boyutu) değişirse sıklığın bir güç yasası olduğu söylenir. Ağ kapsamında sıklık düğümlerin sayısıdır ve nitelik düğümlerin derecesidir. Düğümlerin sıklığı düğüm derecesi yükselirken güç yasasına göre azalır. Ölçekten bağımsız ağlar geniş merkezlerin varlığıyla karakterize edilmiş bir ağ tipidir. Yani, bağlantı sayısı çok fazla olan birkaç düğüm vardır. Yönsüz bir ağ için derece dağılımı aşağıdaki gibi yazılabilir:

$$P_{der}(k) \propto k^{-\gamma} \quad (3.6)$$

Burada k derecesi yükselirken çok daha geniş bir dereceli bir düğümü bulma olasılığı yükseldiğinden $P_{der}(k)$ formu yavaşça azalır. Çizge teorili yaklaşımların çoğu gözle görülebilir seviyede çizge gelişim mekanizmasını modellemek için denenmişlerdir. Buna karşın olasılık ve benzerlik tabanlı yöntemleri gibi bazı diğer yaklaşımlar ise ağ gelişimini mikroskobik seviyede tanımlayabilmişlerdir.

3.2.2. Olasılıksal Yaklaşımlar

Olasılıksal yaklaşımlar temel olarak muhtemel bağlantıların olasılığını tahmin etmeye çalışır. Daha yüksek olasılık bağlantıların oluşma olasılığını artırır. Üstel rasgele çizgeler gibi modeller bütün ağ için olasılık modellerini tahmin etmede kullanılır. Böylece, olasılıksal yaklaşımlar bireysel muhtemel bağlantıların olasılığını tahmin eden modeller ve bir ağın muhtemel yapılarının olasılığını tahmin eden modeller olmak üzere iki gruba ayrılır. Bunun yanında, olasılıksal yöntemler çoğunlukla önceki bölümde tanımlanan çizge teorili yaklaşımlara dayalıdır. Son yıllarda, çoğunlukla modellerin P^* sınıfı olarak ta adlandırılan sosyal ağlar için üstel rasgele çizge modellerinde büyük bir ilgi oluşmuştur. Üstel rasgele çizge modelleri bir ağın genel özellikleri, düğümler ve kenarlar kullanarak bütün bir ağ için olasılıksal modelleri tahmin etmede kullanılan popüler bir yaklaşımdır. Bu modeller belirli ağ yapılarının oluşup oluşmayacağı hakkında çıkarsamaya izin veren istatistiksel modellere dayalıdır. Bu yapısal özellikleri üretebilecek sosyal süreçler

hakkında hipotezler geliştirilebilir. Üstel rasgele çizge modelleri önceki çizge teorili yaklaşımların sınırlamalarının üstesinden gelerek ağlar için makul modeller inşa etme teşebbüsüdür.

En son bağlantı tahmini araştırması çoğunlukla ağların geçici ve yerel örüntülerine odaklanmıştır. Geçici ve yerel örüntüler bağlantının gelişimi üzerine önemli bir etkiye sahiptir. Son zamanlarda yapılan olasılıksal yaklaşımlarda, yerel olasılıksal yöntemler esnekliği ve verimliliği nedeniyle yaygın olarak tanıtılmış ve yükselen bir popülerlik kazanmıştır. Wang ve diğ. [47] birbiriyle yerel yakınlık içinde duran iki düğümün birlikte bulunma olasılığını tahmin etmek için olasılıksal çizge modelleri kullanan yerel bir olasılıksal model önermiştir. Daha spesifik bir şekilde bir düğümün yerel komşuluğunu modellemek için Markov rasgele alanlar kullanmışlardır. Yerel yakınlık yol uzunluğu üzerinde tanımlıdır.

Tylenda ve diğ. [48] mevcut sosyal ağın etkileşimleriyle ilgili geçmiş bilgileri dahil etmenin değerini araştırdı. Bu çalışma özellikle zaman farkındalığı dahil eden Wang ve diğ. [47] tarafından önerilen yerel olasılıksal modelin geliştirilmiş halidir. Onlar özellikle bağlantı tahmini işlerinde sosyal ağların gelişimini göz önüne aldılar. Bu çalışmada tanımlanan zaman-farkındalığı yöntemleri verilen bir düğümle bağlantılı olacak en yüksek k adayları sıralamak için kullanıldı. Olası adaylar düğümün yerel komşusundan seçilir. Bu denetimsiz sıralama yöntemi Adamic-Adar, Ortak komşular, Jaccard ve PageRank gibi diğer denetimsiz sıralama yöntemleriyle karşılaştırıldığında etkili sonuçlar üretir.

Kashima ve diğ. [49] Web, sosyal ağlar ve biyolojik ağlar gibi ağ yapılı alanlar için bağlantı tahmini problemine yeni bir yaklaşım önerir. Onların yaklaşımları düğüm özelliklerinden ziyade ağ yapısının topolojik özelliklerine dayalıdır. Çalışmalarında ağ gelişimi için parametrik terimle ifade edilmiş yeni bir olasılıksal model önerir ve daha sonra düğümler arasındaki bağlantıları tahmin etmede kullanılan böyle modeller için etkili bir artımsal öğrenme algoritması sunar. Bu yöntem bir dereceye kadar Markov kabulüne benzer olan $t + 1$ zamanındaki bir kenarın durumu t durumuna bağlıdır ifadesindeki gibi zamanla bir düğümden diğerine bir kenar oluşturma olasılığını hesaplar. Bazı durumlarda düğüm bilgileri mevcut iken ağların yapısal bilgisi tamamen eksiktir veya kısmen vardır. Böylece topolojik özelliklerin mevcut olmadığı yerlerde bağlantı tahmini ismi daha karmaşık bir hal alır. Leroy ve diğ. [50] yukarıdaki probleme bir çözüm olarak önyükleyici olasılıksal çizgeye dayalı iki aşamalı bir yöntem önermişlerdir. İlk aşamada bu yaklaşım bir önyükleyici olasılıksal çizge inşa eder. Burada kenarlar düğümlerin grup üyelikleri

kullanarak hesaplanan olasılıklara sahiptir. İkinci aşamada olasılıksal önyükleyici çizgenin özellikleri sonuç çıktısı olarak kabul edilecek kenarların yeni olasılıklarını çıkarmada kullanılır. Bu yöntem Flickr veri kümesi üzerinde test edilmiştir.

Popescu ve diğ. [51] bilimsel yayın alanında atıfları tahmin etmek için istatistiksel ilişki öğrenme yöntemini kullandılar. Bu alanda bağlantı tahmini modelleri atıf tavsiye edici servis olarak kullanılabilir. Bu servis muhtemelen süreçteki bir makalenin özetini, yazarların isimlerini ve kısmi referans listesini sağlayan kullanıcılara atıf tavsiyesinde bulunabilir. Bu yöntem iki ana süreçten ibarettir: ilişkisel veriden özellik adayların üretilmesi ve istatistiksel model seçme kriteriyle adayların seçilmesi.

Krzysztof ve diğ. [52] Matkov zincirlerin karışımına dayalı sosyal ağın temel alt çizgelerindeki yapısal değişikliklerin tahmin edici modelini önerdiler. Model bir gün uzunluklu zaman aralıklarda analiz edilmiş geniş sosyal ağdan elde edilen bir veritabanı üzerinde eğitilmiş ve doğrulanmıştır. Sonuç olarak yerel ağ topolojisi üzerinde bağıntıların gelişiminin ayırt edici örüntüleri çıkmıştır. Onların iddiaları bu gibi kısa zaman ölçeklerinde araştırılan ağın son derece dinamik olduğu ve bağlantı tahmininin klasik yöntemlerin ve yapısal analizin bu gibi probleme çözüm üretmediği olmuştur. Bu nedenle karmaşık ağ durumunda dinamik alt çizge madenciliği daha iyi bir tahmin etme uygunluğu sağlayabilir.

Maksimum olasılık yöntemleri ağ gelişimini modellemek için kullanılan popüler bir istatistiksel yöntemdir. Maksimum olasılık yöntemleri prensibine dayanarak bir metodolojiyi kullanmayı öneren Leskovec ve diğ. [53] ağ biçimlenme stratejilerinin geniş bir çeşidini araştırdı. Buna göre çizgedeki kenarın yeri ağların gelişiminde kritik bir rol oynar. Onların bulguları aslında yerel olmayan tercihli bağlılığa dayalı daha önceki ağ modellerinin eksikliklerini giderir. Leskovec ve diğ. [53] Delicious, Flickr, Answers ve LinkedIn'den oluşan dört farklı online sosyal ağı analiz ederek ağ gelişimin detaylı bir çalışmasını sunmuştur.

Yukarıda ifade edilen olasılıksal yöntemler düğüm benzerliklerinin bütün durumlarını kapsamamaktadır. Bu yüzden aşağıda verilecek olan bazı araştırmalar bağlantı tahminini etkili bir şekilde geliştirebilecek diğer benzerlik ölçütlerini keşfetmek için yapılmıştır.

3.2.3. Benzerlik Tabanlı Yaklaşımlar

Bağlantı tahmini probleminde ağdaki aktörler(düğüm) arasındaki benzerlik ölçütleri oldukça önemlidir. Genel yaklaşım herhangi iki düğüm arasında benzerlik ölçütünün değeri ne kadar yüksekse o iki düğüm arasında gelecekte bağlantı olma ihtimali o kadar yüksektir.

Benzerlik ölçütleri genel olarak semantik ve topolojik olarak kategorize edilebilir. Semantik ölçütlerde düğümün içeriği benzerlik ölçütü olarak ele alınır. Örneğin yazar işbirliği ağında makalelerin anahtar kelimelerdeki benzerlikle yazarlar arasında gelecekteki etkileşimler tahmin edilebilir [54]. Semantik ölçütlerden farklı olarak topolojik ölçütler benzerlik ölçütü hesaplamak için düğümlerin içeriklerinden ziyade ağın yapısını kullanırlar. İki düğüm arasındaki ortak komşuların sayısı topolojik ölçütlere bir örnektir. Topolojik ölçütler daha genel oldukları ve içerikle alakalı zengin bir özellik tanımına ihtiyaç duymadıkları için daha yaygın kullanılmaktadırlar [55]. Topolojik ölçütler literatürde genel olarak komşuluk tabanlı ve yol tabanlı diye kategorize edilmiştir [56]. Şimdi bu kategorilerdeki ölçüt birimlerine Liben-Nowell ve Kleinberg [57] çalışması referans alınarak değinilecektir. Bu çalışmada ağdaki herhangi x ve y düğüm çifti arasındaki ölçüt biriminin değeri $skor(x, y)$ olarak ifade edilmiştir.

3.2.3.1 Komşuluk Tabanlı Ölçütler

Komşuluk tabanlı yaklaşımlarındaki temel fikir x ve y düğümlerinin komşuları $\Gamma(x)$ ve $\Gamma(y)$ ne kadar büyük oranda kesişirse gelecekte aralarında bağlantı olma ihtimali de o kadar yüksektir. $\Gamma(x)$, x düğümünün ağdaki komşularının kümesini göstermektedir.

- **Ortak Komşular:** x ve y düğümleri için ortak komşuların sayısını göstermektedir. Newman [58] işbirliği ağlarında t zamanında x ve y 'nin ortak komşuları ile bunlar arasında işbirliği olma olasılığı arasındaki korelasyonu doğrulamak için kullanmıştır. Matematiksel ifadesi denklem (3.7)'de verilmiştir.

$$OK(x, y) = |\Gamma(x) \cap \Gamma(y)| \quad (3.7)$$

- **Jaccard Katsayısı:** x ya da y 'den rasgele seçilen bir f özelliğinin x ve y 'de birlikte bulunma olasılığıdır. Çalışmada [57] özellikler yerine komşuluk alınmıştır. Denklemi (3.8)'de verilmiştir [59].

$$JK(x, y) = |\Gamma(x) \cap \Gamma(y)| / |\Gamma(x) \cup \Gamma(y)| \quad (3.8)$$

Jaccard, ortak komşuların normalleştirilmiş halidir.

- **Adamic/Adar Katsayısı:** Bu ölçüm iki web sayfasındaki içeriklerin birbirlerine ne kadar yakın olduğunu ölçmektedir. Bunu yapmak için sayfaların özellikleri belirlenmiştir. Formülü denklem (3.9)'da verilmiştir [60].

$$AA(x, y) = \sum_{z: x \text{ ve } y \text{ nin özelliği}} \frac{1}{\log(\text{frekans}(z))} \quad (3.9)$$

Bu nicelikte ortak özelliklerden nadir olanların ağırlık oranları artırılmaktadır. Bağlantı tahmini problemi için bu formül denklem (3.10)'daki gibi güncellenmiştir.

$$AA(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log(z)} \quad (3.10)$$

Bağlantı tahmini için yapılan çalışmalarda Adamic/Adar önceki ölçütlere göre daha iyi sonuç vermiştir.

- **Tercihli Bağlılık:** Ağda oluşacak yeni bir bağlantının uç noktalarından birinin x olma ihtimali x düğümünün komşularının sayısı ile orantılıdır. Diğer bir deyişle komşusu çok olan düğümlerin yeni bağlantı oluşturma ihtimali daha yüksektir. Newman [58] ve Barabasi vd. [61] çalışmalarına göre bir işbirliği ağında x ve y arasında işbirliği oluşma ihtimali x ve y 'nin iş birlikteliklerinin sayısı ile orantılıdır. Matematiksel denklemi (3.11)'te verilmiştir.

$$TB(x, y) = |\Gamma(x)| \cdot |\Gamma(y)| \quad (3.11)$$

- **Kaynak Paylaştırma İndeksi:** Kompleks ağlarda kullanılır. Birbiriyle direk bağlantılı olmayan düğüm çiftleri arasındaki bağlantıları ölçer. Birbirleri arasında bağlantı olmadığı halde düğümler ortak komşuları arasında iletim sağlarlar. İletim sağlayan her düğümün kaynak ünitesi vardır ve bu kaynağı komşularına eşit dağıtır. Bu düğümler arasındaki benzerlik birbirlerinden aldıkları kaynaklara göre hesaplanır.

$$KP(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k(z)} \quad (3.12)$$

Liben-Nowell ve arkadaşları [62] ve Zhou ve arkadaşları [63] yerel benzerlik indislerini kullanarak gerçek ağlar üzerinde kıyaslamalar yaptılar. $k(z)$ z 'nin derecesidir.

3.2.3.2. Yol Tabanlı Ölçütler

Yol tabanlı ölçütler iki düğüm arasındaki en kısa yolların sayısını temel almaktadır.

- **Katz:** Leo Katz [64] tarafından önerilen bu ölçüt x ve y düğümleri arasındaki en kısa yolların sayısının toplamını hesaplamaktadır. Benzerlik hesaplamasında uzun yolların katkısının dezavantajından kurtulmak için yolun uzunluğu bir parametreye (β^l) üs olarak eklenmiştir. Böylece uzun yolların hesaplamaya katkısı daha az olmaktadır. Bu ölçüt denklem (3.13)'deki eşitlikle hesaplanır.

$$Katz(x, y) = \sum_{l=1}^{\infty} \beta^l \cdot |yollar_{x,y}^{(l)}| \quad (3.13)$$

$|yollar_{x,y}^{(l)}|$, x ve y düğümleri arasında l uzunluğundaki yolların sayısıdır. $\beta > 0$ 'dır. β parametre değeri çok küçük verildiğinde 3 ve üzeri uzunluğundaki yolların katkısı az olacağından sonuç ortak komşular değerine yakın bir değer olacaktır. Bu formül komşuluk matrisi (M) üzerinde denklem (3.14) ile hesaplanmaktadır.

$$(I - \beta M)^{-1} - I \quad (3.14)$$

I , birim matristir. Bu ölçütün bir dezavantajı hesaplama zamanının yüksek olmasıdır, bu yüzden büyük ağlar için pek uygun değildir. Liben-Nowell ve Kleinberg [57] tarafından ağırlık ve ağırlıksız ağlara uyarlanmıştır. Ağırlıksız ağlarda x ve y düğümleri arasında işbirliği varsa $yollar_{x,y}^{(1)} = 1$, yoksa 0 alınmış, ağırlıklı ağlarda $yollar_{x,y}^{(1)}$ değeri için x ve y düğümleri arasındaki iş birlikteliği sayısı alınmıştır.

- **Ulaşma zamanı:** x düğümünden başlayıp komşuları üzerinden adım adım hareket eden rasgele bir yürüyüş oluşturulur. Ulaşma zamanı $H_{(x,y)}$, x düğümünden başlayan bu rasgele yürüyüşün y düğümüne erişmesi için beklenen adım sayısıdır. Ulaşma zamanının düşük olması x ve y 'nin benzer olduğunu aralarında bağlantı olabileceğini göstermektedir. Bu ölçüt birimi simetrik değildir. O yüzden bunun yerine gidip gelme zamanı (commute time), $C_{(x,y)} = H_{(x,y)} + H_{(y,x)}$, kullanılabileceği belirtilmiştir. y düğümünün çok geniş bir dağılım olasılığı olduğunda $H_{(x,y)}$ çok küçük olacaktır. Bunu dengelemek için ölçüt normalleştirilmiştir. Ulaşma zamanının denklemi (3.15)'de verilmiştir.

$$UZ(x, y) = -(H_{(x,y)} \cdot \pi_y + H_{(y,x)} \cdot \pi_x) \quad (3.15)$$

- **Köklü PageRank:** Ulaşma zamanında x ve y düğümleri çok kısa yollarla birbirlerine bağlı olsalar da ağın geri kalanındaki yolların çok uzun olması bu ölçüt birimi için bir dezavantaj oluşturmaktadır. x ve y arasındaki rasgele yürüyüş her bir adımda α parametresindeki olasılık değeri ile yeniden başa döndürülebilir. Böylece ağın çok uzak kısımlarının gezilmesinden kurtulmuş olunur. Rasgele yürüyüşün belli bir olasılıkla yeniden başlatılması web sayfalarındaki PageRank [65] ölçütünün temelidir. Rasgele yürüyüş $\alpha[0,1]$ olasılığı ile başa döner, $1 - \alpha$ olasılığı ile o an bulunan düğümün komşularından rasgele birine gider. Bu işlem her adımda uygulanır. Liben-Nowell ve Kleinberg [57] tarafından önerilen bu ölçüt birimi matematiksel olarak Hasan ve Zali [56] çalışmasındaki gibi gösterilmiştir:

i düğümünün tüm koşulları için D diagonal derece matrisinde $D[i, i] = \sum_j A[i, j]$ 'dir. $N = D^{-1}A$, komşuluk matrisinin satırlarının 1'e normalleştirilmesidir. Formülü denklem (3.16)'da verilmiştir.

$$KPR = (1 - \beta)(I - \beta N)^{-1} \quad (3.16)$$

β , yukarıda bahsedilen α yerine kullanılmıştır.

- **SimRank:** Jeh ve Widom [66] tarafından önerilen bu ölçütün dayandığı esas şudur: iki nesne benzer nesnelerle ilişkili ise bu nesneler benzerdir. Öz yinelemeli olarak formül aşağıdaki gibi tanımlanmıştır:

$$Benzerlik(x, y) := \gamma \cdot \frac{\sum_{a \in \Gamma(x)} \sum_{b \in \Gamma(y)} Benzerlik(a, b)}{|\Gamma(x)| \cdot |\Gamma(y)|} \quad (3.17)$$

3.2.4. Diğer Yaklaşımlar

Benzerlik tabanlı ve olasılıksal yöntemler hariç, literatürde farklı kuramsal temellerden pek çok değişik yaklaşım daha vardır. Bunlardan bazıları sık örüntü madenciliği, rasgele yürüyüş ve yayma yöntemleridir. Bu bölümde kısaca bu yöntemlere değinilecektir.

Lin ve diğ. [67] seyreklik analizine dayalı denetimsiz bir bağlantı keşfi yöntemi önermiştir. Çalışmalarında ikili ilişkilerin kümesiyle bağlantılı düğümlerin ve ilginç yolların keşfine odaklanmışlardır. Veri kümesindeki her bir nesne ayrı bir varlık olarak ele alınmış ve bu varlıkları bağlayan ikili ilişkilerin farklı tiplerde olabileceği sonucu çıkarılmıştır. Bu çeşit bir veri seviyelenmiş çizgelerle temsil edilebilir. Burada düğümler varlıkların, bağlantılarda ikili ilişkilerin yerini alır. Bu çalışmanın anahtar kabulü verinin zengin bir ilişki sözlüğü kullanmasıdır. Farklı veri bağlantı tipleri farklı semantik ilişkileri temsil eder. Yani, benzer yapıları farklı çizgeler bağlantı tiplerine bağımlı olarak farklı anlamlara sahip olabilir. Bu kabullere dayanarak bağlantı keşfi probleminin 3 farklı sınıfı tanımlanabilir. Bunlar, yeni yol keşfi, yeni döngü keşfi ve önemli düğüm keşfidir. Yeni yol keşfi, çizgede varlıkların bir çifti verildikten sonra onlar arasında en ilginç veya yeni yollar

keşfetmektir. Yeni döngü keşfi, çizgede rasgele bir varlık verildikten sonra onunla başlayıp yine onunla biten en ilginç ve yeni döngüleri bulmaktır. Önemli düğüm keşfi ise, çizgede rasgele bir varlık verildikten sonra ona bağlı en önemli diğer varlıkları belirlemektir.

Kashima ve diğ. [68] “bağlantı yayılma” olarak adlandırılan yarı denetimli bir algoritma sundular. Bu seviye yayılma prensibine benzerdir. Yani eğer iki düğüm birbirlerine benzer ise muhtemelen aynı seviyedirler. Burada, iki düğüm çifti birbirine benzer ise muhtemelen aynı bağlantı tipine sahiptir çıkar. Çalışmalarında seviye yayılma yöntemini çoklu bağlantı tipli düğüm çiftlerine uyguladılar ve düğümler arasındaki ilişkileri tahmin ettiler.

Sosyal ağlarda bilgi yayılması bağlantı gelişimi için temel bir faktördür. Özellikle Twitter gibi bilgi değişimi veya paylaşımı için adanmış sosyal ağların gelişimi önemlidir. Cha ve diğ. [69] Twitter sosyal ağında sosyal ağ gelişimi için bilgi yayılmasının önemini vurgulayan deneysel çalışmalar yürütmüşlerdir. Çalışmalarında Flickr sosyal ağındaki bilgi dağılımının geniş ölçekli izlerini toplayıp analiz ettiler. Analizleri 11 milyon fotoğraf üzerinde 2,5 milyon kullanıcının favori işaretlerinin izlenmesine dayalıdır.

Bağlantı tahmini için rasgele yürüyüş tabanlı yaklaşımlar son yıllarda çok popülerdir. En son araştırmalarda yaygın bir şekilde kullanılmıştır. Backstrom ve diğ. [70] bağlantı tahmini için topolojik ve düğüm/kenar özelliklerini birleştiren denetimli rasgele yürüyüş tabanlı bir yöntem önermiştir. Bu yöntem birleştirilmiş bağlantı tahmini algoritması içinde ağın düğüm ve kenarlarının karakteristikleriyle (nitelikler, özellikler) ağın yapısını birleştirir. Yöntemde rasgele yürüyücü, diğerlerinden ziyade verilen düğümleri (pozitif eğitme örnekleri) ziyaret edecek şekilde bir PageRank benzeri rasgele yürüyüşü nasıl etkileyeceğini öğrenir. Bunu düğüm ve kenarın özelliklerini kullanarak başarır. Ağırlıklı ağlar üzerinde rasgele yürüyüş negatiflerden daha çok pozitif düğümleri ziyaret etme eğilimi gösterir. Bağlantı tahmini kapsamında pozitif düğümler gelecekte yeni kenarların oluşturulacağı düğümlerdir, negatifler ise bütün diğer düğümlerdir. Denetimli öğrenme o zaman: “eğer bir s kaynak düğümü ve düğümler hakkında eğitme örnekleri verilirse, s ’nin gelecekte linkleri oluşturma” işi olarak biçimlendirilir.

Lichtenwalter ve diğ. [67] ağlar üzerinde *PropFlow* olarak adlandırılan yeni bir denetimsiz bağlantı tahmini algoritması önerir. Buna göre v_i düğümünden başlayan sınırlı bir rasgele yürüyüş geçiş olasılıkları olarak bağlantı ağırlıklarını kullanarak l veya daha az adımda v_j ye ulaşır. Sınırlamalar, yürüyüşün hedef düğüm olan v_j ’ye eriştiğinde veya başlangıç düğüm olan v_i ’de içeren herhangi bir düğüm yeniden ziyaret edildiğinde

sonlanmasıdır. Yürüyüş bağlantıları ağırlıklarına göre seçer. Yeni bağlantıların olasılık tahmini olarak hizmet edecek bir s_{ij} skoru üretir. *PropFlow*, *Köklü PageRank*'e benzerdir. Fakat yayılmanın daha yerel bir ölçüsüne sahiptir ve kaynak düğümünden uzaktaki gürültülerden etkilenmez. Köklü PageRank'den farklı olarak PropFlow'un hesaplanması yürüyüşlerin yeniden başlamasına veya yakınsamaya gerek duymaz. Bunun yerine l yüksekliğiyle sınırlı değiştirilmiş bir ilk önce enine aramayı kullanır. Böylece hesaplanması çok daha hızlıdır. Ağırlıklı, ağırlıksız, yönlü veya yönsüz ağlarda kullanılabilir.

Bağlantı tahmini araştırmasında üzerinde tartışılan en önemli problemlerden biri de bir ağda gözlenen bağlantıları kurmak için hangi nitelik alt kümelerinin önemli olduğunu belirlemektir. Bu durum doğru bağlantı tahmini için çok önemlidir. Bu maksatla Scripps ve diğ. [71] bağlantı tahminiyle en iyi şekilde hizalanmış ilgili nitelikleri veya topolojik özellikleri tanımlamaya izin veren esnek bir çerçeve tanımlamıştır. Bu çalışma matris hizalama yaklaşımına dayalı bağlantı tahmini için yeni bir ayrık öğrenme tekniğini ifade eder. Algoritma otomatik olarak düğüm nitelikleri ve komşuluk topolojik özelliklerinden hesaplanan ağırlıklı benzerlik matrisleriyle bir ağın komşuluk matrisini hizalayarak bağlantı yapısının en ayırt edici özelliklerini belirler.

Arkadaşlık fikrinin sosyolojik çalışmalarda göz önüne alınandan daha geniş olan online sosyal ağlarda arkadaşlık bağlantıları daha yoğundur fakat bağlantılar daha gürültülü bilgi içerir (Yani daha zayıf ilişkiler). Buna karşın ağlar gerçek sosyal ağı oluşturan varlıklar arasında ek hareketsel olaylar içerir (Yorumlama, sohbet etme, etiketleme, bağlantı kurma vb.). Bu nedenle Kahanda ve diğ. [72] hareketsel bilgidен bağlantı gücünü tahmin etmek için denetimli bir öğrenme yaklaşımı önermiştir. Bu yaklaşımın yeniliği hareketsel özelliği kullanmaktır. Bu özelliğin önemi büyüktür çünkü oluşan online sosyal ağlar çoklu ilişkilerdir. Bundan dolayı ilişkinin her bir tipi ile oluşan hareketlerin etkililiğini çalışmak değerlidir. Hareketsel özellikler kullanıcılar arasında hareketsel bilgiyi göz önüne alır. Bu yaklaşım Purdue üniversitesinin Facebook ağından kullanıcıya açık verisi üzerinden değerlendirilmiş, güçlü ilişkileri doğru bir şekilde tahmin edebildiği gösterilmiştir.

En yaygın kullanılan ve en başarılı tavsiye etme algoritmalarından biri de işbirliğiyle filtrelemedir. Bu yöntem kullanıcının ilgi ve tercihlerini çıkarmak için kullanıcı-nesne etkileşimi içindeki korelasyonları keşfeder. İşbirliğiyle filtreleme yaklaşımlarının tavsiye etme kalitesi çoğunlukla veri seyreklik problemiyle sınırlıdır. Bu problemin üstesinden gelmek için Huang ve diğ. [73] işbirliğiyle filtreleme tavsiyelerini yapmak için ağ

modelleme literatüründe en son önerilen bağlantı tahmini yaklaşımları kullanarak ve çizgeler olarak kullanıcı-nesne etkileşimlerini temsil ederek çizge tabanlı bir algoritma önermiştir. Bu çalışmada tavsiyeler yapmak için bağlantı ölçülerinin geniş bir aralığını önermişlerdir. Kitap tavsiye veri tabanı üzerinde yapılan ilk deneysel sonuçlar bu ölçülerin bazılarının standart işbirliğiyle filtreleme algoritmalarından önemli bir şekilde daha iyi performans sergilediğini göstermiştir.

4. BAĞLANTI TAHMİNİNDE MAKİNE ÖĞRENMESİ YÖNTEMLERİ

Bağlantı tahmini problemi için önerilen yaklaşımlara denetimsiz ve denetimli yöntemler kategorilerinde değinilecektir. Denetimsiz metotlarda [57] bir benzerlik ölçütü seçilir, ağdaki düğüm çiftlerinin bu benzerlik ölçütüne göre skorları hesaplanıp en iyi skora sahip düğüm çiftleri arasında [67] gelecekte bağlantı olacağı tahmininde bulunulur. Denetimli metotlarda ise farklı benzerlik ölçütlerinin ve ağla ilgili başka özelliklerin birer tahmin edici özellik olarak kullanıldığı bir sınıflandırma işi yapılmaktadır [55]. Zaman tabanlı yaklaşımlarda [55] bağlantıların ne zaman oluştuğu ne zamandır devam ettiği gibi bilgilerden yararlanılarak bağlantı tahmini yapılmaktadır. Zaman içerisinde gerçekleşen olayların bağlantı tahmininde önemli olduğu düşünülmektedir. Şimdi bu kategorilerdeki bazı çalışmalara değinilecektir.

4.1. Denetimsiz Yöntemler

Liben-Nowell ve Kleinberg [57] tarafından yapılan çalışma bu kategorideki çalışmaların temelini oluşturmaktadır. Çalışmada benzerlik ölçütleri düğümler arasında gelecekte oluşacak bağlantıları tahmin etmek için kullanılmıştır. $t_0 < t'_0 < t_1 < t'_1$ olacak şekilde dört ayrı zaman belirlenmiştir. $[t_0, t'_0]$ eğitim aralığı, $[t_1, t'_1]$ test aralığı olarak adlandırılmıştır. $G[t_0, t'_0]$, t_0 ve t'_0 zaman aralıklarında G ağında görülen düğüm ve bağlantılardır. Önerilen algoritmanın çıktısı $G[t_0, t'_0]$ ağında olmayan ancak $G[t_1, t'_1]$ ağında olan bağlantılardır.

Yöntemde, bağlantı tahmininde kullanılan topolojik ölçütler bölümünde bahsedilen benzerlik ölçütleri kullanılmıştır. Bunlarında dışında bir de düğüm çiftleri arasındaki uzaklığı benzerlik ölçütü olarak kullanan çizge uzaklığı ölçütü kullanılmıştır. Aralarında bağlantı olmayan düğüm çiftlerinin benzerlik ölçütü skoru hesaplanır, düğüm çiftlerinin skorları azalan şekilde sıralanır ve en yüksek skora sahip düğüm çiftleri arasında gelecekte bağlantı olacağı öngörüsünde bulunulur.

Çalışmada performans veya kaliteyi artırmak için bazı yaklaşımlarda bulunulmuştur. Bu yaklaşımlar aşağıda verilmiştir.

Low-rank yaklaşımı: Ağın gösterimi için komşuluk matris kullanılabilir. Komşuluk matrisi (M) kullanıldığında benzerlik ölçütlerinin formülleri M matrisinin bir formu olarak yazılır. Mesela x ve y düğümlerinin ortak komşular ölçütü değeri M matrisi üzerinde x . ve y . satırlar ile kolayca bulunur. Ancak Katz gibi ölçütler büyük ağlarda M matrisi üzerinden kolayca elde edilemez, oldukça zaman alıcı bir işlem gerektirir. Bunun için M matrisini temsil edebilecek küçük bir M_k matrisi belirlenir. M yerine M_k 'nın kullanılması gürültü giderme olarak görülebilir. Testlerde Katz ve ortak komşular ölçüt birimleri için M matrisi yerine M_k matrisi de denenmiştir.

Unseen-biagrams: $skor(x, y)$ 'yi hesaplarken x 'e benzer z düğümlerinin $skor(z, y)$ değerleri kullanılabilir. Bu şekilde daha isabetli bir skor değeri elde edileceği düşünülmektedir. Ortak komşuluk ölçütü aşağıdaki gibi güncellenmiştir [57]:

$$skor_{ağırlıksız}^*(x, y) := \left| \left\{ z: z \in \Gamma(y) \cap S_x^{(\delta)} \right\} \right| \quad (4.1)$$

$$skor_{ağırlıklı}^*(x, y) := \sum_{z \in \Gamma(y) \cap S_x^{(\delta)}} score(x, z) \quad (4.2)$$

$S_x^{(\delta)}$, ifadesi x 'e en yakın δ tane düğümün $skor(x, .)$ değerini göstermektedir.

Kümeleme: Bağlantı tahmini için düşük değerlere sahip bağlantılar ağdan silinebilir. Ağdaki tüm bağlantılar için $skor(u, v)$ hesaplanır daha sonra en düşük değere sahip olan bağlantıların $(1 - p)$ oranı kadarı ağdan silinir. $p \in [0,1]$ 'dir. Daha sonra oluşan yeni ağda tüm düğüm çiftlerin $skor(u, v)$ değerleri yeniden hesaplanır. Bu hesaplama ile benzerlik ölçüt değerlerinin daha güvenilir olacağı düşünülmektedir.

4.2.Denetimli Yöntemler

Lichtenwalter ve diğ. [67] yapmış olduğu çalışmada bağlantı tahmini problemi için denetimli genel bir sistem önermişlerdir. Bu sistem ile ağırlıklı, ağırlıksız, yönlü, yönsüz gibi farklı türdeki ağlar ele alınabilmektedir. Aynı zamanda bağlantı tahmini probleminde önem arz eden konulardan metotların genelliği, varyans azaltma, topolojik problemler ve derece dengesizliği, örnekleme yaklaşımı konuları da ele alınmıştır. Ayrıca bağlantı

tahmini problemi için denetimsiz bir algoritma da önerilmiştir. Yazarlara göre denetimsiz metotlar ağdaki dinamikliği, bağımlılıkları ve diğer özellikleri ele almakta yetersizdir.

Test için 712 milyon telefon görüşmesinden oluşan bir telefon ağı ile 1995-2000 yılları arasında fizikte yoğun maddeler konusunda yapılan çalışmaları gösteren 19.464 çoklu etmen olaydan oluşan bir akademik iş birliği ağı kullanılmıştır. Bu ağa *condmat* adı verilmiştir. Telefon ağı yönlü ve ağırlıklı bir ağıdır. Yönler aramayı kimin yaptığı ve kimin arandığı bilgisini, ağırlıklar ise kaç defa arama yapıldığını göstermektedir. Telefon ağında ele alınan zaman aralığındaki ilk beş haftalık görüşmeler özellik çıkarımı için altıncı haftadaki görüşmeler ise test için kullanılmıştır. *condmat* ağı ağırlıklı ve yönsüz bir ağıdır. Yazarlar düğümler, yazarlar arasındaki iş birlikleri bağlantılar olarak gösterilmiştir. Bağlantılardaki ağırlık yazarlar arasında kaç defa iş birliği olduğunu göstermektedir. 1995-1999 yılları arasındaki veriler özellik çıkarımı için 2000 yılındaki veriler ise test için kullanılmıştır.

Denetimsiz yaklaşımlar kategorisinde PropFlow metodu önerilmiştir. Bu metot, v_i düğümünden başlayıp v_j düğümünde biten l uzunluğundaki rasgele yürüyüşün olasılığı ile bir benzerlik ölçütü sunmaktadır. Yürüyüşün bir düğümünden diğer düğüme geçme olasılığı bağlantıların ağırlıkları ile orantılıdır. Ağırlığı fazla olan bağlantı üzerinden gitme olasılığı daha fazladır. Bu olasılığın sonucu i, j düğümleri arasında yeni bağlantı oluşma olasılığı olarak ele alınmaktadır. Köklü Pagerank ölçütüne benzemektedir. Farklı olarak ölçütün hesaplanması yerel bir bölgede olmakta kaynak düğümünden uzak topolojik gürültülerden etkilenmemektedir. Diğer bir farklılık da Köklü Pagerank'taki yeniden başa dönme durumunun olmamasıdır. l uzunluğunda enine arama yapılarak bu ölçüt hesaplanmaktadır. Böylece hesaplama hızlı bir şekilde yapılmaktadır. Önerilen yöntem yönlü ya da yönsüz, ağırlıklı ya da ağırlıksız ağlarda uygulanabilmektedir. Bu yöntem telefon ağında klasik denetimsiz yöntemlere (tercihli bağlılık, ortak komşular, Adamic/Adar, Jaccard katsayısı ve Katz) göre %15, Köklü Pagerank'a göre %8.75 daha iyi sonuç vermiştir. *condmat* ağında etkili olamamıştır. Telefon ağındaki başarının ağın yapısı ile ilgili olduğu düşünülmüştür. Yöntemin transfer ağları gibi bilginin ya da kaynağın ağ üzerinden sıklıkla hareket ettiği ağlarda daha iyi sonuç verdiği belirtilmiştir.

İncelenecek ağın hangi zaman aralıklarını kapsayacağı τ_x ve τ_y parametreleri ile belirlenmektedir. $G_x = (V_x, E_x)$ ile gösterilen ağ t_0 ile $t_{0+\tau_x}$ zaman aralığındaki ağı, $G_y = (V_y, E_y)$ ile gösterilen ağ t_{τ_x+1} ile $t_{\tau_x+\tau_y}$ zaman aralığındaki ağı temsil etmektedir.

G_x ağı topolojik ölçütlerin hesaplanacağı, düğüm niteliklerinin özellik olarak kullanılabileceği ağıdır. G_y ise oluşacak yeni bağlantıların araştırılacağı ağıdır. Bağlantı tahmini $|V_x|^2 - |E_x|$ örnekleri içinden yapılacaktır. τ_x ve τ_y parametreleri modelin başarısını etkilemektedir. τ_x parametre değerinin artması doyuma ulaşmış ağdaki topolojik ölçütlerin kalitesini artıracaktır. τ_x 'nin yeterince büyük olması ile gözlemlenen olayların oluşturacağı topoloji ağ statik bir duruma yakınlaşacak böylece daha tutarlı sonuçlar elde edilebilecektir. Denetimsiz metotların tahmin gücünü artıracak, denetimli metotlarda ilişkilerin daha iyi anlaşılmasını sağlayacak özelliklerin belirlenmesini kolaylaştıracaktır.

Bağlantı tahmini için önerilen metotların genelliği ile ilgili [57] 'deki denetimsiz metotların başarılarının verilere bağımlılık gösterdiği ancak denetimli metotlar kapsamında sınıflandırıcıların birçok ortamda genellik gösterebileceğinden bahsedilmiştir.

Denetimli metotların diğer bir artısı da sınıflandırma algoritmalarının örnekleme yaparak varyans azaltmadan faydalanabilmesidir. Bunu denemek için ortak komşular ölçütü sınıflandırmada özellik olarak kullanılmış ve ağdaki bağlantılarda belli oranlarda örnekleme yapılmıştır. En iyi sonucun hiç örnekleme yapılmadığı durumda elde edildiği görülmüş, örnekleme denemesi başarısız olmuştur. Ancak birden fazla denetimsiz metodun özellik olarak kullanılacağı sınıflandırma yaklaşımında varyans azaltma tekniklerinin verimliliği artıracağı iddia edilmiştir.

Bağlantı tahmininde verideki dengesizlik bir problemdir. Gerçek ağlar seyrek ağlardır. Seyrek ağlarda bağlantı sayısı ($|E|$) olası bağlantı sayısından ($|V|^2$) oldukça düşüktür. Bu durum dengesizlik problemini doğurmaktadır. Sınıflandırma yaklaşımlarında da bu durum bir problemdir. Örneğin iki sınıfın olduğu bir veri kümesinde, bir sınıfa ait verilerin sayısı diğer sınıfa ait verilerin sayısından oldukça fazla olduğunda bir dengesizlik problemi ortaya çıkar. Denetimsiz metotlar bu dengesizliğin üstesinden gelemeyen denetimli metotlar veriyi dengeleyebilirler.

Lichtenwalter ve diğ. [67] yapmış olduğu çalışmada veri madenciliği için makine öğrenmesi ile ilgili birçok algoritmanın bulunduğu bir araç olan WEKA [74] kullanılmıştır. Sınıflandırıcılardan C4.5 [75] , J8, Naive Bayes kullanılmıştır. Varyans azaltma ve örnekleme için bagging [76] ve rasgele orman [77] metotları kullanılmıştır. Sınıflandırmada özellik kümesi için temel topolojik özellikler ve denetimsiz metotlarda kullanılan klasik ölçütler kullanılmıştır. Bunların listesi Tablo 4.1'de verilmiştir. Gelen-Bağlantı(i) i düğümüne gelen bağlantı sayısıdır, Giden-Bağlantı(i) i düğümünden çıkan bağlantı sayısıdır. Gelen-Yoğunluk(i) i düğümüne gelen bağlantıların ağırlıkları toplamıdır,

Giden-Yoğunluk(i) i düğümünden çıkan bağlantıların ağırlıkları toplamıdır. Maksimum Akış (i,j) i düğümünden j düğüme 5 adımda dolaşılabilen noktalardır. Yüksek performanslı bağlantı tahmininde (HPLP) denetimsiz metotlar kullanılmamış temel topolojik özellikler kullanılmıştır. HPLP+'da ise tüm özellikler kullanılmıştır. Sınıflandırma ve aynı zamanda varyans azaltma için bagging metodu [76] ve performansı artırmak için de rasgele orman metodu [77] kullanılmış, yaklaşık %8.11 gibi bir gelişme görülmüştür. Test sonuçlarında önerilen yöntem ele alınan denetimsiz metotlardan %30 daha iyi sonuç vermiştir.

Davis ve diğ. [20] tarafından farklı türlerde bağlantıların ya da düğümlerin olduğu heterojen ağlarda bağlantı tahmini problemi ele alınmıştır. Çalışmada heterojen ağlar için Adamic/Adar ölçütü ağırlıklı ağlarda kullanılabilecek şekilde genişletilmiştir.

Tablo 4.1. Özelliklerin Listesi [67].

İsim	Parametreler	HPLP	HPLP+
Gelen-Bağlantı(i)	-	✓	✓
Gelen-Yoğunluk(i)	-	✓	✓
Gelen-Bağlantı(j)	-	✓	✓
Gelen-Yoğunluk(j)	-	✓	✓
Giden-Bağlantı(i)	-	✓	✓
Giden-Yoğunluk(i)	-	✓	✓
Giden-Bağlantı(j)	-	✓	✓
Giden-Yoğunluk(j)	-	✓	✓
Ortak Komşular(i,j)	-	✓	✓
Maksimum Akış(i,j)	$l = 5$		✓
En Kısa Yollar (i,j)	$l = 5$		✓
PropFlow(i,j)	$l = 5$		✓
Adamic/Adar(i,j)	-		✓
Jaccard Katsayısı(i,j)	-		✓
Katz(i,j)	$l = 5, \beta = 0.005$		✓
Tercikli Bağlılık(i,j)	-		✓

Gerçek hayatta birçok sistemin farklı türde nesne ve ilişkilerle karmaşık bir ağ oluşturduğuna değinilmiştir. Bu yapıdaki ağlara heterojen ağlar denilmektedir. Bağlantı tahmininde bu ağlarda ilişkiler genellikle ayrı ayrı ele alınmış ve aralarındaki ilişki ya da bağımlılıklar dikkate alınmamıştır. Bu durum için bir örnek verilmiştir: Arkadaşlık, aile, iş arkadaşı gibi ilişkiler sosyal ağlarda ayırt edici bir şekilde genelde modellenmemektedir. Gerçekte bir kişinin iş arkadaşının iş arkadaşı ile bir etkileşimde bulunma ihtimali o kişinin annesinin iş arkadaşı ile etkileşimde bulunma ihtimalinden daha yüksektir. Biyoloji

alanında da aynı durumun söz konusudur. Genellikle hücreler tek bir açıdan ele alınmaktadır. Örneğin sadece protein etkileşimi açısından ele alınmaktadır. Halbuki hücreler arasında birçok etkileşim vardır. Önerilen yöntem youtube, hastalık-gen ve iklim ağlarında test edilmiştir. Youtube ağı kullanıcıların video paylaşım ağıdır. Bu ağda düğümler kullanıcılardır. Bağlantıların adları ve anlamları aşağıda verilmiştir:

CN kullanıcıların bağlantı ağı, FR kullanıcının bağlantı ağının dışında yabancı ağlardaki kullanıcılarla paylaştığı bağlantılar, SBN kullanıcıların paylaştığı abonelikler, SBR paylaşılan aboneler, VID paylaşılan favori videolardır. Hastalık-gen ağında hastalık ve gen olmak üzere iki farklı türde düğüm vardır. G hastalık ile gen düğümleri arasında hastalık gen birlikteliği, P hastalık düğüm çiftleri arasında hastalıkların birlikte görülme durumu, F hastalık düğüm çiftleri arasında üst ve alt kategori bilgisidir. İklim ağında düğümler fiziksel yerleri bağlantılar ise iklimsel özellikleri göstermektedir. SST sıcaklık, SLP deniz seviyesindeki basınç, HWS ile VWS yatay ve dikey rüzgar hızı, PW yağış, RH bağıl nem, GH yüksekliktir. Değerlendirme için 10-kat çapraz doğrulama kullanılmıştır. Bu yaklaşımda her katta orijinal ağın rasgele %10'unu bulunmakta ve her kat orijinal ağ ile aynı dağılımı göstermektedir.

Hastalık-gen ağındaki G bağlantısı ikili ağ (bipartite network) yapısındadır. Çalışmada tercihli bağlılık ve yol tabanlı ölçütler ikili ağda rahatlıkla uygulanmış ancak komşuluk tabanlı ölçütler ikili ağa göre güncellenmiştir. Bunun için aralarında benzerlik hesaplanacak kaynak (s) ve hedef (t) düğümü arasında 3 uzunluğunda bir yapı (s, n_1, n_2, t) önerilmiştir. Ortak komşuluk ölçütü bu yapıda s ve t düğümleri arasındaki 3 uzunluğunda birbirinden farklı yolların sayısıdır. Jaccard Katsayısı s ve t düğümleri arasında 3 uzunluğunda birbirinden farklı yolların sayısının s ya da t 'den başlayan 3 uzunluğundaki yolların sayısına bölünmesi ile uyarlanmıştır. Adamik/Adar ölçütündeki logaritmik kısım $\log(|N_{n_1}| + |N_{n_2}|)$ ile uyarlanmıştır.

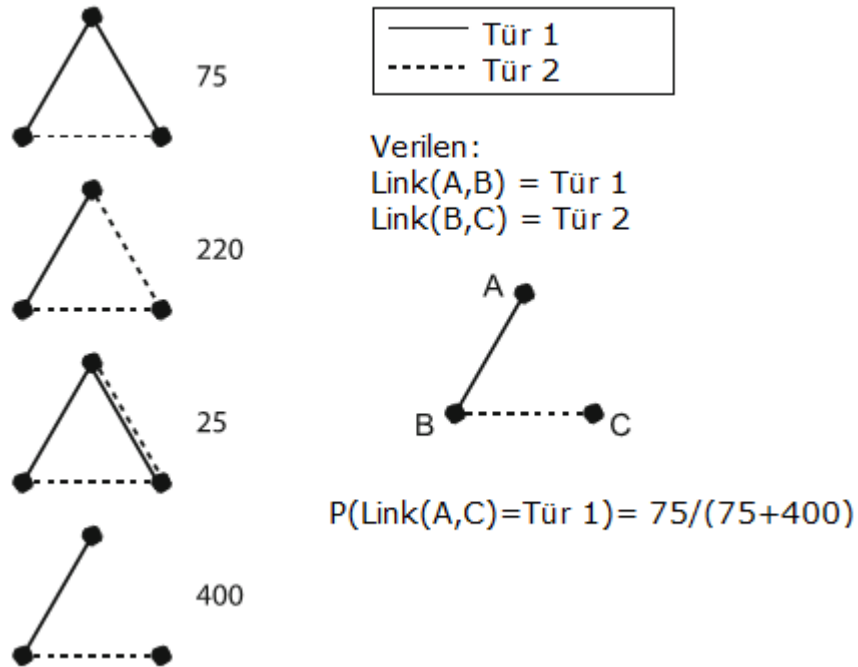
Davis ve diğ. [20] tarafından birden fazla bağlantı türünü bir arada ele almak için çok ilişkili bağlantı tahmini yöntemi önerilmiştir (MRLP). Bu yöntem bağlantıların ağırlıklı olduğu yapı için de uygundur. Yöntemde üçlü düğüm yapıların sayısı hesaplanmıştır. (s, n, t) düğümleri ve x türündeki bağlantı için hesaplama aşağıdaki gibi yapılmaktadır:

(s, n, t) üçlüsü ve x bağlantısı ile aynı örüntüde olan yapıların sayısı ve (s, n, t) üçlüsü ile aynı örüntüde olan üçlülerin sayısı hesaplanır ve buna ilk kısımda bulunan sayı ilave edilir. Daha sonra ilk kısım ikinci kısma bölünür. Matematiksel olarak denklem (4.3)'te gösterilmiştir.

$$P(x \subset \text{link_türü}(s, t) | \text{örüntü}(s, n, t))$$

(4.3)

Bu hesaplama Şekil 4.1'deki bir örnekle de açıklanmıştır. Şekilde sol tarafta aynı düğümler için farklı bağlantı yapılarından oluşan üçlüler ve bunların ağıdaki sayıları verilmiştir. Sağ tarafta ise verilen üçlü yapıda A ve C düğümleri arasında 1. türden bağlantının olasılığı hesaplanmıştır. Hesaplamaya sol taraftaki üçlülerin sadece en üstündeki ve en altındaki üçlü dahil edilmiştir.



Şekil 4.1. MRLP hesaplamasına bir örnek [20].

Bunun nedeni üçlü yapıya uygun üçlülerin bunlardan ibaret olduğundandır. Diğer üçlülerde A ve C düğümleri arasında 2. türden bağlantılar vardır. İstenen ise sadece 1. türdür.

MRLP'nin ortak komşuluk ölçütünün genişletilmiş hali olduğu ifade edilmiştir. s ve t düğümlerinin olduğu üçlüler aslında s ve t 'nin ortak komşularıdır. Her bir bağlantı türü için ağırlıklı ağıda formül denklem (4.4)'de gösterilmiştir.

$$skor(s, t) = \sum_{n \in N_s \cap N_t} w_n \quad (4.4)$$

(4.4) denklemindeki w_n çoklu bağlantılar için aşağıdaki gibi uyarlanmıştır:

$$w_n = \frac{\sigma |P(x) - P(x \subset link_türü(s, t) | örüntü(s, n, t))|}{P(link_türü(s, n) P(link_türü(t, n))} \quad (4.5)$$

$$\sigma = \begin{cases} 1 & P(x \subset link_türü(s, t) | örüntü(s, n, t)) > P(x) \\ 0 & P(x \subset link_türü(s, t) | örüntü(s, n, t)) = P(x) \\ -1 & P(x \subset link_türü(s, t) | örüntü(s, n, t)) < P(x) \end{cases} \quad (4.6)$$

Bu formül Adamic/Adar ölçütündeki ilke esas alınarak aşağıdaki gibi genişletilmiştir:

$$skor(s, t) = \sum_{n \in N_s \cap N_t} w_n \frac{1}{\log \begin{cases} |N_n(t1)| & t1 = t2 \\ |N_n(t1) + N_n(t2)| & t1 \neq t2 \end{cases}} \quad (4.7)$$

Denetimsiz metotlar güncellenmiş halleriyle test edilmiştir. Test işlemi iki ayrı şekilde yapılmıştır. İlkinde (homojen) her bir bağlantı türü için ayrı ayrı hesaplamalar yapılmış, ikincisinde (birleşik) ise bağlantı türlerinin hepsi eşit kabul edilerek hesaplama yapılmıştır. Denetimsiz metotların hiçbirisi tüm ağlarda tutarlı bir şekilde iyi sonuç verememiştir. Yerel komşuluk tabanlı metotlar youtube ağında baskın olmuşlardır. İklim ağında Jaccard katsayısı daha iyi sonuç vermiştir. Bunun iklim ağında coğrafik olarak birbirine yakın bölgelerin iklim olarak benzerlik göstermesinden kaynaklandığı belirtilmiştir. Hastalık ağında ise baskın bir metot olamamıştır. Önerilen MRLP yöntemi çoğunlukla Adamic/Adar ölçütünden daha iyi sonuç vermiştir. Homojen ağlar ise birleşik ağlardan daha iyi sonuç vermiştir.

Çıkarılan genel sonuç denetimsiz metotların ağa bağımlı olduğudur. Örneğin ortak komşuluk ölçütü paylaşımların artması ile yeni bağlantıların oluşma şansının arttığı ağlarda başarılıdır. Biyolojik yollarda ya da dil yapılarında diğer ağlardaki gibi başarı göstermeleri

güçtür. Heterojen yapılarda da bu durum geçerlidir. Hastalık-gen ağında baskın bir metot olamamıştır. Bu dezavantajlara karşı denetimli öğrenmelerle gerek homojen gerek heterojen ağlarda birden fazla unsurun bağlantı oluşumuna etkisinin daha iyi anlaşılabilmesi belirtilmiştir. İyi tasarlanmış bir sınıflandırma sistemi ile verilere bağımlılıktan kurtulabilir. Denetimli öğrenmedeki en önemli kısıt ağın topolojisini iyi bir şekilde yansıtmak özelliklerin seçimidir. Denetimli öğrenme yaklaşımının diğer bir avantajı verilerdeki dengesizlikte ortaya çıkmaktadır. Denetimsiz metotlar bu durumu ele alamamaktadırlar. Denetimli öğrenmede ise bu sıkıntının etkisi azaltılabilir.

Davis ve diğ. [20] yaptığı çalışmada heterojen ağlar için MRLP ölçütü denetimli öğrenme metodunda uygulanmıştır. Komşuluk tabanlı ölçütlerin değerinin artması bağlantı tahmini için doğrusal bir katkı oluşturacağı varsayımında bulunulmuştur. Yani ölçütün değeri ne kadar yüksekse ele alınan iki düğüm arasında bağlantı olma ihtimali o kadar yüksektir. Birden fazla bağlantı türünün olduğu özellikle bazı bağlantı türlerinin tahminine katkısının diğerlerinden fazla olduğu ağlarda bu varsayım genellikle başarısız olacaktır. İklim ağındaki dikey rüzgar hızı bağlantısı (VWS) buna örnek olarak verilmiştir. Bu örnekte komşuluk sayısı artması bağlantı tahmininde doğrusal bir artış göstermemiş farklı bağlantı türleri için farklı sonuçlar vermiştir.

Bu durumun üstesinden gelmek için MRLP ölçütü sınıflandırma için geliştirilmiştir (CW-MRLP). MRLP’de sayıları hesaplanan üçlü yapılardan özellik vektörü oluşturulmuştur. Her bir bağlantı türü için işlemler ayrı ayrı yapılmıştır. Üçlü düğümler arasında bağlantı türlerinin sayısına bağlı olarak çok fazla kombinasyon olabilir. Tüm kombinasyonlar ele alınmamıştır, var olan kombinasyonlar özellik vektörü için kullanılmıştır. Denetimli öğrenmede eğitim veri kümesi için ağdaki tüm düğüm çiftlerinin özellik vektör değerleri oluşturulmuştur.

Denetimli öğrenme hem homojen ağlar hem de heterojen ağlar için yapılmıştır. Homojen ağlarda özellik vektörü olarak temel topolojik özellikler ile [57] çalışmasındaki ölçütlerin kombinasyonu olan HPLP [67] kullanılmıştır. Bu özellikler şunlardır: düğüm derecesi, PageRank, ortak komşular, Jaccard katsayısı, Adamic/Adar, Köklü PageRank, PropFlow. HPLP ağdaki bir ilişki durumunu göstermek için bir ya da daha fazla ilişkiyi birleştirerek ele almaktadır. Bu da heterojen ağların daraltılmış bir şeklidir. HPLP’yi heterojen ağlarda uygulamak için HPLP’nin bir tek ilişki için kullandığı özellik vektörleri birleştirilmiştir. HPLP’nin genişletilmiş versiyonuna çok ilişkili HPLP (MR-HPLP) denmiştir.

Sınıflandırıcı olarak bagging metodu kullanılmıştır [76]. Bu metotla varyansın azaltılması da hedeflenmiştir. Bağlantı tahmininde veri dengesizliğini önlemek için eğitim veri kümesinin %25'i alınarak örnekleme yapılmıştır. Bagging sınıflandırıcının performansını artırmak için rasgele orman kullanılmıştır [77].

Bağlantı tahmini sonuçları denetimli öğrenmede CW-MRLP, HPLP ve MR-HPLP yöntemleri ve denetimsiz öğrenme metotlarının en iyi sonuç verenleri ile karşılaştırılmıştır. Denetimli metotlar denetimsiz metotlara göre iyi sonuç vermiştir. Çoklu bağlantı açısından sınıflandırıcı ile ele alınan yaklaşım (CW-MRLP) denetimsiz yaklaşımdaki çoklu bağlantı metodundan (MRLP) çoğunlukla daha iyi sonuç vermiştir. Heterojen ağlar için önerilen sınıflandırma yaklaşımları (CW-MRLP, MR-HPLP) homojen ağda kullanılan sınıflandırma yaklaşımına (HPLP) göre çoğunlukla ya aynı ya da daha iyi sonuç vermiştir. Denetimsiz yöntemlerdeki veriye bağımlılık problemi denetimli yaklaşımda tam olarak aşılanamamıştır. CW-MRLP iklim ağında baskın iken HPLP youtube ve hastalık-gen ağında daha iyi sonuç vermiştir.

Al Hasan ve diğ. [78] çalışmasında bağlantı tahminini bir denetimli öğrenme olarak ele alınmıştır. Etkili özellikler kullanılarak ikili sınıflandırma uygulanmıştır. Denetimli öğrenme için belirlenen özelliklerin hesaplaması kolay olmuş, aynı zamanda iyi sonuçlar da yakalanmıştır. Yazar iş birliği ağlarında bilinen sınıflandırma algoritmaları (karar ağaçları, k-NN, çok katmanlı algılayıcı, SVM, RBF ağı) bağlantı tahmini için kullanılmış, SVM metodunun en iyi sonuçları verdiği gözlemlenmiştir. Bağlantı tahmini problemi ile ilgili en temel çalışmalardan biri olan [57]'deki topolojik özellikler ile bağlantı tahminin doğruluğunu artıracak topolojik olmayan özellikler kullanılmıştır. Topolojik olmayan özelliklerin de sonuçların doğruluğunu artırabileceği vurgulanmıştır. Önerilen yaklaşım biyoloji ve bilgisayar bilimleri alanında yazar iş birliği ağlarında test edilmiştir. Biyoloji ağı için 1998-2002 yılları arasındaki veriler alınmış, ilk 4 yıl eğitim için son yıl test için kullanılmıştır. Bilgisayar ağı için 1990-2004 yılları arasındaki yayınlardan oluşan bir yazar iş birliği ağı oluşturulmuş, ilk 11 yıl eğitim son 4 yıl test için kullanılmıştır. Biyoloji ağı 831.478 yayından ve 156.561 yazardan, bilgisayar ağı ise 540.459 yayından 1.564.617 yazardan oluşmaktadır.

Özellikler üç kategoride belirlenmiştir: benzerlik, toplamsal ve topolojik.

1. **Benzerlik özellikleri:** Ağdaki aktörlerin ilgi alanları ne kadar yakınsa aralarında bağlantı olma ihtimali o kadar yüksektir. Biyoloji ağında anahtar kelimeler mevcut olduğundan bu özellik sadece bu ağda elde edilebilmiştir.

Anahtar kelimelerin eşleşme sayısı: Yazarların makalelerindeki tüm anahtar kelimeler belirlenir. Daha sonra yazarların anahtar kelimelerinin kesişimi hesaplanır. En çok kesişime sahip olan yazarlar gelecekte aralarında iş birliği olabilecek en güçlü adaylardır.

2. **Toplamsal özellikler:** Bu kategoride ele alınan özellikler tek bir düğümle ilgili özelliklerdir. Toplamsal bir fonksiyonla bu özellikler düğüm çiftleri için elde edilmiştir.

Yayınların toplamı: Öğrenme yılında yazarların işbirliğinde bulundukları yayın sayısıdır. Bu değerin yüksek olması yazarın üretken olduğunu göstermektedir. Ele alınan yazar çiftlerinden biri ya da ikisi üretkense bunlar arasında iş birliği olma ihtimali rasgele ele alınan iki yazar arasında iş birliği olma ihtimalinden daha yüksektir.

Komşuların toplamı: Yazarların iş birliği içinde olduğu toplam yazar sayısıdır. Bir yazar için bu özellik ne kadar yüksekse bu yazarın yeni işbirlikleri oluşturma ihtimali o kadar yüksektir.

Anahtar kelimelerin toplamı: Yazarın yapmış olduğu tüm çalışmalarındaki farklı anahtar kelimelerin toplamıdır. Disiplinler arası çalışmaları çok olan yazarların bu özelliği yüksektir. Bu yazarların yeni iş birliktelikleri kurma ihtimali de fazladır.

Kod sınıflarının toplamı: Yayınlar araştırma alanlarına göre kategorize edilmektedirler. Bu kategoriler de belirlenmiş kodlarla ifade edilmektedirler. Bir yayındaki kategori kodlarının fazla olması bunun disiplinler arası olduğunu göstermektedir. Böyle çalışmaları çok olan yazarların yeni iş birliktelikleri kurma ihtimalleri yüksektir.

Logaritmik toplam: Bilimsel araştırma alanındaki iş birlikteliklerinde birinci derece ilişkilerin yanı sıra ikinci derece ilişkiler de önemlidir. İkinci derece ilişkiler yazar işbirliği ağında yazarın komşusunun komşusu yani ikincil komşu anlamına gelmektedir. Bu özellikte ikincil komşuların sayısı logaritmik olarak toplanmaktadır. Bu değerin yüksek olması yeni iş birlikteliklerin ihtimalini artırmaktadır.

4.3. Öğrenme Modellerini Değerlendirme Ölçütleri

Öğrenme model başarımını değerlendirme ölçütleri Hata oranı, Anma, Kesinlik ve F-ölçütü'dür.

4.3.1. Sınıflandırma Hatası

Sınıflandırma yöntemlerinin hatalarını ölçmek için başarı, hata, doğruluk ve hata oranı gibi ölçüler kullanılır. Başarı, örneğin doğru sınıfa atanmasıyla elde edilir. Hata durumunda ise örnek yanlış sınıfa atanmıştır. Bu durumda;

Doğruluk: Bir sınıflandırıcı için *doğruluk*; doğru sınıflanmış örneklerin toplam örnek sayısına oranından bulunur.

Hata Oranı: Hata sayısının toplam örnek sayısına bölünmesidir. Hata oranı test kümesi kullanılarak hesaplanır ve aynı zamanda $hata\ oranı = 1 - doğruluk$ 'tur.

Model başarımını değerlendirme ölçütleri modelin ne kadar doğru sınıflandırma yaptığını ölçer. Hız, ölçeklenebilirlik gibi özellikleri değerlendirmez.

4.3.2. Karışıklık Matrisi

Aşağıdaki tabloda sınıflandırma modelinin başarımının daha kolay değerlendirilmesi için karışıklık matrisi verilmiştir.

Tablo 4.2. Karışıklık Matrisi

	ÖNGÖRÜLEN SINIF		
		Sınıf= 1	Sınıf= -1
	DOĞRU SINIF	Sınıf= 1	DP
		Sınıf= -1	YP
			DN

DP (Doğru Pozitif): Sınıflandırıcının pozitif olarak sınıflandırdığı ve bunun doğru olduğu test örneklerinin sayısıdır.

YN (Yanlış Negatif): Sınıflandırıcının negatif olarak sınıflandırdığı ve bunun yanlış olduğu test örneklerinin sayısıdır.

YP (Yanlış Pozitif): Sınıflandırıcının pozitif olarak sınıflandırdığı ve bunun yanlış olduğu test örneklerinin sayısıdır.

DN (Doğru Negatif): Sınıflandırıcının negatif olarak sınıflandırdığı ve bunun doğru olduğu test örneklerinin sayısıdır.

Bu durumda;

$$\text{Doğruluk} = \frac{DP + DN}{DP + YN + YP + DN} \quad (4.8)$$

$$\text{Hata Oranı} = \frac{YN + YP}{DP + YN + YP + DN} \quad (4.9)$$

Bazı durumlarda sadece doğruluk ve hata oranı sınıflandırıcıyı değerlendirmede yeterli olmaz. Örnek vermek gerekirse;

Sınıflandırıcı A	
DP=25	YN=25
YP=25	DN=25

Doğruluk =%50

Sınıflandırıcı B	
DP=50	YN=0
YP=25	DN=25

Doğruluk =%75

Sınıflandırıcı C	
DP=25	YN=25
YP=0	DN=50

Doğruluk =%75

Sadece doğruluk ve hata oranına bakılırsa yukarıdaki sınıflandırıcılardan B ve C, A'dan daha iyi bir sınıflandırıcı olduğu söylenebilir. Buna karşın B ve C arasında hangisinin daha iyi bir sınıflandırıcı olduğu sadece doğrulukla belirlenemez. Bu nedenle sınıflandırıcı model başarımını değerlendirmede kullanılan bazı ölçülere daha gereksinim vardır. Bunlar kesinlik, anma, özgüllük ve F-ölçütü'dür.

Kesinlik (Pozitif Kestirim Değeri): Doğru sınıflandırılmış pozitif örnek sayısının, pozitif sınıflandırılmış örnek sayısına oranıdır. Test sonucu pozitif olanın, gerçek bir hasta olma olasılığı nedir sorusuna cevap arar. Pozitif kestirim değeri olarak ta bilinir.

$$Kesinlik = \frac{\text{Doğru sınıflandırılmış pozitif örnek sayısı}}{\text{Pozitif sınıflandırılmış örneklerin sayısı}} = \frac{DP}{DP + YP} \quad (4.10)$$

Anma (Doğru Pozitif Oranı, Duyarlılık): Doğru sınıflandırılmış pozitif örnek sayısının, pozitif örneklerin sayısına oranıdır. Testin, gerçek hastalar içinden hastaları ayırma yeteneğidir. Doğru pozitif oranı veya Duyarlılık olarak ta adlandırılır.

$$Anma = \frac{\text{Doğru sınıflandırılmış pozitif örnek sayısı}}{\text{Pozitif örneklerin sayısı}} = \frac{DP}{DP + YN} \quad (4.11)$$

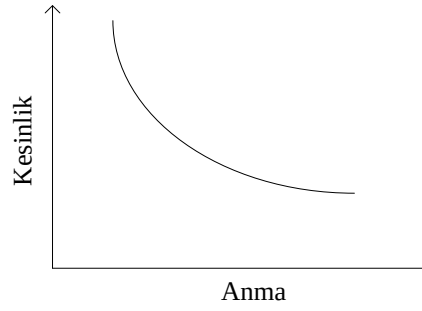
Özgüllük (Doğru Negatif Oranı): Testin, gerçek sağlamlar içinden sağlamları ayırma yeteneğidir. Analizin doğru olarak gösterdiği belirli bir hastalığa sahip olmayanların oranıdır. Aranılan hastalığın hastada bulunmaması durumunda test sonucunun negatif olma olasılığıdır. Doğru negatif oran olarak ta bilinir.

$$\text{Özgüllük} = \frac{\text{Doğru sınıflandırılmış negatif örnek sayısı}}{\text{Negatif örneklerin sayısı}} = \frac{DN}{DN + YP} \quad (4.12)$$

F-Ölçütü: Kesinlik ve anma değerlerinin harmonik ortalamasını alan bir ölçüdür. Bir sınıflandırıcının F-ölçütü aşağıdaki formülle hesaplanır:

$$F - \text{ölçütü} = \frac{2 * Kesinlik * Anma}{Kesinlik + Anma} \quad (4.13)$$

Herhangi bir sınıflandırıcıdaki kesinlik ve anma değerleri diğer sınıflandırıcıdan daha yüksek ise daha iyi bir sınıflandırıcıdır. Şekil 1’de gösterildiği gibi kesinlik ve anma arasında ters orantı vardır. C sınıflandırıcısının kesinliği daha fazladır. Buna karşın B sınıflandırıcısının anma değeri daha yüksektir. Böyle durumlarda hangisinin F-ölçütü değeri daha fazla ise o sınıflandırıcı daha iyi bir modeldir.



Şekil 4.2. Kesinlik ve anma arasındaki ilişki

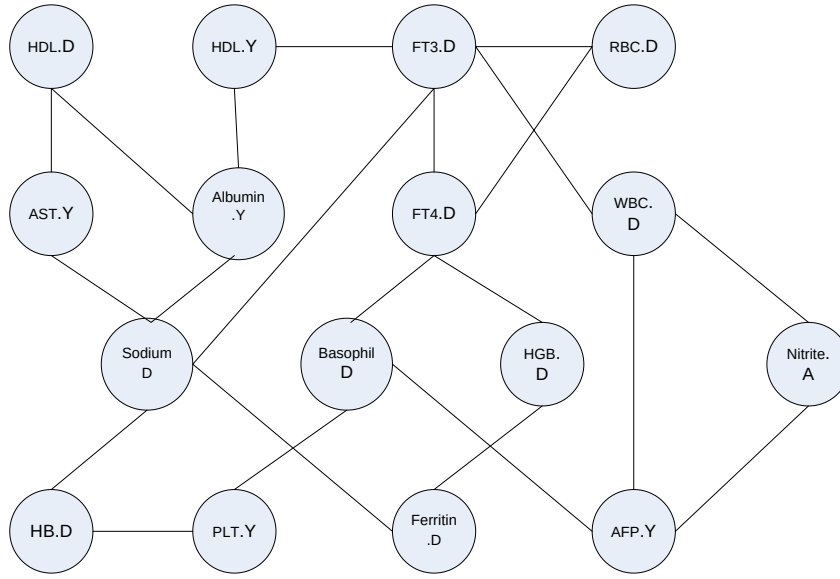
5. GELİŞEN DURUMLU HASTALIK BELİRTİLERİ AĞINDA BAĞLANTI TAHMİNİ

Bu bölümde öncelikle hastalık belirti ağı oluşturulacaktır. Hastalık belirtilerinin sosyal ağ ile modellenmesine daha önce hiçbir çalışmada rastlanmamıştır. Ağ oluşturulduktan sonra önerilen denetimli ve denetimsiz bağlantı tahmini yöntemleriyle yaşa göre hastalık belirtilerinin birbirini nasıl tetikledikleri gözlenecektir. Önerilen yöntem JAVA programlama dili kullanılarak gerçekleştirilmiştir. Veri tabanı olarak MySQL kullanılmıştır.

5.1. Hastalık Belirti Ağının Oluşturulması

Hastalık belirtileri çoğunlukla laboratuvar testlerinin sonuçlarından oluşur. Bu verilerde, bir hasta için test edilmiş birçok tıbbi parametre (hastalık belirtisi) vardır. Bu parametrelerin bütün değerleri referans aralığı içerisindeyse, hastaneye gelen kişinin bir hastalığa sahip olmadığı söylenebilir. Fakat onlardan biri veya birkaçı referans aralığı içinde değilse, diğer bir deyişle, referans aralığını aşarsa veya aralığın minimum değerinden daha düşük olursa, o zaman bu parametre anormal bir durum sergiler ve bu laboratuvar sonucuna sahip kişinin, bir hastalığa sahip olması olasıdır. Hastalık belirti ağında, düğümler referans aralığı içinde olmayan anormal parametrelerdir. Örneğin, X parametresinin değeri referans aralığının maksimum değerinden daha büyükse, bu düğüm X.Y ile temsil edilir. Burada Y yüksek anlamına gelir. Buna karşın, X parametresinin değeri, referans aralığının minimum değerinden daha düşük ise, düğüm X.D ile ifade edilir. Burada D düşük anlamını taşır. Eğer X parametresi kategorik veya ikilik bir değere sahipse, düğüm X.A ile gösterilir. Burada A, anormal anlamı ifade eder. Oluşturulan hastalık belirti ağında, kenarlar aynı hastada görülen bu anormal parametre çiftlerini bağlar. Böylece iki anormal parametre arasındaki ilişki, onlar ne zaman bir hastada birlikte görülürse oluşur. Bu tezde öncelikle, bu amaçları içeren bir hastalık belirti ağı oluşturuldu ve bir tahmin modelini inşa etmek için, anormal parametreler arasındaki bağıntılar kullanıldı.

Bu modelde, bir hastalığa sebep olan anormal parametrelerin tahmin edicisini üretmek için bağlantı tahmini yöntemlerinden faydalanıldı. Şekil 5.1, tezde oluşturulan hastalık belirti ağının bir parçasını gösterir. Hastalık kayıtları, aynı hastayı etkileyen, anormal parametrelerin birlikte meydana gelmesi hakkında önemli bir bakışı içerir. Bu veri setiyle düğümler arasındaki bağlantıların anormal parametre çiftinin en az bir hastayı etkilediği kenarlar olan hastalık belirti ağı inşa edilmiştir. Kenarlar her iki parametreyi gösteren hasta sayısı ile ağırlıklandırılmıştır.



Şekil 5.1. Hastalık belirti ağının bir parçası

5.2. Hastalık Belirti Ağları için Önerilen Denetimli Bağlantı Tahmini

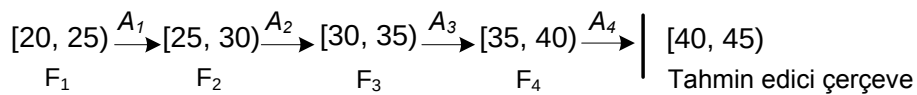
Sosyal ağlar, yeni düğümlerin veya düğümler arasındaki yeni bağlantıların eklenmesi ile zamanla gelişen dinamik yapılardır. Bundan dolayı, bu ağların gelişimi gelecekteki bağlantıların ne olacağı sorusunu akla getirir. Bu çalışmada, hastalık belirti ağı gibi ağırlıklandırılmış bir ağın gelişen yapısını ve ağdaki düğüm çiftleriyle ilgili gelişen durumları göz önüne alan denetimli stratejiye dayalı yeni bir bağlantı tahmin yöntemi önerilmiştir. Tezin amacı iki düğüm arasındaki bağlantının gücü onların birbirleriyle ne kadar sık etkileştiği ile orantılı olduğunu ifade eden Homans'ın iddiasını [79] Newman'ın yöntemiyle [80] birleştirmektir. Newman yönteminde [80], iki düğüm arasındaki ortak komşuların sayısı ne kadar fazla ise bu düğümlerin gelecekte bağlantılı olma olasılıklarının

o derece yüksek olduğunu ifade eder. Bu tezde, ağırlıklandırılmış ağ kullanıldığından dolayı, Newman'ın yöntemi şu şekilde geliştirilmiştir: “İki düğüm arasındaki ortak komşuların toplam ağırlığı ne kadar fazla ise bu iki düğümün gelecekte bağlantılı olma olasılıkları o kadar yüksektir”.

Önerilen yöntem, iki düğüm ve onların komşuları arasındaki yakınlık skorunu güçlendirecek veya zayıflatacaktır. Ağ gelişirken, belirli bir oranda bağlantıların ağırlıklarının artması veya azalması bir gelişen durum olarak tanımlanır. Ağın gelişen yapısı farklı yaş aralıklarında ağın ardışıl çerçeveleri çıkarılarak oluşturulur. Daha sonra tahmin edici nitelikler, çerçevelerin geçişleri boyunca oluşan durumla belirlenen yakınlık skoru ile her bir aday düğüm çiftleri ile tanımlanır.

5.2.1. Hastalık Belirti Ağının Gelişen Yapısı

Gelişen yapı, hastalık belirti ağında hastaların farklı yaş aralıklarına göre ağın durumlarını temsil eder. Bu maksatla tezde ağ öncelikle birkaç yaş dilimli parçalara ayrılmıştır. Daha sonra ardışıl parçalar gruplandırılarak çerçeveler inşa edilir. Örneğin, {[20-25]} yaş aralığındaki bir çerçeve (F_i), 20-25 yaşları arasındaki tüm insanların tıbbi verilerinin anormal parametreleri arasında oluşan bütün bağlantıları içeren bir alt çizgedir. Hastalık belirti ağının gelişen yapısı ve tahmin edici çerçevesinin bir örneği Şekil 5.2’de gösterilmiştir. Burada önemli bir noktaya dikkat çekmek gerekir: Her bir çerçevenin boyutu tahmin edici pencerenin uzunluğu ile aynıdır. Bu da bağlantı tahmin problemlerinde önceden belirlenir.



Şekil 5.2. Beş çerçeveli bir hastalık belirti ağının gelişen yapısı

5.2.2. Gelişen Durumlar

Gelişen bir durum genellikle düğüm çiftlerini bir durumdan (Bağlantılı veya bağlantısız) diğerine taşıyan bir harekettir. Bir gelişen durum hastalık belirti ağı gibi bir ağırlıklandırılmış ağda bir çerçeveden diğerine iki düğüm arasındaki bir bağlantının ağırlığının artması veya azalması olarak tanımlanır. Bu durumlar üç farklı tipten biri olabilirler. Bunlar tutarlı, güçlendiren ve zayıflatan durumlardır.

5.2.2.1. Tutarlı Durum

Hastalık belirti ağında bir tutarlı tutum ağ gelişirken iki düğüm arasındaki ilişkinin ağırlığı önceden belirlenen bir durum değişim oranından daha fazla yükselmez veya bu orandan daha fazla azalmazsa meydana gelir. Diğer bir deyişle, iki anormal parametre bir çerçevede bir bağlantıya sahipse, bu bağlantının güçlülüğü daha sonraki çerçevede de korunur. Böyle bir durumda $(k-1)$ 'inci çerçeveden k 'nıncı çerçeveye geçiş boyunca bir tutarlı durumu ele almak için her bir (x, y) düğümde tutarlı durum skoru $C_k(x, y)$ aşağıdaki gibi tanımlanır:

$$C_k(x, y) = \begin{cases} c & \text{if } w_k(x, y) < w_{k-1}(x, y) + r \cdot w_{k-1}(x, y) \text{ OR } w_k(x, y) > w_{k-1}(x, y) - r \cdot w_{k-1}(x, y) \\ 0 & \text{aksi takdirde} \end{cases} \quad (5.1)$$

Burada $w_{k-1}(x, y)$ ve $w_k(x, y)$, sırasıyla F_{k-1} ve F_k çerçevelerindeki x ve y düğümleri arasındaki bağlantının ağırlıklarıdır. r ($0 < r \leq 1$) önceden belirlenen durum değişim oranıdır. Çerçeveler arasındaki durumlar bu orana göre değişir. C sabiti ($0 < c \leq 1$) iki düğüm arasındaki bağlantının gücü korunduğundan tutarlı durumlar için pozitif bir skordur.

5.2.2.2. Güçlendiren Durum

Güçlendiren durum farklı çerçevelerde iki düğüm arasındaki bir bağlantının ağırlığı önceden belirlenen durum değişim oranından daha fazla yükselmesi şartında oluşur. Diğer bir deyişle, bir çerçevede iki anormal parametrenin ağırlığı sonraki çerçevede durum

değişim oranından daha fazla artarsa güçlendiren durum oluşur. (x, y) düğüm çifti ve F_k çerçevesiyle ilişkili $S_k(x, y)$ güçlendiren durum skoru aşağıdaki gibi hesaplanır:

$$S_k(x, y) = \begin{cases} s & \text{if } w_k(x, y) > w_{k-1}(x, y) + r \cdot w_{k-1}(x, y) \\ 0 & \text{aksi takdirde} \end{cases} \quad (5.2)$$

Burada $s(0 < s \leq 1)$ güçlendiren durumlar için bir ödülü gösterir. Bu değer iki düğüm arasındaki bağıntıyı güçlendirdiğinden dolayı pozitif olmalıdır.

5.2.2.3. Zayıflatan Durum

Zayıflatan durum güçlendiren durumun tersidir. Bu durum çeşidinde iki düğüm arasındaki var olan bir bağlantının gücü bir çerçeveden bir sonrakine azalır. Zayıflatan durum skoru $D_k(x, y)$ aşağıdaki gibi hesaplanır:

$$D_k(x, y) = \begin{cases} d & \text{if } w_k(x, y) < w_{k-1}(x, y) - r \cdot w_{k-1}(x, y) \\ 0 & \text{aksi takdirde} \end{cases} \quad (5.3)$$

Burada $d(-1 \leq d < 0)$ bir azalma değeridir ve iki düğüm arasındaki bağlantının gücü azaldığından dolayı negatif olmalıdır.

5.2.3. Denetimli Bağlantı Tahmini için Tahmin Edici Niteliklerin ve Sınıf Seviyesinin Belirlenmesi

Bu tezde göz önüne alınan düğüm çiftleriyle doğrudan ilgili gelişen durumlar olan birincil durumların skorlarıyla, düğümün komşularında gözlenen gelişen durumlar olan ikincil durumların skorları birleştirilmiştir. Verilen (x, y) düğüm çiftiyle ilgili bu skor aşağıdaki gibi hesaplanır:

$$A_k(x, y) = P(x, y, k) + \alpha S(x, y, k) \quad (5.4)$$

$$P(x, y, k) = \begin{cases} C_k(x, y) & \text{Eğer tutarlı durum gözlenirse} \\ S_k(x, y) & \text{Eğer güçlendiren durum gözlenirse} \\ D_k(x, y) & \text{Eğer zayıflatılan durum gözlenirse} \end{cases} \quad (5.5)$$

$$S(x, y, k) = \sum_{z \in I(x) \cap I(y)} P(x, z, k) + P(z, y, k) \quad (5.6)$$

$P(x, y, k)$, $k-1$ çerçevesinden k çerçevesine geçişte gözlemlenen (x, y) düğüm çifti için gelişen durum skorunu hesaplar. Bu durumda gözlemlenen durum, tutarlı, güçlendiren ve zayıflatılmış durumlardan biri olabilir. $S(x, y, k)$, (x, y) düğüm çiftinin ikincil durumların toplam skorlarını verir. Burada $I(x)$, oluşturulan ağda x düğümünün komşu dizi setidir. α parametresi, ikincil durumların x ve y arasındaki ilişkiyi ne kadar güçlü etkilediğini gösteren bir kontrol faktörüdür.

Çerçevelerin her bir geçişi boyunca birincil ve ikincil durumlardan elde edilen bu birleştirilmiş skorlar düğüm çiftlerinin özelliğini tanımlayan tahmin edici nitelik değeri olarak kullanılır. Şekil 5.2’de gösterildiği gibi her bir geçiş için bir (A_i) niteliği belirlenir. Bu tezde, düğümler etrafındaki bağıntılarla ilgili bilgi bizim için önemli olduğundan komşuluk tabanlı ölçü kullanıldı. Tutarlı, güçlendiren ve zayıflatılan durumlar için verilen c , s ve d değerleri geçerlilik kümesi bağlantı tahmini performansı değerlendirilerek hesaplanır. Geçerlilik kümesi tahmin edici çerçeveden önceki en son çerçevedir. Bağlantı tahmininin amacı, bir ağda yeni bir bağlantıyı veya ağırlığı yükselmiş bir bağlantıyı araştırmak olduğundan göz önüne alınacak bağlantılı olmayan yada aralarındaki bağlantısı zayıf olan düğüm çiftleri geçerlilik kümesinden seçilir. İki düğüm arasındaki bir bağlantının zayıf olması demek, bu bağlantının ağırlığının o çerçevedeki bütün ağırlıkların toplamının *Eşik* olarak adlandırılan belirli bir yüzdesinden daha küçük olması demektir. Fakat ağın gelişimi boyunca doğrudan veya dolaylı etkilenmemiş bağlantılı olmayan çiftler elenir. Eğer geçerlilik kümesinde bağlantılı olmayan bir çift, tahmin edici çerçevede bağlantılı bir çifte dönüşürse veya önceden belirlenen *Eşik* değerinden daha küçük bir ağırlığa sahip bir çiftin ağırlığı geçerlilik kümesinden tahmin edici çerçeveye durum değişim oranından (r) daha fazla bir oranda artarsa; düğüm çifti pozitif sınıf olarak kabul edilir. Aksi durumda sınıf negatiftir.

Yöntemin performansını değerlendirmek için önerilen durum tabanlı skorlar bağlantı tahmini için literatürde iyi bilinen ölçülerle karşılaştırıldı. Bu tezde oluşturulan ağın kenarları anormal parametreye sahip hastaların sayılarıyla ağırlıklandırıldığından dolayı

karşılaştırmada ağırlıklı ağlar için geliştirilen yakınlık skorları kullanıldı. Bu nedenle beş yakınlık skoru göz önüne alındı. Bunlar Ortak Komşular [4], Jaccard Katsayısı [7], Tercihli Bağlılık [5], Adamic–Adar Katsayısı [6] ve Kaynak Paylaştırma İndeksidir [81]. Bu tezde oluşturulan tüm ağlar ağırlıklı olduklarından yakınlık skorlarının tanımı da buna göre değişiklik gösterir. Aşağıda tez çalışması boyunca kullanılacak beş yakınlık ölçütünün ağırlıklı ağlar için geliştirilmiş hali verilmektedir. Buna göre: x , G ağırlıklı ağının bir düğümü olsun ve $\Gamma(x)$ G 'de x 'in komşular kümesini göstere. Kullanılan yaklaşımların çoğu “ x ve y düğümlerinin komşuları olan $\Gamma(x)$ ve $\Gamma(y)$ ne kadar fazla kesişirse iki düğümün gelecekte bir bağlantı oluşturma olasılığı o kadar fazladır” fikrine dayalıdır. Diğer bir deyişle, eğer ağda (x,z) ve (y,z) mevcutsa (x,y) 'nin bağlantı oluşturma olasılığı fazladır.

- **Ortak Komşular (OK):** Hastalık belirti ağı gibi ağırlıklı ağlar için ortak komşular ölçüsü aşağıdaki gibi hesaplanır:

$$OK(x, y) = \sum_{z \in |\Gamma(x) \cap \Gamma(y)|} \omega(x, z) + \omega(y, z) \quad (5.7)$$

- **Jaccard Katsayısı (JK):** Ağırlıklı ağlarda Jaccard benzerlik ölçütü aşağıda verilen denklemdeki gibi hesaplanır:

$$JK(x, y) = \sum_{z \in |\Gamma(x) \cap \Gamma(y)|} \frac{\omega(x, z) + \omega(y, z)}{\sum_{a \in \Gamma(x)} \omega(a, x) + \sum_{b \in \Gamma(y)} \omega(b, y)} \quad (5.8)$$

- **Tercihli Bağlılık (TB):** Tezde oluşturulan ağlarda tercihli bağlılık bir x düğümünden yeni bir bağlantıyı oluşturma olasılığı, y düğümünün toplam ağırlığıyla ilişkili olduğunu söyler. Diğer bir deyişle tercihli bağlılık aşağıdaki gibi tanımlanır:

$$TB(x, y) = \sum_{a \in \Gamma(x)} \omega(a, x) * \sum_{b \in \Gamma(y)} \omega(b, y) \quad (5.9)$$

- **Adamic-Adar Katsayısı (AA):** Adamic-Adar benzerlik ölçütü Jaccard benzerlik ölçütünün değiştirilmesiyle formüle edilmiştir. Bu ölçüt az komşulara sahip ortak komşulukların önemini artırır. Bu şekilde, özel (ya da güçlü) bir ortak komşuya sahip olmak düğüm çiftlerinin ileride kendi aralarında bağlantı oluşturacağı ihtimalini artırır.

Ağırlıklı ağlarda Adamic-Adar benzerlik ölçütü aşağıdaki gibi hesaplanır:

$$AA(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{\omega(x, z) + \omega(y, z)}{\log(1 + \sum_{c \in \Gamma(z)} \omega(z, c))} \quad (5.10)$$

- **Kaynak Paylaştırma İndeksi (KP):** Ağırlıklı ağlarda kaynak paylaştırma indeksi aşağıdaki gibi bulunur:

$$KP(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{\omega(x, z) + \omega(y, z)}{\sum_{c \in \Gamma(z)} \omega(z, c)} \quad (5.11)$$

5.2.4. Gelişen Durumlu Hastalık Belirti Ağında Denetimli Bağlantı Tahmini Algoritması

Yaşa göre çerçeveslendirilen ve bu sayede bağlantıları arasında farklı durumlar gözlenen hastalık belirti ağı için önerilen denetimli bağlantı tahmini algoritması aşağıda verilmiştir:

1. Geçerlilik kümesinde bağlantılı olmayan çiftler veya önceden belirlenen eşik değerinden daha küçük ağırlık derecesine sahip 400 çift rasgele seçilir ve ilgili nitelikler böylece 400 örnek oluşturur. Bu örneklerin 200'ü pozitif geri kalanı negatif olarak seçilir. Bu şekilde sınıflar denetimli öğrenme için eşit bir şekilde dağıtılır.
2. Seçilen çiftlerin nitelik sayısı önerilen yöntemde çerçeve sayısı kadardır. Bu durumda her bir geçiş için birincil ve ikincil durumlardan elde edilen birleştirilmiş skor (A_i) değeri hesaplanır. Geleneksel yaklaşımlarda ise bu nitelik değerleri farklı bağlantı tahmin edicilerle sadece geçerlilik kümesinden elde edilir. Yani ağın gelişen durumu göz önüne alınmaz. Bu tezde karşılaştırma amaçlı 5 farklı tahmin

edici kullanıldığından nitelikler çerçeve sayısına bağlı olarak rasgele 5 tahmin ediciden belirlenir.

3. Farklı nitelik belirleme yöntemlerine göre göz önüne alınan bütün sınıflandırıcıların performans ölçüleri hesaplanır.

5.2.5. Uygulama Sonuçları

Bu bölümde önerilen yöntemi değerlendirmek için gerçek tıp verileri kullanarak bazı uygulamalar gerçekleştirilmiştir. Bu maksatla öncelikle 210.134 hastadan oluşan dört farklı laboratuvar sonuçları bir araya getirilmiştir. Bu laboratuvarlar tam kan sayımı, klinik biyokimya kan testi, idrar testi ve hormon testidir. Daha sonra, oluşturulan hastalık belirti ağının değişik versiyonları için bu testlerin anormal parametreleri arasında bağlantıları tahmin etme üzerine odaklanılmıştır.

5.2.5.1. Hastalık Belirti Ağının Versiyonları

Yöntemin etkinliğini değerlendirmek için yapılan uygulamalarda Fırat Üniversitesi Hastanesinden elde edilen laboratuvar sonuçları kullanılmıştır. Hastaneden girişimsel olmayan etik kurul raporu alınarak, hastaların tıbbi kayıtları elde edilmiştir. Daha önceden ifade edildiği gibi, düğümler anormal parametreleri gösterirken, kenarlar bu anormal parametrelerin çifti en az bir hastayı etkilemesi durumunda meydana gelir. Sürekli değere sahip bir parametre referans aralığı içinde değilse, normalden daha düşük veya daha yüksek şeklinde iki düğümle temsil edildiğinden 235 hastalık parametresinden toplam 386 düğüm elde edilmiştir. Bu çalışmada ağın 3 farklı versiyonu kullanılmıştır.

- (1) Ağın ağırlıklandırılmamış versiyonu: Burada en az bir hastada birlikte görülen anormal parametre çiftleri arasındaki bağlantılar sadece sunulur.
- (2) Ağın ağırlıklı versiyonu: Burada her bir bağlantı anormal parametrelerin her ikisini gösteren hastaların sayısıyla ağırlıklandırılır. Durum değişim oranı (r) %20'dir.
- (3) Ağın ağırlıklı farklı versiyonu: Durum değişim oranı (r) %40'dır.

Bağlantıdaki ağırlıklar, parametre arasındaki sadece ikili ilişkiyi göz önüne almak yerine iki hastalık belirtisi arasındaki bağlantının ne kadar güçlü olduğunu gösterir. Ağın değişik versiyonlarıyla uygulama yapmanın amacı, ilişkideki ağırlıkları kullanmanın ve farklı

durum deęişim oranlarının daha iyi tahmin sonuçlarını sağlayıp sağlamadığını incelemektir.

5.2.5.2. Deęerlendirme Sonuçları

Uygulamalarda çerçeveler literatürdeki çalışmalardan farklı olarak hastanın yaşına göre belirlenir. Bundan dolayı, analizin amacı hastalık belirti ağının hastaların yaşlarına göre gelişip gelişmediğinin saptanmasıdır. Bu maksatla iki farklı uygulama kümesi gerçekleştirilmiştir. İlk uygulama kümesi daha genç hastalardan oluşan yapı üzerinedir. Diğer yandan daha yaşlı hastalardan oluşan çerçeve uygulamaların ikinci kümesini oluşturmaktadır. Her bir küme üç farklı uygulama içerir. Bu uygulamalarda bağlantı tahmini için NaiveBayes (Naive Bayes sınıflandırıcının gerçekleştirilmesi), LibSVM (Destek vektör makineleri için yaygın bir kütüphane), IBk (K-NN algoritmasının gerçekleştirilmesi) ve J48 (C4.5 karar ağacı algoritmasının gerçekleştirilmesi) sınıflandırma yöntemleri kullanılmıştır. Bu algoritmaların tamamı WEKA ortamında mevcuttur. Bütün uygulamalar WEKA'nın başlangıç parametreleriyle yürütülmüştür. Önerilen algoritmalar ve geleneksel yaklaşımlar üç farklı ağ versiyonu üzerinde WEKA'nın uygulayıcısında çalıştırılmıştır. Yöntemler 10 kat çapraz geçерleme ile deęerlendirilmiştir. Bu süreç daha gerçekçi sonuçların elde edilmesi için her bir algoritma ve veri tabanı için 10 defa tekrarlanmıştır.

Uygulamaların ilk kümesinde öncelikle hastalık belirti ağının gelişen yapısı oluşturulmuştur. Bu oluşumda gelişen yapının çerçeveleri laboratuvar sonuçları için [20,25), [25,30), [30,35) ve [35,40) yaş aralıklarında belirlenmiştir. [40,45) yaş aralığındaki durum ise tahmin edici çerçeve kullanılmıştır. Geçerlilik kümesinde bağlantılarına bakılacak düğüm çiftlerinin maksimum ağırlığı, o çerçevedeki toplam ağırlığın *Eşik* olarak adlandırılan yüzdesi kadardır. Bu tezde hastalık belirti ağlarıyla ilgili yapılan bütün uygulamalarda *Eşik* %0,5 olarak belirlenmiştir. Yani geçerlilik kümesindeki tüm ağırlıkların toplamının %0,5'inden daha küçük ağırlıklı bağlantılar ele alınmıştır. Bu 5 boyutlu bir çerçevede yaklaşık olarak 50'ye karşılık gelmektedir. Bağlantı tahmini için c , s , d ve α parametrelerin en uygun deęerlerini deneysel olarak deęerlendirmek ve belirlemek için geçerlilik kümesi kullanılmıştır. Yapılan testlerle en uygun deęerler olarak $c = 0,45$, $s = 0,85$, $d = -0,35$ ve $\alpha = 0,07$ bulunmuştur. Bu sonuçlar tutarlı (c) ve zayıflayan (d)

değerlerin neredeyse birbirini dengelediğini gösterir. En iyi performans değerleri $c < s$ durumunda elde edilmiştir.

Bu uygulamada ağın gelişen yapısındaki çerçeve sayısı 4 olduğundan nitelik sayısı da 4 olarak belirlenmiştir. Yani her bir çerçeve geçişindeki skorlar ilgili düğümün nitelik değeri olarak alınmıştır. Ağırlıklı ağlar için geliştirilmiş geleneksel tahmin edicilerle önerdiğimiz durum tabanlı yöntemi karşılaştırmak için denetimli bağlantı tahmininde tahmin edici nitelikler olarak OK, JK, TB, AA ve KP'dan dördü rastgele olarak belirlenmiştir. Hastalık belirti ağının oluşturulmasında kullanılan parametreler daha önce de ifade edildiği gibi 4 farklı laboratuvar sonucundan çıkarılmıştır. Tez'de 235 parametreyi ve referans aralıklarının tamamını vermek yerine her bir laboratuvardan geçerlilik kümesi [35,40) yaş aralığı olan küme için şikayetleri en çok görülen ilk 10 parametreye ve bunların referans aralıklarına yer verilmiştir. Bu bağlamda Tablo 5.1 tam kan sayımında, Tablo 5.2 klinik biyokimyada, Tablo 5.3 idrar testinde ve Tablo 5.4 hormon testindeki ilgili parametreleri ve görülme sıklığını gösterir.

Tablo 5.1. Tam kan sayımı testindeki ilk 10 anormal parametre ve bunların görülme sıklığı

Anormal Parametre	Referans Aralığı	Sıklığı
MCHC.Yüksek	28-34	5458
HGB.Düşük	10-16	5103
HCT.Düşük	33-57	4847
WBC.Yüksek	3,8-10,3	4241
PLT.Düşük	140-360	3877
BA#.Yüksek	0-0,6	3235
MPV.Yüksek	7-9	2780
MGH.Yüksek	24-31	2498
RBC.Düşük	4,1-6,0	2311
RDW-Düşük	12-15	1892

Tablo 5.2. Klinik Biyokimya testindeki ilk 10 anormal parametre ve bunların görülme sıklığı

Anormal Parametre	Referans Aralığı	Sıklığı
Trigliserit.Yüksek	40-180	8447
LDL Kolesterol.Yüksek	< 130	8015
Kolesterol.Yüksek	120-200	6889
LDH.Yüksek	120-246	6097
Glukoz.Yüksek	75-115	5754
Kalsiyum.Düşük	8,5-10,8	5025
AST.Yüksek	5-40	4132
ALT.Yüksek	5-40	3814
Potasyum.Yüksek	3,5-5,5	2408
Albumin.Düşük	3,5-5,3	2358

Tablo 5.3. İdrar testindeki ilk 10 anormal parametre ve bunların görülme sıklığı

Anormal Parametre	Referans Aralığı	Sıklığı
Lökosit.Anormal	Negatif	3974
Protein.Anormal	Negatif	3012
PH.Düşük	6,0-8,0	2877
Dansite.Yüksek	< 1,030	2408
Keton.Anormal	Negatif	2107
Nitrite.Anormal	Negatif	1780
Protein.Anormal	Negatif	1552
Uric Asit.Yüksek	1,5-8,0	1124
Bilirubin.Anormal	Negatif	754
Glukoz.Anormal	Negatif	447

Tablo 5.4. Hormon testlerindeki ilk 10 anormal parametre ve bunların görülme sıklığı

Anormal Parametre	Referans Aralığı	Sıklığı
Beta HCG.Yüksek	0-5	3874
T3.Düşük	1,57-4,71	2987
TSH.Yüksek	0,5-5,5	2741
T4.Düşük	0,93-1,71	2630
Folik Asit.Düşük	3,1-17,5	2252
B12.Yüksek	174-878	2068
Ferritin.Düşük	7,0-306,5	1827
Kortizol.Yüksek	2,3-19,4	1308
Parathormon.Yüksek	19,8-74,9	1164
PSA.Yüksek	0-6,5	943

İlk uygulama önerilen durum tabanlı skorla geleneksel yakınlık skorlarını karşılaştırmak için hastalık belirti ağının üç farklı versiyonunun doğruluk değerini bulmayla ilgilidir. Tablo 5.5 bu uygulamanın sonuçlarını rapor eder. Tablo 5.5’den görülebileceği gibi, en iyi tahmin performansları ağın ağırlıklandırılmış birinci versiyonu ($r=0.20$) üzerinde bizim bağlantı tahminimizle başarılımıştır. Bu tecrübeye ağın ağırlıklandırılmamış versiyonunda daha kötü sonuçlar elde edilir.

Tablo 5.5. Hastalık belirti ağında karşılaştırılan yöntemlerin doğruluk değerleri (%)

Yöntem	Ağırlıklandırılmamış Ağ		Ağırlıklandırılmış Ağ 1		Ağırlıklandırılmış Ağ 2	
	<u>Önerilen Yöntem</u>	<u>Geleneksel Ölçülere dayalı</u>	<u>Önerilen Yöntem</u>	<u>Geleneksel Ölçülere dayalı</u>	<u>Önerilen Yöntem</u>	<u>Geleneksel Ölçülere dayalı</u>
NaiveBayes	75,39	71,27	78,14	73,15	76,15	70,11
LibSVM	72,47	70,10	73,21	70,80	70,81	68,55
IBk	68,83	66,77	70,35	65,47	67,19	68,86
J48	77,45	73,55	80,21	74,55	73,74	69,48

Benzer şekilde ikinci uygulamada ağın her üç versiyonunun kesinlik değerlerini bulur. Tablo 5.6’da ağın ağırlıklandırılmış birinci versiyonu, geleneksel yakınlık ölçüleriyle öğrenmede daha kötü sonuçlar sergilerken, önerdiğimiz yöntemde daha iyi sonuçlar gösterdiği gözlenir. İlk kümenin son uygulaması ağın her üç versiyonunun anma değerlerini karşılaştırır. Tablo 5.7’de verilen durum tabanlı skorlar göz önüne alınan bütün öğrenme yöntemlerinde en iyi sonucu yakalamışlardır.

Tablo 5.6. Hastalık belirti ağında karşılaştırılan yöntemlerin kesinlik değerleri (%)

	Ağırlıklandırılmamış Ağ		Ağırlıklandırılmış Ağ 1		Ağırlıklandırılmış Ağ 2	
	<u>Önerilen Yöntem</u>	<u>Geleneksel Ölçülere dayalı</u>	<u>Önerilen Yöntem</u>	<u>Geleneksel Ölçülere dayalı</u>	<u>Önerilen Yöntem</u>	<u>Geleneksel Ölçülere dayalı</u>
NaiveBayes	72,49	70,18	75,86	72,71	73,41	70,57
LibSVM	65,15	62,54	68,24	65,48	66,27	63,06
IBk	62,11	60,77	65,53	64,50	65,21	62,48
J48	70,88	67,15	74,92	70,44	72,19	68,25

Tablo 5.7. Hastalık belirti ağında karşılaştırılan yöntemlerin anma değerleri (%)

	Ağırlıklandırılmamış Ağ		Ağırlıklandırılmış Ağ 1		Ağırlıklandırılmış Ağ 2	
	<u>Önerilen Yöntem</u>	<u>Geleneksel Ölçülere dayalı</u>	<u>Önerilen Yöntem</u>	<u>Geleneksel Ölçülere dayalı</u>	<u>Önerilen Yöntem</u>	<u>Geleneksel Ölçülere dayalı</u>
NaiveBayes	80,22	76,79	83,21	80,04	85,50	83,10
LibSVM	79,53	70,84	81,18	70,28	74,56	71,25
IBk	76,18	74,11	73,52	66,88	71,52	66,88
J48	81,15	79,15	85,29	82,22	86,14	84,22

Uygulamaların ikinci kümesi daha yaşlı hastalarda yukarıda tanımlanan aynı süreci araştırır. Bu yeni uygulamalarda hastalık belirti ağı oluşturulurken, [40,45), [45,50), [50,55), [55,60) yaş aralığındaki çerçeveler ağın gelişen yapısı olarak belirlendi. [60,65) yaş aralığındaki durum ise tahmin edici çerçeve olarak kullanılmıştır. İlgili uygulama sonuçları bu yeni çerçevelerde yukarıdaki üç durumu açıklayan Tablo 5.8-5.10'da gösterilmiştir. Gözlemler daha genç hastalı çerçeveler için tanımlamalar benzerdir. Fakat uygulamaların ilk kümesinden farklı olarak daha yaşlı hastaların ağırlıklandırılmış ikinci versiyonu ($r=40\%$) hemen hemen bütün durumlarda daha iyi performans elde etmiştir. Bu beklenen durumdur. Çünkü bu yeni ağda bağlantıların sayısı ve ağırlıkların değeri önceki olandan daha fazladır. Böyle bir ortamda, durum değişim oranını yükseltmek daha az durum geçişlerine (Yani, tutarlı durumdan zayıf duruma doğru) izin verir. Bu davranış açıkça gösterir ki, ağ topolojisinde tahmin etme maksatlı kullanılabilecek saklı bir bilgi vardır. Sonuç olarak uygulamalar hastalık belirti ağına uygulandığı zaman bağlantı analizinin somut ilgisini kanıtlar. Ayrıca çıkarılabilecek diğer bir sonuç, gelecek hastalık belirtilerini bulmada hastalık belirti ağının topolojisi korelasyonların geniş bir sayısını barındırmasıdır.

Tablo 5.8. Daha yaşlı hastalı hastalık belirti ağında karşılaştırılan yöntemlerin doğruluk değerleri(%)

Yöntem	Ağırlıklandırılmamış Ağ		Ağırlıklandırılmış Ağ 1		Ağırlıklandırılmış Ağ 2	
	<u>Önerilen Yöntem</u>	<u>Geleneksel Ölçülere dayalı</u>	<u>Önerilen Yöntem</u>	<u>Geleneksel Ölçülere dayalı</u>	<u>Önerilen Yöntem</u>	<u>Geleneksel Ölçülere dayalı</u>
NaiveBayes	74,55	70,23	75,48	72,21	81,14	75,66
LibSVM	73,46	68,75	71,87	69,53	79,12	73,06
IBk	68,23	66,59	70,21	67,77	73,56	69,58
J48	72,44	69,88	74,12	73,21	80,23	77,95

Tablo 5.9. Daha yaşlı hastalı hastalık belirti ağında karşılaştırılan yöntemlerin kesinlik değerleri (%)

	Ağırlıklandırılmamış Ağ		Ağırlıklandırılmış Ağ 1		Ağırlıklandırılmış Ağ 2	
	<u>Önerilen Yöntem</u>	<u>Geleneksel Ölçülere dayalı</u>	<u>Önerilen Yöntem</u>	<u>Geleneksel Ölçülere dayalı</u>	<u>Önerilen Yöntem</u>	<u>Geleneksel Ölçülere dayalı</u>
NaiveBayes	71,12	68,22	73,23	70,88	78,66	73,57
LibSVM	70,46	65,51	70,85	67,57	74,88	71,23
IBk	66,23	63,09	66,55	64,16	69,27	66,70
J48	69,28	66,47	72,21	69,47	78,21	74,12

Tablo 5.10. Daha yaşlı hastalı hastalık belirti ağında karşılaştırılan yöntemlerin anma değerleri (%)

	Ağırlıklandırılmamış Ağ		Ağırlıklandırılmış Ağ 1		Ağırlıklandırılmış Ağ 2	
	<u>Önerilen Yöntem</u>	<u>Geleneksel Ölçülere dayalı</u>	<u>Önerilen Yöntem</u>	<u>Geleneksel Ölçülere dayalı</u>	<u>Önerilen Yöntem</u>	<u>Geleneksel Ölçülere dayalı</u>
NaiveBayes	75,18	72,10	84,57	75,67	83,88	79,12
LibSVM	76,21	70,23	74,48	72,03	80,78	75,78
IBk	70,64	71,69	72,48	70,10	76,14	72,73
J48	77,54	74,19	87,15	76,14	84,12	77,51

5.3. Hastalık Belirti Ağları için Önerilen Denetimsiz Bağlantı Tahmini Algoritması

Bu bölümde ağırlıklı olarak oluşturulan hastalık belirti ağları için yeni bir denetimsiz bağlantı tahmini algoritması önerilmiştir. Hastalık belirti ağları yaşa göre gelişen bir şekilde çerçevelere ayrılmıştır. Her bir çerçevede belirli yaş grupları arasındaki hastaların anormal parametreleri tutulmaktadır. Tahmin edilecek düğüm çiftleri için yakınlık değerleri aşağıdaki gibi belirlenir:

$$Skor(x, y) = \sum_{k=2}^n P(x, y, k) + \alpha S(x, y, k) \quad (5.12)$$

$$P(x, y, k) = \begin{cases} C_k(x, y) & \text{Eğer tutarlı durum gözlenirse} \\ S_k(x, y) & \text{Eğer güçlendiren durum gözlenirse} \\ D_k(x, y) & \text{Eğer zayıflatıcı durum gözlenirse} \end{cases} \quad (5.13)$$

$$S(x, y, k) = \sum_{z \in I(x) \cap I(y)} P(x, z, k) + P(z, y, k) \quad (5.14)$$

Önceki bölümde olduğu gibi $P(x, y, k)$ $k-1$ 'nci çerçeveden k 'inci çerçeve geçişte gözlenen (x, y) düğüm çifti için birincil durum skorudur. Bu anda gözlemlenen durum üç gelişen durumdan biri olabilir. $S(x, y, k)$ ise (x, y) ile ilgili ikincil durum skorudur. Uygulamalarda, önerilen yakınlık ölçüsü daha önce bilinen ve ağırlıklı ağlar için değiştirilmiş olan beş farklı yakınlık ölçütüyle karşılaştırılacaktır. Bunlar yine Ortak Komşular, Jaccard Katsayısı, Tercihli Bağlılık, Adamic-Adar ve Kaynak Paylaşımli indekslemedir.

Hastalık belirti ağı için önerilen denetimsiz bağlantı tahmini algoritması aşağıdaki gibidir:

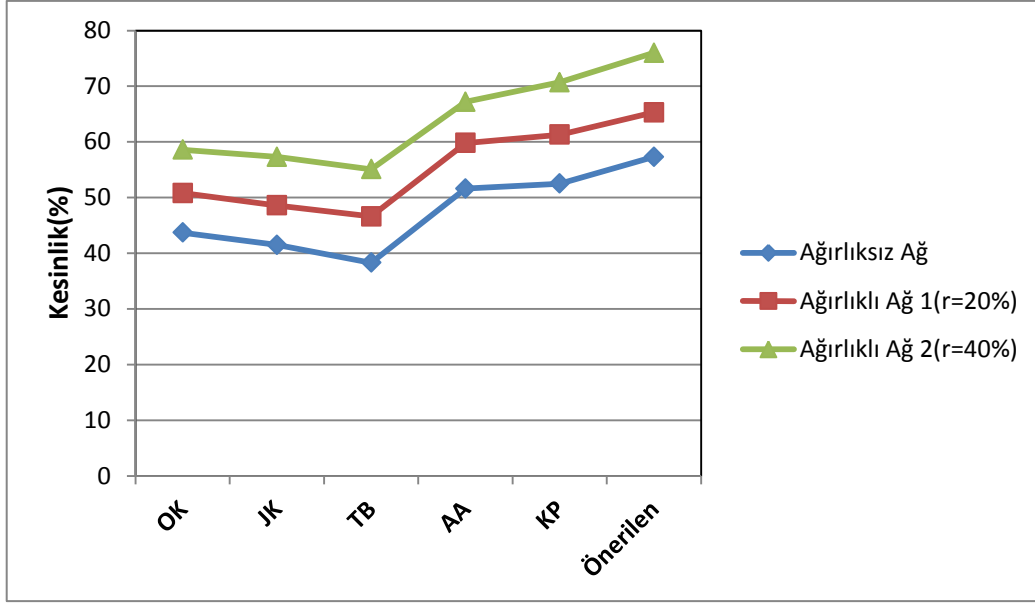
1. Geçerlilik kümesinden bağlantılı olmayan çiftler veya önceden belirlenen eşik değerinden daha küçük ağırlık derecesine sahip çiftler seçilir. Buna karşın ağın gelişimi boyunca herhangi bir durumla doğrudan veya dolaylı olarak etkilenmemiş bağlantılı olmayan çiftler yine bir bağlantı oluşturmaz düşüncesiyle elenir.
2. Seçilen çiftlerin yakınlık skorları altı farklı tahmin ediciye göre hesaplanır.
3. Yakınlık değeri en yüksek ilk 400 çift test için seçilir. Bu yüksekten azalan şekilde sıralanmış çiftler muhtemelen oluşacağı düşünülen gelecek bağlantılar olarak kabul edilir.
4. Eğer geçerlilik kümesinde bağlantılı olmayan bir çift, tahmin edici çerçevede bağlantılı bir çifte dönüşürse veya önceden belirlenen eşik değerinden daha küçük bir ağırlığa sahip bir çiftin ağırlığı geçerlilik kümesinden tahmin edici çerçeveye durum değişim oranından (r) daha fazla bir oranda artarsa; düğüm çifti pozitif sınıf olarak kabul edilir. Aksi durumda sınıf negatiftir.
5. Göz önüne alınan bütün tahmin edicilerin performans ölçüleri hesaplanır.

5.3.1. Değerlendirme Sonuçları

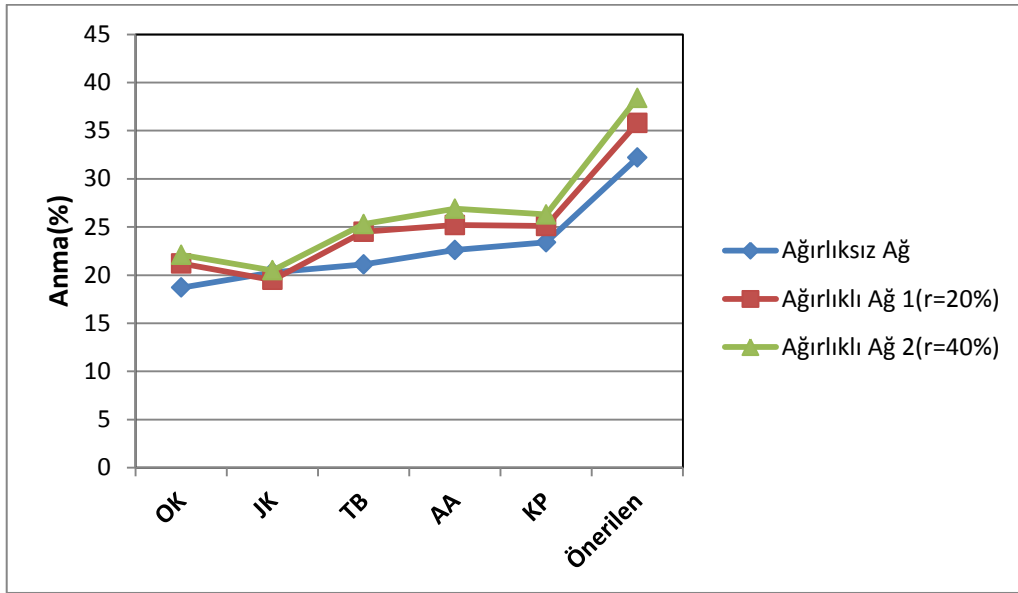
Önerilen yöntemin performansını değerlendirmek için önceki bölümde verilen hastalık belirti ağının ikisi ağırlıklı olan yine üç versiyonu kullanılmıştır. Çerçeveler hastaların yaşlarına göre oluşturulmuştur. Uygulamaların ilk kümesi daha geniş aralıklı çerçevelerden oluşan yapı için gerçekleştirilmiştir. Daha küçük aralıklı çerçeveler ikinci uygulama kümesinde ele alınmışlardır. Her bir küme için üç farklı test yapılmıştır.

Uygulamaların ilk kümesinde hastalık belirti ağının gelişen yapısı oluşturulmuştur. Bu yapıda [20,30), [30, 40) ve [40, 50) yaş aralığındaki çerçeveler belirlenmiştir. [50, 60) yaş aralığındaki ağ yapısı ise tahmin edici çerçeve olarak kullanılmıştır. Geçerlilik kümesi tahmin edici çerçeveden önceki son çerçevedir. Ağın gelişimi boyunca doğrudan ve dolaylı olarak herhangi durumdan etkilenmemiş bağlantılı olmayan düğüm çiftleri göz önüne alınmaz. Düğümleri bağlayan kenarın ağırlığı geçerlilik kümesinden tahmin edici çerçeveye durum değişim oranından daha fazla artarsa tüm tahmin ediciler için pozitif olarak tahmin edilir. Aksi halde bağlantının tahmin edici çerçevede de devam etmediği kabul edilir. En iyi performans $c = 0,4$, $s = 0,9$, $d = -0,3$ ve $\alpha = 0,07$ değerlerinde elde edilmiştir. *Eşik* olarak yine %0,5 belirlenmiştir. Ancak çerçeve boyutu 10 olan bir geçerlilik kümesinde zayıf bağlantı 100'den daha küçük ağırlıklar için ifade edilmiştir.

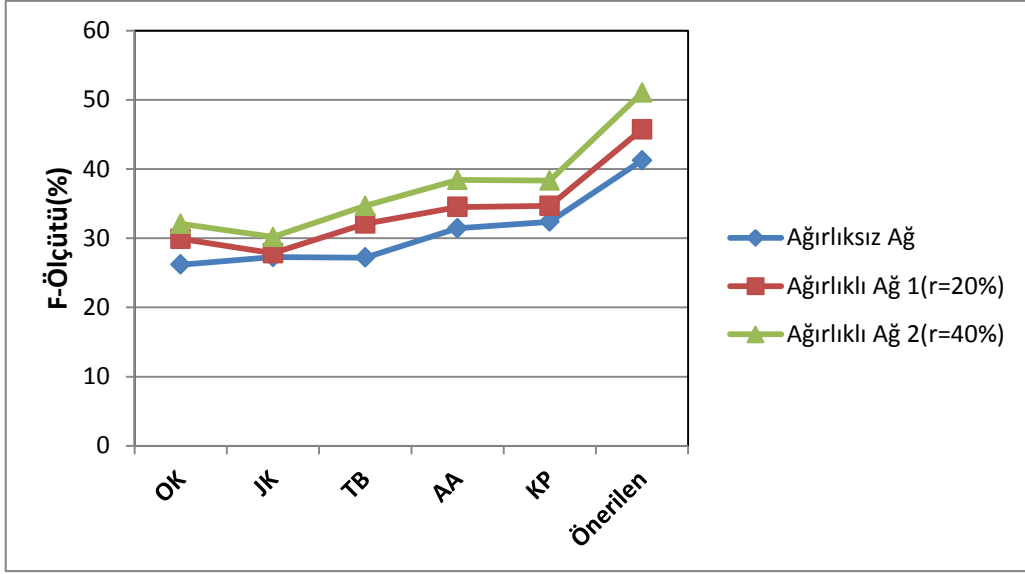
İlk uygulama önerilen durum tabanlı skoru ağırlıklı ağlar için geliştirilmiş geleneksel yakınlık skorlarıyla karşılaştırmak için hastalık belirti ağının üç farklı versiyonunun kesinlik değerini bulmakla ilgilidir. Şekil 5.3 bu uygulamanın sonuçlarını gösterir. Şekil 5.3'den görüleceği gibi en iyi tahmin ağırlıklı ağın ikinci versiyonu üzerinde tahmin edicimiz tarafından elde edilmiştir. Bu uygulamada ağın ağırlıksız versiyonu en kötü sonuçları gösterendir. Benzer şekilde ikinci uygulama ağın her üç versiyonunun anma değerlerini bulur. Şekil 5.4'deki sonuçlar ağın ikinci ağırlıklı versiyonunun geleneksel yakınlık ölçütlerinde diğerlerine göre fazla bir başarı elde edemediği, yöntemimiz için ise daha iyi sonuçlar yakaladığını göstermiştir. İlk kümenin son uygulaması ağların F-ölçütü değerlerini karşılaştırır. Şekil 5.5'de eğrilere göre durum tabanlı skorumuz göz önüne alınan bütün yöntemler içerisinde en iyi sonucu verendir.



Şekil 5.3. Hastalık belirti ağında farklı tahmin edicilerin kesinlik değerleri



Şekil 5.4. Hastalık belirti ağında farklı tahmin edicilerin anma değerleri



Şekil 5.5. Hastalık belirti ağında farklı tahmin edicilerin F-ölçütü değerleri

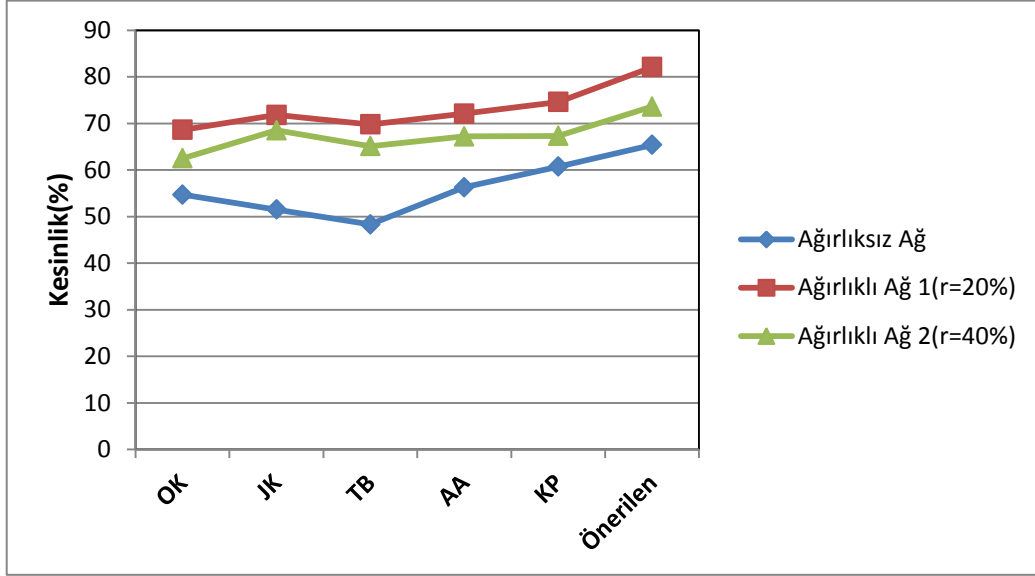
Tablo 5.11’de ise ilk uygulama için bulunan en yüksek skorlu 10 ilişki verilmektedir. [40,50) yaşları arasında bağlantı ağırlıkları eşik değer olan %0,5’den daha küçük olan ve bu şekilde zayıf olduğu kabul edilen bu ilişkilerin tahmin edici yaş aralığı olan [50,60)’da artacağı tahmin edilmiş ve nitekim Tablo 5.8’de gösterildiği gibi bu artışların varlığı da gözlenmiştir. Tablodaki değişim (%) kolonu geçerlilik kümesinden tahmin edici çerçeveye bu artışın yüzdesini vermektedir. Dikkat edilirse burada önerilen yöntemle göre bulunan bu en yüksek skorlu 10 bağlantının ağırlığının 7-11 kat arası arttığı gözlenmiştir. Bu sonuç diğer normal olmayan parametrelerle ortak komşulukları olan iki anormal parametre arasındaki ilişkinin geleceği hakkında fikir vermesi bakımından önemlidir. Bu sayede bir hastanın daha erken yaşlarda şikayetçi olduğu parametrelerin gelişimine bakarak ileride hangi parametrelerinin anormal değerlerde seyredeceği tahmin edilebilecek ve buna göre tedbir alınabilecektir. Tablo 5.11’de verilen Glukoz.Yüksek-Trigliserit.Yüksek bağlantısı denetimsiz bağlantı tahmini yöntemine göre bulunan en yüksek skorlu ilişkisidir. Zaten tabloda da en yüksek değişimin gözlemlendiği bağlantıdır. Bu sonuç tahmin edilen en güçlü ilişkinin artışın en yüksek olduğu bağlantıyı bulduğunu ifade eder.

Tablo 5.11. [40,50) Geçerlilik kümesinde tahmin edilen en yüksek skorlu 10 bağlantı türü

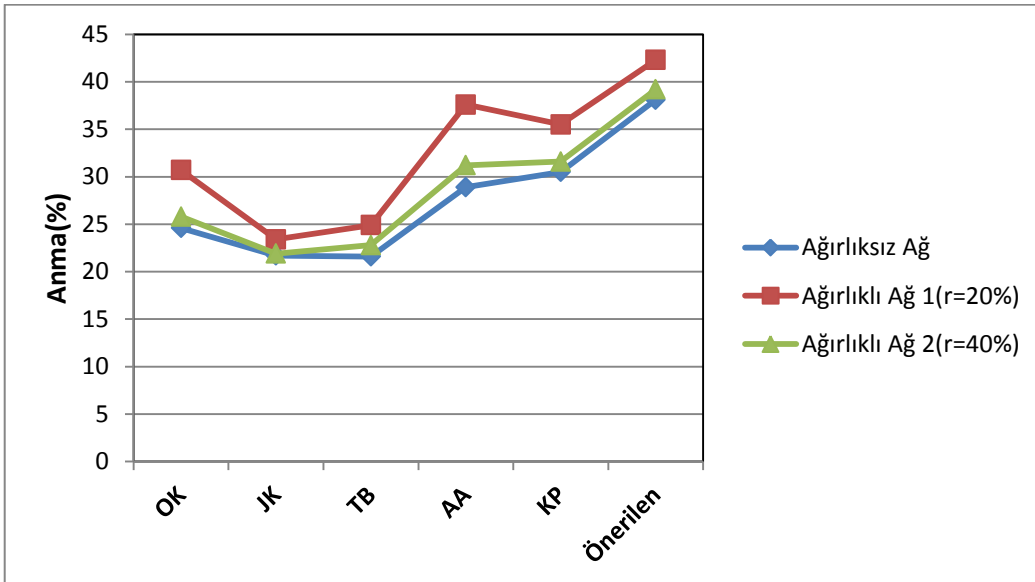
Bağlantı	Geçerlilik kümesindeki ağırlığı	Tahmin edici çerçeveadaki ağırlığı	Değişim (%)
Glukoz.Yüksek- Trigliserit.Yüksek	46	486	1056,5
Protein.Yüksek- Trigliserit.Yüksek	45	383	851,1
B12.Düşük- Folik Asit.Düşük	68	394	579,4
PSA.Yüksek- Üre.Yüksek	84	628	747,6
Kolesterol.Yüksek- Potasyum.Yüksek	36	321	891,6
MCHC.Yüksek- PLT.Düşük	26	274	1053,8
Potasyum.Yüksek- Kalsiyum Düşük	18	175	972,2
BUN.Yüksek-ALP Yüksek	74	694	937,8
Glukoz.Yüksek- Keton.Anormal	34	256	752,9
Ferritin.Yüksek- Gastrin.Yüksek	21	186	885,7

Uygulamaların ikinci kümesi yukarıda verilen aynı süreci daha küçük aralıklı çerçeveler için tekrarlar. Bu yeni uygulamalar kümesinde çerçeve uzunluğu 5 yıl olarak belirlenmiştir. Diğer bir deyişle hastalık belirti ağı oluşturulurken [20, 25), [25, 30), [30, 35), [35, 40), [40, 45), [45, 50), [50, 55) yaş aralığı ağın gelişen durumunu, [55, 60) ise tahmin edici çerçevenin yaş aralığını verir. Şekil 5.6-5.8 bu yeni çerçeveler için yukarıda tanımlanan üç farklı testin sonuçlarını verir. İlk uygulama kümesinden farklı olarak daha küçük aralıklı fakat daha fazla çerçeveli ağlarda, ağın birinci ağırlıklı versiyonu hemen hemen bütün durumlar için daha iyi sonuçlar elde etmiştir. Bu aslında beklenen bir durumdur çünkü ağırlıklı ağ-1’de durum değişim oranı

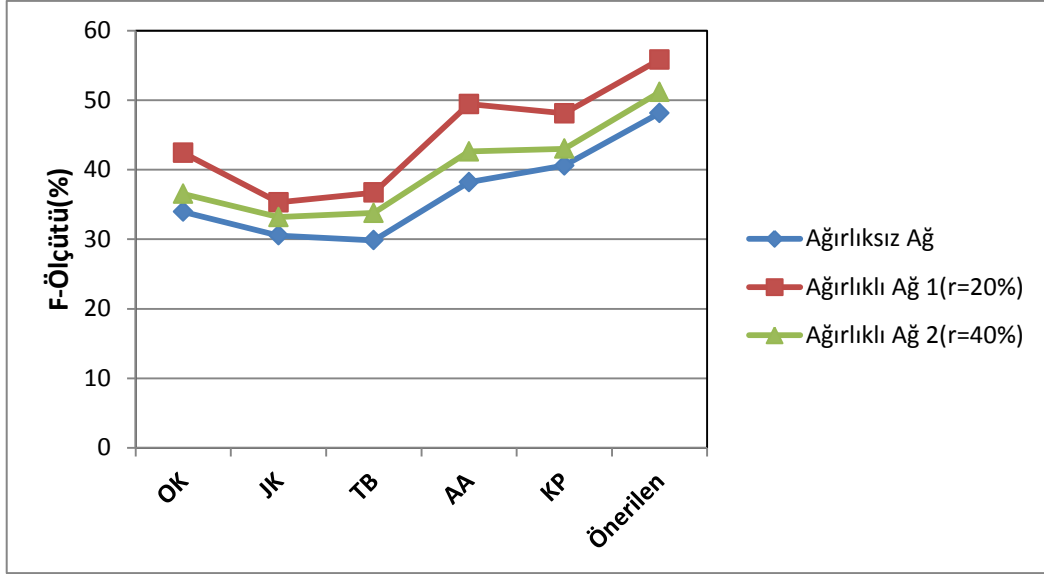
diğerlerinden daha küçüktür bu da çerçeveler arasındaki ağırlıkların değişiminin daha küçük olduğu anlamını taşır.



Şekil 5.6. Daha küçük aralıklı çerçevelerde farklı tahmin edicilerin kesinlik değerleri



Şekil 5.7. Daha küçük aralıklı çerçevelerde farklı tahmin edicilerin anma değerleri



Şekil 5.8. Daha küçük aralıklı çerçevelerde farklı tahmin edicilerin F-ölçütü değerleri

Tablo 5.12 daha küçük aralıklı çerçeveleri içeren uygulama için bulunan en yüksek skorlu 10 ilişkiyi göstermektedir. [50,55) yaşları arasında bağlantı ağırlıkları zayıf olan bu ilişkilerin tahmin edici yaş aralığı olan [55,60)’de arttığı gözlenmiştir. Ancak gerek bağlantıların ağırlığı ve gerekse değişim oranları önceki uygulamadan daha düşüktür. Bunun nedeni her bir çerçevenin boyutunun daha küçük olmasından dolayı hasta sayısının da buna paralel az olmasından kaynaklanmaktadır. Ayrıca bu yeni uygulamada öncekinin bulduğu bazı bağlantılarla birlikte ondan farklı yeni bağlantılarda mevcuttur. Tabloda dikkat çeken diğer bir ayrıntı ise değişim oranının en yüksek olduğu bağlantının yine önerilen yöntemin bulduğu en yüksek skorlu ilişki olduğudur.

Tablo 5.12. [50,55) Geçerlilik kümesinde tahmin edilen en yüksek skorlu 10 bağlantı türü

Bağlantı	Geçerlilik kümesindeki ağırlığı	Tahmin edici çerçeveadaki ağırlığı	Değişim (%)
AST.Yüksek- Kolesterol.Yüksek	27	296	1096,2
Protein.Yüksek- Trigliserit.Yüksek	24	164	683,3
B12.Düşük- Folik Asit.Düşük	37	184	497,2
HGB.Düşük- RBC.Düşük	33	97	293,9
TSH.Yüksek- T3.Düşük	44	103	234,0
Potasyum.Yüksek- Kalsiyum Düşük	21	174	828,5
Albumin.Düşük- Toplam Protein.Düşük	23	151	656,5
Kortizol.Yüksek- PSA.Yüksek	2	78	390,0
MCHC.Yüksek- PLT.Düşük	35	73	208,5
BUN.Yüksek-ALP Yüksek	46	114	247,8

5.3.2. Son Çerçeveleri Daha Önemli Kılarak Elde Edilen Sonuçlar

Tezin bu bölümünde en son çerçevelerin öneminin artırılması durumunda önerilen yöntemin sonuçlarının nasıl değiştiği irdelenmiştir. Bu maksatla çerçeveler arasında monoton bir şekilde artan fonksiyonlar kullanılmıştır. Bu durumda önerilen yöntemin yakınlık değerini hesaplayan fonksiyon aşağıdaki gibi değişikliğe uğramıştır:

$$Score(x, y) = \sum_{k=2}^n I(k)[P(x, y, k) + \alpha S(x, y, k)] \quad (5.15)$$

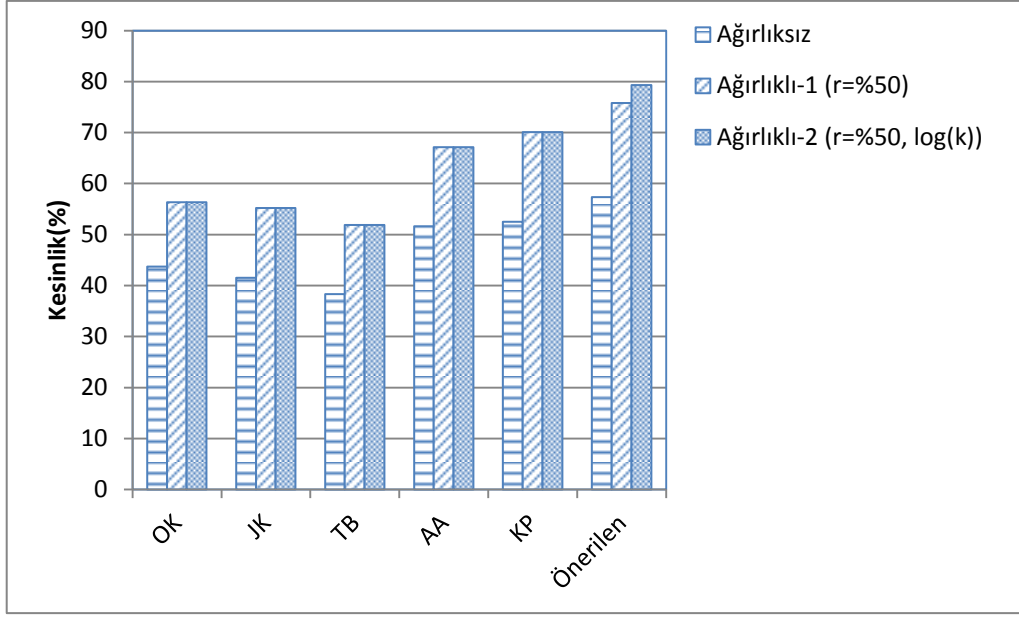
Burada $I(k)$ çerçevenin önem değerini temsil eder. Bu çalışmada $I(k)$ ilk olarak $\log(k)$ olarak belirlenmiştir. Yapılan uygulama için ağırlıklı versiyonları ise aşağıdaki gibi tanımlanmıştır:

(1) Ağırlıklı versiyon-1: Burada her bir link anormal parametrelerin her ikisini gösteren hastaların sayısı ile ağırlıklandırılır. Durum değişim oranı (r) %50 ve $I(k)=1$ 'dir.

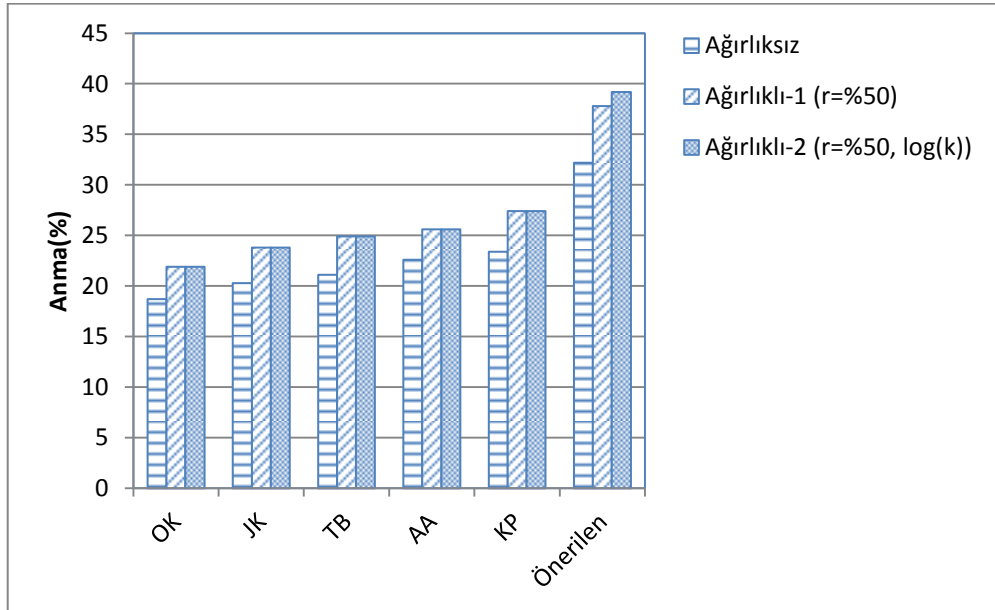
(2) Ağırlıklı versiyon-2: Durum değişim oranı (r) %50 ve $I(k)=\log(k)$ 'dir.

Hastalık belirti ağı oluşturulurken $[20, 30)$, $[30, 40)$, $[40, 50)$ yaş aralığındaki çerçeveler ağırlıklı gelişen yapısı olarak belirlenmişken, $[50, 60)$ yaş aralığı ise tahmin edici çerçeve olarak kullanılmıştır. Eğer geçerlilik kümesinde bağlantılı olmayan bir düğüm çifti tahmin edici çerçevede bağlantılı bir çiftte dönerse veya önceden tanımlanan eşik değerinden daha küçük ağırlığa sahip bir çift geçerlilik kümesinden tahmin edici çerçeveye durum değişim oranından daha fazla yükselirse düğüm çifti pozitif sınır olarak kabul edilir. Aksi halde sınıf negatiftir. Daha önceki uygulamada da olduğu gibi en iyi performans $c = 0,4$, $s = 0,9$, $d = -0,3$ ve $\alpha = 0,07$ değerlerinde elde edilmiştir.

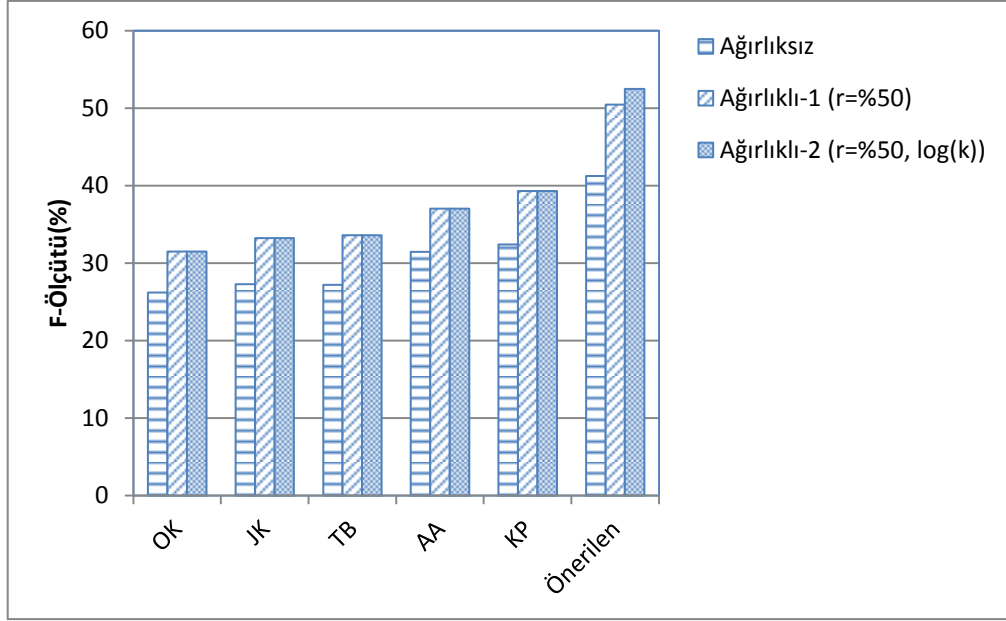
İlk uygulama ağırlıklı ağlar için geliştirilen geleneksel yakınlık fonksiyonlarıyla önerilen durum tabanlı fonksiyonu karşılaştırmak için yapılmıştır. Şekil 5.9 hastalık belirti ağının üç farklı versiyonun kesinlik değerlerini gösterir. Şekilden de görüldüğü gibi en iyi tahmin performansı ağırlıklı ikinci versiyonu üzerinde önerilen tahmin edici tarafından başarılmıştır. Bu uygulamada ağırlıksız versiyonu yine kötü sonuçlar elde etmiştir. Benzer şekilde ikinci uygulama ağı her üç versiyonunun anma değerleriyle ilgilidir. Şekil 5.10'da görüleceği gibi geleneksel yakınlık ölçüleri için ağırlıklı olmanın avantajı nispeten azalmıştır. Yani, bilinen yakınlık ölçülerinde ağırlıksız ile ağırlıklı ağların anma değeri birbirine yaklaşmıştır. Son uygulama ağların F-ölçüt değerlerini bulmak için yapılmıştır. Şekil 5.11'de verilen durum tabanlı yöntemin F-ölçüt değerleri göz önüne alınan bütün yöntemler içinde en iyi olanıdır. Bu uygulamalar ayrıca $I(k)=\log(k)$ durumundaki değerlerin $I(k)=1$ 'e göre daha üstün olduğunu göstermiştir. Bu durum önemli bir gerçeği ortaya çıkarır ki: Son durumlar bağlantıların oluşmasında daha fazla bilgi taşırlar.



Şekil 5.9. Önem çerçevesi hastalık belirti ağında altı farklı tahmin edicinin kesinlik değeri.



Şekil 5.10. Önem çerçevesi hastalık belirti ağında altı farklı tahmin edicinin anma değeri.



Şekil 5.11. Önem çerçeveli hastalık belirti ağında altı farklı tahmin edicinin F-ölçütü değeri.

Tablo 5.13. Farklı önem fonksiyonlarına göre altı farklı yöntemin kesinlik değerleri (%)

Önem Fonk./ Yöntem	OK	JK	TB	AA	KP	Önerilen
$I(k)=1$	62,5	68,5	65,1	67,2	67,3	73,6
$I(k)=k$						75,8
$I(k)=\log(k)$						77,2

Bu alanda yapılan diğer bir uygulamada ise üç farklı önem fonksiyonuyla $r = \%40$ durum değişim oranına sahip ağırlıklı bir ağın göstereceği değişiklikler izlenmiştir. Burada $[20, 25)$, $[25, 30)$, $[30, 35)$, $[35, 40)$, $[40, 45)$, $[45, 50)$, $[50, 55)$ yaş aralıkları ağın gelişen yapısı olarak belirlenmiş, $[55, 60)$ yaş aralığı ise tahmin edici çerçeve olarak kullanılmıştır. Yani bir önceki uygulamalara göre çerçevelerin boyutu 10'dan 5'e inerken durum değişim oranında ise %10'luk bir azalma olmuştur. Böyle ağırlıklı bir ağda altı farklı yöntemin üç önem fonksiyonuna göre kesinlik değerleri Tablo 5.13'de verilmiştir. Geleneksel yöntemler bağlantı tahminlerini yaparken önem fonksiyonundan etkilenmezler. Buna

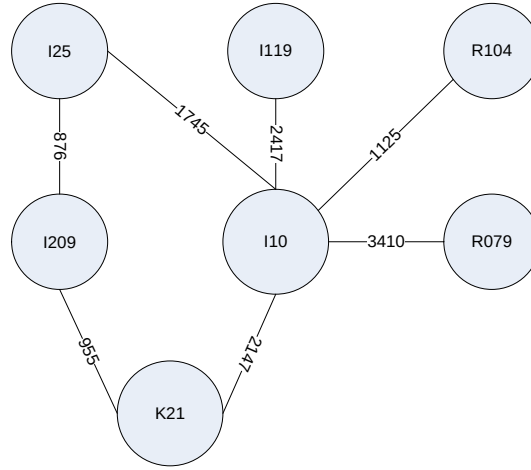
karşın durum tabanlı yöntemde $I(k) = k$, $I(k) = 1$ 'e göre daha iyi sonuç elde ederken, son durumlara daha fazla önem veren $I(k) = \log(k)$ en iyi sonucu bulmuştur.

6. GELİŞEN DURUMLU HASTALIK AĞINDA YAŞ SERİLERİ TABANLI BAĞLANTI TAHMİNİ

Önceki bölümde hastalık parametrelerinin birbirini nasıl etkilediği incelendikten sonra bu bölümde doğrudan hastalıkların birbirlerini nasıl tetiklediği araştırılacaktır. Bu maksatla öncelikle ağırlıklandırılmış bir hastalık ağı oluşturulacaktır. Bu ağ yaş serilerine göre gruplandırılmıştır. Böylece genç yaşlardan itibaren hastalıkların birbirini etkileme süreci incelenecek ve belirli bir yaştan üstündeki kişilerin hangi hastalıklara yakalanma risklerinin olduğu önceden tahmin edilmeye çalışılacaktır.

6.1. Hastalık Ağının Oluşturulması

Hastalık ağında düğümler hastalıklardır. Kenarlar ise aynı hastada görülen hastalık çiftlerini birbirine bağlar. Önceki bölümde verilen hastalık belirti ağından farklı olarak iki düğüm arasındaki ilişki, aynı hastada farklı zamanlarda görülmüş olsa bile meydana gelir. Dolayısıyla kenarlardaki ağırlıklar farklı zamanlarda yakalanmış olsa bile her iki hastalığa yakalanan hasta sayısını gösterir. Ağ, yaş dilimli bölümlere ayrılmıştır. Bu bölümler farklı yaş aralıklarındaki ağın durumlarını temsil eder. Bundan sonra ardışık bölümler gruplandırılarak yaş serilerini oluşturur. Bu seriler aynı zamanda çerçeve olarak ta isimlendirilir. Düğümler ICD-10 ile kodlanmıştır. ICD Kodları (*Hastalıkların Uluslararası Sınıflandırılması*) hastalıkların ve sağlık sorunlarının uluslararası platformda standart bir biçimde temsil edilebilmesini sağlayan istatistik kodlar bütünüdür. ICD-9 olarak bilinen önceki sürüm güncellenerek ICD-10 kodları oluşturulmuş ve günümüzde kullanılması sağlanmıştır. 1990 yılı Mayıs ayında 43. Dünya Sağlık Kongresi'nde onaylanan ICD-10 kodları 1994 yılında Dünya Sağlık Örgütü (WHO) üye ülkelerinde kullanılmaya başlamıştır. Kodlama sisteminin başlangıcı 1850'li yıllara kadar uzanmaktadır. Hazırlanan ilk liste Uluslararası Ölüm Sebepleri Listesi olarak 1893 yılında duyurulmuştur. Dünya Sağlık Örgütü, 1948 yılında bu listenin devam ettirilmesi sorumluluğunu üzerine almış ve 6. liste güncellemesiyle birlikte ICD Kodlarının temelleri atılmıştır. ICD-10 kodları Türkiye Cumhuriyeti Sağlık Bakanlığı tarafından Türkçeleştirilerek kullanılmaya başlanmıştır. Şekil 6.1'de ICD-10 ile kodlanmış hastalık ağının bir parçası gösterilmiştir.



Şekil 6.1. ICD-10 kodlu hastalık ağının bir parçası

6.2. Yaş Serileri Tabanlı Denetimsiz Bağlantı Tahmini Algoritması

Bu bölümde ağdaki düğüm çiftleriyle ilgili gelişen durumları ve ağırlıklı ağın gelişen yapısını göz önüne alan denetimsiz öğrenme tabanlı bağlantı tahmin yöntemi anlatılmıştır. Ağın gelişen yapısı, geçerlilik kümesini de içeren birkaç yaş dilimli çerçeveden oluşur. Örneğin, [20, 25) yaş aralığındaki bir çerçeve, 20 ile 25 yaş aralığındaki bütün insanların hastalıkları arasında gözlenen bağlantıları içeren bir alt-çizgedir. Hastalık belirti ağındaki çerçeveye benzer şekilde, her bir çerçevenin boyutu tahmin edici çerçevenin uzunluğuyla aynıdır ve bu uzunluk daha önceden belirlenmiştir.

Yaş serileri arasındaki durumların geçişleri ve hastalık ağının yaş serileri skoru daha önceki bölümde verilen durum tabanlı yakınlık fonksiyonlarıyla aynıdır. Sonuç olarak hastalık ağı için geliştirilen denetimsiz bağlantı tahmini algoritması aşağıdaki gibi önerilmiştir:

1. Geçerlilik kümesinden bağlantılı olmayan çiftler veya önceden belirlenen eşik değerinden daha küçük ağırlık derecesine sahip çiftler seçilir. Buna karşın ağın gelişimi boyunca herhangi bir durumla doğrudan veya dolaylı olarak etkilenmemiş bağlantılı olmayan çiftler yine bir bağlantı oluşturmaz düşüncesiyle elenir.
2. Seçilen çiftlerin yakınlık skorları altı farklı tahmin ediciye göre hesaplanır.

3. Yakınlık değeri en yüksek ilk 1000 çift test için seçilir. Bu yüksekten azalan şekilde sıralanmış çiftler muhtemelen oluşacağı düşünülen gelecek bağlantılar olarak kabul edilir.
4. Eğer geçerlilik kümesinde bağlantılı olmayan bir çift, tahmin edici çerçevede bağlantılı bir çifte dönüşürse veya önceden belirlenen eşik değerinden daha küçük bir ağırlığa sahip bir çiftin ağırlığı geçerlilik kümesinden tahmin edici çerçeveye durum değişim oranından (r) daha fazla bir oranda artarsa; düğüm çifti pozitif sınıf olarak kabul edilir. Aksi durumda sınıf negatiftir.
5. Göz önüne alınan bütün tahmin edicilerin performans ölçüleri hesaplanır.

Tutarlı, güçlendiren ve zayıflatan durumlar için verilen c , s ve d değerleri geçerlilik kümesi bağlantı tahmini performansı değerlendirilerek hesaplanır.

6.2.1. Uygulama Sonuçları

Bu bölümde, önerilen yöntemi değerlendirmek için gerçek tıp verileri kullanarak bazı uygulamalar gerçekleştirilmiştir. 210134 hasta kaydını içeren veri kümesi Fırat Üniversitesi Hastanesi'nden elde edilmiştir. Hastaların medikal kayıtları hastanenin etik kurallarına uygun olarak alınmıştır. Bundan sonraki kısımda kullanılan hastalık ağı ve denetimsiz öğrenme stratejili süreç ele alınmıştır.

6.2.1.1. Hastalık Ağı

Daha önce de ifade edildiği gibi hastalık ağında her bir düğüm bir hastalık iken, düğümler arasındaki bağlantılar hastalığın aynı hastada farklı zamanlarda bile olsa var olduğunu ifade eder. Hastalık ağı oluşturulurken, kayıtların alındığı dönem içerisinde sadece bir defa hastaneye giden hastalar silinmiştir. Ayrıca bir defadan fazla hastaneye giden fakat aynı teşhis konulan hastalar da ele alınmamıştır. Bu temizleme işlemlerinden sonra bir defadan daha fazla hastaneyi ziyaret eden hasta sayısı 168006 olarak belirlenmiştir. Bu durumda düğüm sayısı 3114'dür. Bu sayı toplam hastalık sayısını ifade eder. Oluşturulan ağın versiyonları olarak ta aşağıdaki 3 yapı kullanılmıştır:

- (4) Ağın ağırlıklandırılmamış versiyonu: Burada en az bir hastada görülen hastalık çiftleri arasındaki bağlantılar sadece sunulur.
- (5) Ağın ağırlıklı versiyonu-1: Burada her bir bağlantı hastalığın her ikisini gösteren hastaların sayısı ile ağırlıklandırılır. Durum değişim oranı (r) %25'dir.
- (6) Ağın ağırlıklı versiyonu-2: Önceki ile aynıdır tek fark durum değişim oranı (r) %50'dir.

Ağın farklı versiyonlarıyla uygulama yapmanın amacı, ilişkideki ağırlıkları kullanmanın ve farklı durum değişim oranlarının daha iyi tahmin sonuçları üretip üretmeyeceğini görmektir. Ağın yaş serilerine gelince, bu bölümde karakteristikleri farklı dört ayrı yaş serileri kullanılmıştır. Bu özellikler önerilen yöntemin kesinlik, anma ve F-ölçütü değerlerini test etmek için spesifik bir şekilde belirlenmiştir. Tablo 6.1'de görüldüğü gibi, ikinci uygulama kümesinde kullanılacak yaş serisinin çerçeve boyutu birinci yaş serisinden daha büyüktür. Benzer şekilde, üçüncü uygulamada kullanılacak yaş serileri daha genç hastalarla, dördüncününki ise daha yaşlı hastalarla ilgilenir. Bağlantı tahmini için c , s , d , ve α , parametrelerin en uygun değerlerini deneysel olarak değerlendirmek ve belirlemek için geçerlilik kümesi kullanılmıştır. Bazı testlerle en uygun başarımlar $c = 0,45$, $s = 0,85$, $d = -0,3$ ve $\alpha = 0,07$ değerlerinde bulunmuştur. Önceki uygulamalarda olduğu gibi eşik değeri yine 50 olarak belirlenmiştir. Bu sonuçlar tutarlı (c) ve zayıflayan (d) değerlerin neredeyse birbirini dengelediğini gösterir. En iyi performans değerleri $c < s$ durumunda elde edilmiştir.

Tablo 6.1. Uygulamalarda kullanılan yaş serileri

Uygulama Kümesi	Yaş Serileri		Açıklama
	Ağın Gelişen Yapısı	Tahmin Edici Çerçeve	
1	[20, 25), [25, 30), [30, 35), [35, 40), [40, 45), [45, 50), [50, 55)	[55,60)	Küçük boyutlu çerçeveler
2	[20, 30), [30,40)[40,50)	[50,60)	Geniş boyutlu çerçeveler
3	[20, 25), [25, 30)[30, 35)	[35, 40)	Genç hastalar
4	[40, 45), [45, 50)[50, 55)	[55, 60)	Yaşlı hastalar

6.2.1.2. Değerlendirme Sonuçları

Hastalık ağlarıyla ilgili Tablo 6.1’de verilen her bir yaş serisi için üç farklı uygulama yapılmıştır. Bu uygulamalarda amaç önerilen yöntem ve karşılaştırma için verilen 5 iyi bilinen tahmin edicinin kesinlik, anma ve F-ölçüt değerlerini incelemektir. Uygulamaların ilk kümesinde [20, 25), [25, 30), [30, 35), [35, 40), [40, 45), [45, 50), [50, 55)’den oluşan gelişen yapıllı ağ değerlendirilmiştir. Bu ağda çerçeve boyutu 5’dir ve [55,60) yaş aralığındaki çerçeve tahmin etme amaçlı kullanılmıştır. Tablo 6.2 bu uygulamanın geçerlilik kümesinde en sık görülen 20 hastalığı vermektedir.

Tablo 6.2. Birinci uygulamanın geçerlilik kümesinde en sık görülen 20 hastalık

Hastalık	Sıklığı
R10.4 Karın Ağrısı	13143
K21 Gastro-Özofajial Reflü Hastalığı	11252
I10 Esansiyel (Primer) Hipertansiyon	7010
I25 Kronik İskemik Kalp Hastalığı	6446
R07.4 Göğüs Ağrısı, Tanımlanmamış	6359
I20.9 Angina Pectoris, Tanımlanmamış	6118
N39.0 Üriner Sistem Enfeksiyonu, Yeri Tanımlanmamış	5944
M25.5 Eklem Ağrısı	5866
M79.1 Miyalji	5399
E78.4 Hiperlipidemi, Diğer	4708
R50.9 Ateş, Tanımlanmamış	4255
E11.9 İnsülin Bağımlı Olmayan Diyabetes Mellitus, Komplikasyonları Olmayan	3991
R06.0 Dispne	3463
I50 Kalp Yetmezliği	3215
R11 Bulantı ve Kusma	3156
K59.0 Kabızlık	2992
D64 Anemi, Diğer	2945
E55 Vitamin D Eksikliği	2840
N30 Sistit	2638
D50 Demir Eksikliği Anemisi	2629

Ağın üç farklı versiyonu için altı yakınlık skoru tarafından bulunan kesinlik değerleri Tablo 6.3’de rapor edilmiştir. Tablodan da görülebileceği gibi, ağın ağırlıklı versiyonları bütün yöntemler için daha iyi sonuçlar elde etmiştir. Önerilen yaklaşım her üç versiyon için diğer iyi bilinen tahmin edicilere üstünlük sağlamıştır. Dikkat çekilmesi gereken diğer bir husus ise en iyi tahmin edicinin ağın ağırlıklı birinci versiyonu üzerinde, önerilen yöntem ile başarılı olmuş olmasıdır. Ağırlıklı ağın birinci versiyonunda durum değişim oranı nispeten küçüktür ($r=\%25$). Çerçeveler arasındaki bağlantıların durum değişiklikleri (Tutarlı, güçlendiren veya zayıflatan) bu oranla belirlendiğinden, böyle bir oranda 5 gibi düşük boyutlu çerçeveler için daha fazla sayıda durum değişikliği gözlemek mantıklıdır. Diğer bir deyişle, çerçevenin boyutu sabit kalırken durum değişim oranı yükselirse çerçeveler arasında daha az sayıda durum değişikliği gözlenir. Bu durum tahmin edicinin uygunluğunu doğrudan etkiler. İlk kümenin ikinci uygulamasında ağın üç versiyonu için diğer yöntemlerle önerilen yöntemin anma değerleri karşılaştırılmıştır. Tablo 6.4’den geliştirilen yöntemin en iyi sonuçları elde ettiği görülür. Kümenin son uygulaması ise tüm yöntemlerin F-ölçütü değerlerini karşılaştırmasıdır. Sonuçlar Tablo 6.5’de verilmiştir. Bu uygulamanın bir sonucu olarak, önerilen yöntemin diğer yaklaşımlara üstünlük sağladığı belirlenmiştir. Tablo 6.6 ise uygulamanın birinci kümesinde önerilen yöntem ile bulunan en yüksek 10 skorlu bağlantıyı göstermektedir. Bu tabloya göre önerilen yöntem için en yüksek skorlu ilişki “*İnsülin bağımlı olmayan diyabetes mellitus, komplikasyonları olmayan*” ve “*Esansiyel (Primer) Hipertansiyon*” arasında olanıdır. Bu aynı zamanda geçerlilik kümesinden tahmin edici çerçeveye en yüksek değişimin gözlemlendiği ilişkidir.

Tablo 6.3. Küçük boyutlu çerçevelerde altı farklı yöntemin kesinlik değerleri (%)

Yöntem	Ağırlıksız	Ağırlıklı-1 ($r=\%25$)	Ağırlıklı-2 ($r=\%50$)
OK	43,5	44,6	43,7
JK	41,1	42,5	41,8
TB	44,1	44,7	44,7
AA	45,8	47,1	46,8
KP	47,2	50,1	49,9
Önerilen	52,0	58,2	54,8

Tablo 6.4. Küçük boyutlu çerçevelerde altı farklı yöntemin anma değerleri (%)

Yöntem	Ağırlıksız	Ağırlıklı-1 (r=%25)	Ağırlıklı-2 (r=%50)
OK	77,5	75,7	74,5
JK	78,2	80,4	77,8
TB	77,1	78,8	77,7
AA	74,8	74,6	73,4
KP	73,6	75,1	73,9
Önerilen	74,5	82,2	76,8

Tablo 6.5. Küçük boyutlu çerçevelerde altı farklı yöntemin F-ölçütü değerleri (%)

Yöntem	Ağırlıksız	Ağırlıklı-1 (r=%25)	Ağırlıklı-2 (r=%50)
OK	55,7	56,1	55,0
JK	53,8	55,6	54,3
TB	56,1	57,0	56,7
AA	56,8	57,7	57,1
KP	57,5	60,1	59,5
Önerilen	61,2	68,1	63,9

Uygulamaların ikinci kümesinde önerilen durum tabanlı yakınlık skoru daha geniş boyutlu çerçevelere sahip yaş serilerine uygulanmıştır. Bu yeni yaş serilerinde çerçeve boyutu 10'dur. Ağın gelişen yapısı [20,30), [30,40)[40,50) çerçevelerden ibarettir. [40,50) yaş aralığındaki çerçeve geçerlilik kümesi iken, [50,60) tahmin edici çerçevedir. İlk uygulama altı farklı yöntemin kesinlik değerini bulmayla ilgilidir. Tablo 6.7 bu uygulamanın sonuçlarını verir. Durum-tabanlı skor diğer yöntemlere üstünlük sağlar ve ağın ağırlıklı ikinci versiyonu üzerinde biri hariç en yüksek kesinliği çıkarır. Önceki yaş serilerinden farklı olarak daha büyük durum değişim oranının daha geniş boyutlu çerçevelerde daha etkili tahminler yakaladığı görülmüştür. Uygulamalar sayesinde gözlenen diğer önemli bir nokta, düşük değişim oranlarında durum değişiklikleri çerçeveler boyunca sıklıkla değişir. Buna karşın, bu yaş serilerinde her bir çerçevedeki hasta sayısı birinci yaş serisindekine göre daha fazladır. Aynı durum anma ve F-ölçütü ile ilgili yapılan diğer uygulamalarda da geçerlidir. Her iki uygulama için, önerilen yöntem Tablo 6.8 ve 6.9'da da gösterildiği gibi diğer yöntemlerle karşılaştırıldığında en iyi sonuçları verir. Tablo 6.10 ise uygulamanın

ikinci kümesinde önerilen yöntem ile bulunan en yüksek 10 skorlu bağlantıyı göstermektedir.

Tablo 6.6. Uygulamanın birinci kümesinde önerilen yöntem tarafından bulunan en yüksek 10 ilişki

Bağlantı	Geçerlilik kümesindeki ağırlığı	Tahmin edici çerçeveedeki ağırlığı	Değişim (%)
İnsülin bağımlı olmayan diyabetes mellitus, komplikasyonları olmayan- Esansiyel (Primer) Hipertansiyon	32	166	518,7
Angina Pectoris, Tanımlanmamış- Kronik İskemik Kalp Hastalığı	28	134	478,5
Gastro-Özofajial Reflü Hastalığı- Karın Ağrısı	44	147	334,0
Esansiyel (Primer) Hipertansiyon- Kronik Böbrek Yetmezliği, Tanımlanmamış	42	152	361,9
Esansiyel (Primer) Hipertansiyon- Gastro-Özofajial Reflü Hastalığı	30	118	393,3
Kronik İskemik Kalp Hastalığı- Göğüs Ağrısı, Tanımlanmamış	19	97	510,5
Üriner Sistem Enfeksiyonu, Yeri Tanımlanmamış- Diyabetes mellitus, diğer tanımlanmış	22	86	390,9
Hiperlipidemi, Diğer- Kronik İskemik Kalp Hastalığı	27	112	414,8
Esansiyel (Primer) Hipertansiyon- Yaygın Anksiyete Bozukluğu	38	154	405,2
Esansiyel (Primer) Hipertansiyon- Karın Ağrısı	34	103	302,9

Tablo 6.7. Geniş boyutlu çerçevelerde altı farklı yöntemin kesinlik değerleri (%)

Yöntem	Ağırlıksız	Ağırlıklı-1 (r=%25)	Ağırlıklı-2 (r=%50)
OK	40,4	42,7	44,7
JK	39,5	40,5	41,8
TB	43,8	43,5	42,4
AA	44,9	45,4	49,3
KP	44,2	46,8	48,5
Önerilen	50,1	54,3	56,0

Tablo 6.8. Geniş boyutlu çerçevelerde altı farklı yöntemin anma değerleri (%)

Yöntem	Ağırlıksız	Ağırlıklı-1 (r=%25)	Ağırlıklı-2 (r=%50)
OK	75,8	77,1	76,3
JK	76,4	76,5	76,8
TB	73,7	76,3	78,7
AA	72,1	73,9	72,8
KP	72,8	72,5	74,5
Önerilen	72,6	76,2	80,3

Tablo 6.9. Geniş boyutlu çerçevelerde altı farklı yöntemin F-ölçütü değerleri (%)

Yöntem	Ağırlıksız	Ağırlıklı-1 (r=%25)	Ağırlıklı-2 (r=%50)
OK	52,7	54,9	56,3
JK	52,0	52,9	54,1
TB	54,9	55,4	55,1
AA	55,3	56,2	58,7
KP	55,0	56,8	58,7
Önerilen	59,2	63,4	65,9

Tablo 6.10. Uygulamanın ikinci kümesinde önerilen yöntem tarafından bulunan en yüksek 10 ilişki

Bağlantı	Geçerlilik kümesindeki ağırlığı	Tahmin edici çerçevedeki ağırlığı	Değişim (%)
Kronik Böbrek Yetmezliği, Tanımlanmamış- İnsülin bağımlı olmayan diyabetes mellitüs, komplikasyonları olmayan	83	302	363,8
İnsülin bağımlı olmayan diyabetes mellitüs, komplikasyonları olmayan- Esansiyel (Primer) Hipertansiyon	78	326	417,9
Angina Pectoris, Tanımlanmamış- Kronik İskemik Kalp Hastalığı	91	286	314,2
Gastro-Özofajial Reflü Hastalığı- Kronik İskemik Kalp Hastalığı	62	221	356,4
Esansiyel (Primer) Hipertansiyon- Gastro-Özofajial Reflü Hastalığı	29	104	358,6
Esansiyel (Primer) Hipertansiyon Kronik Böbrek Yetmezliği, Tanımlanmamış	47	128	272,3
Vitamin B12 eksikliği anemisi, tanımlanmamış – Baş dönmesi (Vertigo)	70	203	290,0
Üriner Sistem Enfeksiyonu, Yeri Tanımlanmamış- Diyabetes mellitüs, diğer tanımlanmış	34	89	261,7
Esansiyel (Primer) Hipertansiyon- Yaygın Anksiyete Bozukluğu	59	108	183,0
Gastro-Özofajial Reflü Hastalığı - Akut bronşit, tanımlanmamış	24	75	312,5

Tablo 6.11. Genç hastalarda altı farklı yöntemin kesinlik değerleri (%)

Yöntem	Ağırlıksız	Ağırlıklı-1 (r=%25)	Ağırlıklı-2 (r=%50)
OK	35,4	37,3	35,9
JK	36,3	36,8	36,5
TB	37,5	38,4	37,7
AA	39,7	40,7	40,6
KP	40,1	43,2	42,9
Önerilen	44,7	50,9	48,7

Tablo 6.12. Genç hastalarda altı farklı yöntemin anma değerleri (%)

Yöntem	Ağırlıksız	Ağırlıklı-1 (r=%25)	Ağırlıklı-2 (r=%50)
OK	70,9	69,5	68,4
JK	71,8	71,4	70,9
TB	72,3	72,4	71,5
AA	70,6	69,8	68,7
KP	69,9	69,3	67,8
Önerilen	74,5	75,7	73,6

Tablo 6.13. Genç hastalarda altı farklı yöntemin F-ölçütü değerleri (%)

Yöntem	Ağırlıksız	Ağırlıklı-1 (r=%25)	Ağırlıklı-2 (r=%50)
OK	47,2	48,5	47,0
JK	48,2	48,5	48,1
TB	49,3	50,1	49,3
AA	50,8	51,4	51,0
KP	50,9	53,2	52,5
Önerilen	55,8	60,8	58,6

Uygulamaların bundan sonraki iki kümesinde ilk uygulamada verilen yaş serileri iki parçaya bölünmüştür. Bu uygulamaların amacı daha genç ve daha yaşlı hastalar için oluşturulan ağlar üzerinde bağlantı tahmini yönteminin performans ölçülerini değerlendirmektir. Bu maksatla, uygulamaların üçüncü kümesi parçalanmış yaş serisinin ilk parçasını ele alır. Bu parça aynı zamanda daha genç hastaların yaş serileridir. Göz

önüne alınan ağın gelişen yapısı [20, 25), [25, 30)[30, 35) yaş serilerini içerirken, tahmin edici çerçeve [35, 40) yaş aralığıdır. İlk uygulama daha genç hastalar için önerilen yöntem ve diğer yaklaşımların kesinlik değerlerini araştırır. Tablo 6.11’de verilen sonuçlar geliştirilen yöntemin ağın her üç versiyonu için daha iyi sonuçlar elde ettiğini ispatlar.

Tablo 6.14. Uygulamanın üçüncü kümesinde önerilen yöntem tarafından bulunan en yüksek 10 ilişki

Bağlantı	Geçerlilik kümesindeki ağırlığı	Tahmin edici çerçevedeki ağırlığı	Değişim (%)
Karın ağrısı diğer ve tanımlanmamış- Diabetes mellitus, diğer tanımlanmamış	31	68	219,3
Kabızlık- Dispepsi	24	54	225,0
Hipotroidizm tanımlanmamış, diğer- Vitamin B12 eksikliği anemisi	18	37	205,5
Nöralji ve nörit, tanımlanmamış- Gastro-Özofajial Reflü Hastalığı	26	50	192,3
İnsülin bağımlı olmayan diyabetes mellitus, komplikasyonları olmayan- Nöralji ve nörit, tanımlanmamış	45	88	195,5
Kırgınlık ve yorgunluk- Vitamin D eksikliği	19	46	242,1
Esansiyel (Primer) Hipertansiyon- Yaygın Anksiyete Bozukluğu	27	60	222,2
Çarpıntı- Bulantı ve kusma	34	69	202,9
Gastro-Özofajial Reflü Hastalığı - Akut bronşit, tanımlanmamış	16	28	175,0
Kronik İskemik Kalp Hastalığı- Angina Pectoris, Tanımlanmamış	21	46	219,0

Dahası, önerilen yöntemin ağın birinci ağırlıklı versiyonu üzerindeki sonuçları kesinlik değerleriyle ilgili bütün uygulamalar içinde önerilen yöntem için en kötüsüdür. Bunun iki

nedeni vardır. Birincisi, genç hasta sayısı yaşlı hastalarinkinden azdır. İkincisi çerçeve boyutu nispeten küçüktür. Tablo 6.12 ve 6.13 bu yeni yaş serileri için yapılan tahmin edicilerin anma ve F-ölçütü değerlerini gösterir. Sonuçlar genel olarak önceki iki uygulamadan daha kötüdür. Tablo 6.14 uygulamanın üçüncü kümesinde önerilen yöntem ile bulunan en yüksek 10 skorlu bağlantıyı göstermektedir.

Tablo 6.15. Yaşlı hastalarda altı farklı yöntemin kesinlik değerleri (%)

Yöntem	Ağırlıksız	Ağırlıklı-1 (r=%25)	Ağırlıklı-2 (r=%50)
OK	36,8	38,4	40,9
JK	37,7	38,2	39,1
TB	40,2	40,8	41,1
AA	41,3	42,2	45,0
KP	40,8	43,7	45,8
Önerilen	46,6	50,9	53,7

Tablo 6.16. Yaşlı hastalarda altı farklı yöntemin anma değerleri (%)

Yöntem	Ağırlıksız	Ağırlıklı-1 (r=%25)	Ağırlıklı-2 (r=%50)
OK	73,1	74,8	75,1
JK	72,8	74,3	74,6
TB	70,8	74,6	76,2
AA	70,7	73,3	70,4
KP	71,6	70,0	72,7
Önerilen	71,5	73,9	77,8

Tablo 6.17. Yaşlı hastalarda altı farklı yöntemin F-Ölçütü değerleri (%)

Yöntem	Ağırlıksız	Ağırlıklı-1 (r=%25)	Ağırlıklı-2 (r=%50)
OK	48,9	50,7	52,9
JK	49,6	50,4	51,3
TB	51,2	52,7	53,3
AA	52,1	53,5	54,9
KP	51,9	53,8	56,1
Önerilen	56,4	60,2	63,5

Tablo 6.18. Uygulamanın dördüncü kümesinde önerilen yöntem tarafından bulunan en yüksek 10 ilişki

Bağlantı	Geçerlilik kümesindeki ağırlığı	Tahmin edici çerçevedeki ağırlığı	Değişim (%)
İnsülin bağımlı olmayan diyabetes mellitus, komplikasyonları olmayan- Esansiyel (Primer) Hipertansiyon	32	91	284,3
Kronik Böbrek Yetmezliği, Tanımlanmamış- İnsülin bağımlı olmayan diyabetes mellitus, komplikasyonları olmayan	43	107	248,8
Angina Pectoris, Tanımlanmamış- Kronik İskemik Kalp Hastalığı	28	84	300,0
Postmenapozal osteoporoz- Vitamin D eksikliği	19	55	289,4
Esansiyel (Primer) Hipertansiyon- Kronik Böbrek Yetmezliği, Tanımlanmamış	42	114	271,4
Benign prostat hiperplazisi – Esansiyel (Primer) Hipertansiyon	37	106	286,4
Hiperlipidemi, Diğer- Kronik İskemik Kalp Hastalığı	27	59	218,5
Üriner Sistem Enfeksiyonu, Yeri Tanımlanmamış- Diyabetes mellitus, diğer tanımlanmış	22	64	290,9
Vitamin B12 eksikliği anemisi, tanımlanmamış – Baş dönmesi (Vertigo)	44	81	184,0
Gut, tanımlanmamış – Eklem ağrısı	17	43	252,9

Uygulamaların dördüncü kümesinde, yaşlı hastaları içeren bölünmüş yaş serilerinin ikinci parçası ele alınmıştır. Bu yaş serilerinde daha fazla sayıda hasta görülür. Bu aslında beklenen bir durumdur çünkü yaşlı hastaların bir hastalığa yakalanma olasılığı genç hastalarinkinden daha fazladır. İlgili uygulama sonuçları Tablo 6.15-6.17’de verilmiştir. Her ne kadar bu yaş serilerinde çerçeve boyutu küçük olsa da ağın ikini versiyonu yaşlı hastalar nedeniyle her bir çerçevedeki hasta sayılarının fazla olmasından dolayı nispeten daha iyi sonuçlar elde edilmiştir. Tablo 6.18 uygulamanın dördüncü kümesinde önerilen yöntem ile bulunan en yüksek 10 skorlu bağlantıyı göstermektedir.

Son uygulama ağırlıklı ağın her iki versiyonu için düğüm çiftlerinde görülen son durumların önemini test etmeye adanmıştır. Bu maksatla eski olanlara nazaran daha sonra meydana gelen durumlara ağırlıklar vermek için iki monoton olarak artan fonksiyon kullanılmıştır. Bu fonksiyonların kullanılmasıyla önerilen tahmin edicinin yakınlık aşağıdaki gibi güncellenmiştir:

$$Skor(x, y) = \sum_{k=2}^n I(k)[P(x, y, k) + \alpha S(x, y, k)] \quad (6.1)$$

Burada k çerçeve sayısı ve $I(k)$ çerçevenin önem değerini temsil eder. Bu çalışmada $I(k)$, k ve $\log(k)$ olarak belirlenmiştir. Tablo 6.19’dan görülebileceği gibi, en iyi tahmin performansı ağın birinci ağırlıklı versiyonu üzerinde $\log(k)$ önemli birinci uygulama kümesinde başarılmıştır. Uygulamada ayrıca $I(k) = \log(k)$ olması durumundaki performans kazancının $I(k) = k$ ‘yı kullanan durum tabanlı skora göre daha yüksek olduğu belirlenmiştir. Bu durum önemli bir gerçeği ortaya çıkarır ki son yaş serileri bağlantıların tahmin edilmesinde daha fazla önemlidirler.

Tablo 6.19. Farklı önem fonksiyonları için önerilen yöntemin kesinlik değerleri (%)

Uygulama	Ağırlıklı-1		Ağırlıklı-2	
	k	$\log(k)$	k	$\log(k)$
1	59,7	60,8	55,2	57,3
2	56,0	57,5	57,9	58,6
3	51,4	52,0	49,1	49,8
4	51,8	52,7	54,3	55,2

6.3. Yaş Serileri Tabanlı Denetimli Bağlantı Tahmini Sonuçları

Tezin bu bölümünde Tablo 6.1’de verilen yaş serileri denetimli öğrenme yöntemi kullanılarak test edilmiştir. Hastalık belirti ağında olduğu gibi bağlantı tahmini için WEKA ortamındaki NaiveBayes, LibSVM, IBk ve J48 sınıflandırıcıları kullanılmıştır. Her bir yaş serisi için ağırlıklı versiyonları kullanılmıştır. Yöntemler 10 kat çapraz geçerleme ile değerlendirilerek 10 defa tekrarlanan sınıflandırma işinde performansların ortalaması alınmıştır. Geçerlilik kümesinde bağlantılı olmayan çiftler veya önceden belirlenen eşik değerinden daha küçük ağırlık derecesine sahip 400 çift rasgele seçilir ve ilgili nitelikler böylece 400 örnek oluşturur. Bu örneklerin 200’ü pozitif geri kalanı negatif olarak seçilir. Bu şekilde sınıflar denetimli öğrenme için eşit bir şekilde dağıtılır. Çerçeveler arasındaki durum geçişlerini belirleyen parametreler ve Eşik değeri hastalık ağının denetimsiz öğrenme yönteminkiyle aynıdır. 2, 3 ve 4 numaralı yaş serilerinde ağın gelişen yapısındaki çerçeve sayısı 3 olduğundan nitelik sayısı da 3 olarak belirlenmiştir. Karşılaştırma için kullanılacak sınıflandırma verisi için geleneksel tahmin edicilerden OK, JK, TB, AA ve KP’dan üçü rastgele seçilmiştir. 1 numaralı yaş serisinin gelişen yapısındaki çerçeve sayısı 7 olduğundan yine karşılaştırma için tüm tahmin ediciler nitelik belirlenmesinde kullanılmıştır. Tablo 6.20-6.23 dört farklı yaş serisi için oluşturulan veri kümesinin sınıflandırma işleminden sonra elde ettiği doğruluk ve kesinlik değerlerini gösterir. Tablolarda görüldüğü gibi önerilen yöntemle niteliği belirlenmiş veri kümelerinde sınıflandırmalar, geleneksel ölçülere dayalı olarak elde edilenlere üstünlük sağlamıştır. Genel olarak çerçeve boyutlarının küçük olduğu durumlarda ağın ilk versiyonu daha başarılı sonuçlar üretirken, boyutun daha fazla olduğu çerçevelerde ise durum değişim oranının daha yüksek olduğu ağın ikinci versiyonu daha iyi sonuçlar bulmuştur. Tek istisnai durum 4 numaralı yaş serisinde gözlenmiştir. Burada çerçeve boyutu küçük olmasına rağmen yaşlı hasta sayısı fazla olduğundan dolayı bazı sınıflandırıcılarda ağın birinci versiyonu bazılarında ise ağın ikinci versiyonu daha yüksek doğruluk ve kesinlik değerleri elde etmiştir.

Tablo 6.20. 1 nolu yaş serisinde karşılaştırılan yöntemlerin doğruluk (D) ve kesinlik (K) değerleri (%)

Yöntem	Ağırlıklandırılmış Ağ-1 (r=%25)		Ağırlıklandırılmış Ağ-2 (r=%50)	
	<u>Önerilen Yöntem</u>	<u>Geleneksel Ölçülere dayalı</u>	<u>Önerilen Yöntem</u>	<u>Geleneksel Ölçülere dayalı</u>
NaiveBayes	D: 67,6 K: 62,1	D: 62,3 K: 59,7	D: 62,5 K: 60,4	D: 58,1 K: 55,4
LibSVM	D: 65,5 K: 64,4	D: 62,2 K: 60,9	D: 62,7 K: 61,8	D: 58,3 K: 58,0
IBk	D: 62,2 K: 60,8	D: 60,0 K: 57,6	D: 58,3 K: 57,8	D: 57,5 K: 53,4
J48	D: 69,5 K: 67,6	D: 64,1 K: 63,4	D: 64,4 K: 63,7	D: 60,7 K: 60,1

Tablo 6.21. 2 nolu yaş serisinde karşılaştırılan yöntemlerin doğruluk (D) ve kesinlik (K) değerleri (%)

Yöntem	Ağırlıklandırılmış Ağ-1 (r=%25)		Ağırlıklandırılmış Ağ-2 (r=%50)	
	<u>Önerilen Yöntem</u>	<u>Geleneksel Ölçülere dayalı</u>	<u>Önerilen Yöntem</u>	<u>Geleneksel Ölçülere dayalı</u>
NaiveBayes	D: 60,4 K: 58,5	D: 54,3 K: 53,8	D: 63,0 K: 60,7	D: 57,6 K: 57,2
LibSVM	D: 57,9 K: 57,0	D: 55,5 K: 54,2	D: 62,6 K: 62,3	D: 56,8 K: 55,1
IBk	D: 55,2 K: 52,5	D: 50,6 K: 48,7	D: 58,2 K: 56,8	D: 57,5 K: 53,4
J48	D: 63,2 K: 60,9	D: 56,3 K: 55,4	D: 65,2 K: 64,3	D: 62,8 K: 59,5

Tablo 6.22. 3 nolu yaş serisinde karşılaştırılan yöntemlerin doğruluk (D) ve kesinlik (K) değerleri (%)

Yöntem	Ağırlıklandırılmış Ağ-1 (r=%25)		Ağırlıklandırılmış Ağ-2 (r=%50)	
	<u>Önerilen Yöntem</u>	<u>Geleneksel Ölçülere dayalı</u>	<u>Önerilen Yöntem</u>	<u>Geleneksel Ölçülere dayalı</u>
NaiveBayes	D: 61,0 K: 59,7	D: 55,6 K: 53,9	D: 58,0 K: 56,3	D: 52,6 K: 51,7
LibSVM	D: 60,2 K: 58,3	D: 53,3 K: 52,7	D: 55,4 K: 54,1	D: 52,6 K: 51,7
IBk	D: 55,6 K: 52,7	D: 48,3 K: 45,8	D: 49,2 K: 49,7	D: 44,0 K: 42,4
J48	D: 61,4 K: 60,2	D: 56,2 K: 55,0	D: 58,5 K: 56,8	D: 55,8 K: 53,2

Tablo 6.23. 4 nolu yaş serisinde karşılaştırılan yöntemlerin doğruluk (D) ve kesinlik (K) değerleri (%)

Yöntem	Ağırlıklandırılmış Ağ-1 (r=%25)		Ağırlıklandırılmış Ağ-2 (r=%50)	
	<u>Önerilen Yöntem</u>	<u>Geleneksel Ölçülere dayalı</u>	<u>Önerilen Yöntem</u>	<u>Geleneksel Ölçülere dayalı</u>
NaiveBayes	D: 60,8 K: 59,0	D: 52,3 K: 52,1	D: 61,4 K: 61,3	D: 57,1 K: 56,2
LibSVM	D: 61,5 K: 60,8	D: 52,6 K: 51,8	D: 61,1 K: 60,6	D: 56,0 K: 55,3
IBk	D: 56,8 K: 54,7	D: 48,6 K: 48,5	D: 55,2 K: 53,8	D: 52,6 K: 51,8
J48	D: 62,3 K: 61,0	D: 55,8 K: 54,9	D: 63,3 K: 61,8	D: 58,5 K: 57,9

7. SONUÇLAR VE ÖNERİLER

Sosyal ağlarda bağlantı tahmini ve geçmiş hasta kayıtlarından hastalıkların tahmin edilmesi, son yıllarda bilgisayar bilimleri dalında en sık çalışılan araştırma konularındandır. Bu tezde, bu iki alan bağlantı tahmini üzerinde birleştirilerek yeni öğrenme sistemleri geliştirilmiştir. Bu bağlamda tezde sunulan katkılar aşağıda sıralanmıştır:

- Öncelikle hastalık belirti ağı ve hastalık ağ adı altında iki farklı ağ oluşturulmuştur. Bu ağlar ağırlıklıdır ve düğümler arasındaki bağlantıların oluşma biçimi her bir ağ için farklıdır.
- Gelecek t^l zamandaki yeni bağlantıları tahmin etmek için t zamandaki ağı dikkate alan klasik statik yaklaşımlardan farklı olarak önerilen yöntemler yaşa göre ağın nasıl geliştiğini izler ve tahmin yapmak için ağın gelişen yapısını dikkate alır.
- Yaşlara göre ağların gelişen yapısını göz önüne alan yeni bağlantı tahmini yöntemleri önerilmiştir. Önerilen bağlantı tahmini algoritmaları, denetimli ve denetimsiz öğrenme için ayrı ayrı geliştirilmiştir.
- Eski olanlara nazaran daha sonra meydana gelen durumlara ağırlık vererek son çerçevelerdeki durumların önemi test edilmeye çalışılmıştır.

Fırat Üniversitesi Hastanesi'nden etik kurallara dikkat edilerek alınan 210.134 hasta kaydı üzerinde yapılan uygulamalarla aşağıdaki sonuçlar elde edilmiştir:

- Hastalık belirti ağı üzerinde parametreler arasındaki bağlantıların hastalıkların yaş değişikliğine bağlı olarak bir gelişim gösterdiği saptanmış ve önerilen yöntem ile tahmin edilen bağlantıların diğer yöntemlere göre daha doğru sonuçlar verdiği tespit edilmiştir.
- Hastalık belirti ağı ve hastalık ağında denetimli öğrenmenin, denetimsiz öğrenmeye göre nispeten daha iyi sonuçlar verdiği belirlenmiştir.
- Oluşturulan hastalık ağı üzerinde hastalıklar arasındaki önerilen bağlantıların yaş değişikliğine bağlı olarak yine bir gelişim gösterdiği ve bu bağlantıların tahmininin önerilen yöntemle daha doğru sonuçlar verdiği izlenmiştir.

- Çerçeve sayısı veya yaş serileri arttığı zaman, doğruluğu yüksek sonuçlar bulunmuştur. Bunun nedeni, her bir durum değişikliğinin daha sık izleniyor olmasıdır.
- Son çerçevelere veya yaş serilerine daha fazla önem verilerek, bağlantının doğru tahmin edilme başarımı artırılmıştır.
- Çerçeve boyutunun küçük olduğu yerlerde düşük durum değişim oranlı ağ daha iyi sonuçlar verirken, çerçeve boyutunun büyük alındığı yerlerde yüksek durum değişim oranlı ağ daha iyi sonuçlar bulmuştur.

Gelecek çalışmalarda, durum tabanlı olarak önerilen yakınlık fonksiyonunun aşağıdaki şekilde geliştirilmesi planlanmaktadır:

- Önerilen yöntem, yönlü hastalık ağlarına uygulanacaktır. Bu sayede sadece hastalıklar arası tetiklenme belirlenmeyecek, hangi hastalığın öncü hangisinin ise sonradan gelişen olduğu tespit edilmeye çalışılacaktır.
- Bu tezde oluşturulan hastalık belirti ağı ve hastalık ağının her ikisinde düğümler aynı türdendir. Yani ya hastalık belirtisini veya hastalığı göstermektedir. Sonraki çalışmada bir tarafın hastalık diğer tarafında hastalık belirtisi olduğu çoklu mod ağ yapısı kullanılacaktır. Bu sayede hangi hastalığı hangi parametrelerin etkilediği belirlenmiş olacaktır.
- Son olarak önerilen durum tabanlı yöntem, yazar-konu ağına da uygulanacaktır. Bu sayede yazarların geçmişte çalıştıkları konulara bakarak gelecekte hangi konuları çalışmaya aday oldukları belirlenmeye çalışılacaktır.

KAYNAKLAR

- [1] **Wasserman, S. and Faust, K.,** 1994. Social Network Analysis: Methods and Applications, Cambridge University Press, England.
- [2] **Xiang, E.W.,** 2008. A Survey on Link Prediction Models for Networked Data, PhD Qualifying Exam, Department of Computer Science and Engineering, The Hong Kong University of Science and Technology.
- [3] **Hasan, M.A. and Zaki, M.J.,** 2011. A Survey of Link Prediction in Social Network, Social Network Data Analytics, pp. 243-275.
- [4] **Newman, M.E.J.,** 2001. Clustering and preferential attachment in growing networks, Phys. Rev. E 64 , 025102.
- [5] **Barabasi, L. and Bonabeau, E.,** 2003. Scale Free Networks, **288 (5)**, 50-59.
- [6] **Adamic, L.A. and Adar, E.,** 2003. Friends and neighbors on the web, Social Networks **25 (3)**, 211–230.
- [7] **Tan, P.-N. and Kumar, V.,** 2005. Introduction to Data Mining, Addison Wesley.
- [8] **Katz, L.,** 1953. A new status index derived from sociometric analysis, Psychometrika, **18(1)**, 39-43.
- [9] **Liben-Nowell D., and Kleinberg J.,** 2007. The link-prediction problem for social networks, J. Am. Soc. Inform. Sci. Technol., **58(7)**, 1019–1031.
- [10] **Lu, L. and Zhou, T.,** 2011. Link prediction in complex networks: a survey, Phys. A: Statist. Mech. Appl., **390 (6)**, 1150–1170.
- [11] **Hasan, M.A., Chaoji, V., Salem, S. and Zaki, M.,** 2006. Link prediction using supervised learning, *SDM'06: Workshop on Link Analysis Counterterrorism and Security*.
- [12] **Folino, C. F. and Pizzuti.,** 2012. Link Prediction Approaches for Disease Networks, *ITBAM*, pp. 99–108.
- [13] **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, pp. 243-252.

- [14] **Sa, H.R. and Pruencio, R.B.C.**, 2011. Supervised link prediction in weighted networks, in: *Proc. of International Joint Conference on Neural Networks*, pp. 2281–2288.
- [15] **Tylenda, T., Angelova, R. and Bedathur S.**, 2009. Towards time-aware link prediction in evolving social networks, *SNA-KDD'09*, pp. 1-9.
- [16] **Brigmann, B., Berlingerio, M., Bonchi, F. and Gionis, A.**, 2010. Learning and predicting the evolution of social networks, *IEEE Intell. Syst.*, **25 (4)**, 26-35.
- [17] **Huang, Z. and Lin, D.K.J.**, 2009. The time-series link prediction problem with applications in communication surveillance, *INFORMS J. Comput.*, **21(2)**, 286–303.
- [18] **Oyama, S., Hayashi, K. and Kashima, K.**, 2011. Cross-temporal link prediction, in: *IEEE International Conference on Data Mining*, pp. 1188-1193.
- [19] **Soares, P.R.S. and Prudencio, R.B.C.**, 2013. Proximity measures for link prediction based on temporal events, *Expert Syst. Appl.*, **40 (16)**, 6652–6660.
- [20] **Davis, D., Lichtenwalter R., and Chawla N.V.**, 2013. Supervised methods for multi-relational link prediction, *Social Network Anal. Min.*, **3 (2)**, 127-141.
- [21] **Li, X. and Chen, H.**, 2013. Recommendation as link prediction in bipartite graphs: a graph kernel-based machine learning approach, *Decis. Support Syst.*, **54 (2)**, 880-890.
- [22] **Symeonidis, P., Iakovidou, N., Mantas, N. and Manolopoulos, Y.**, 2013. From biological to social networks: link prediction based on multi-way spectral clustering, *Data Knowl. Eng.*, **87**, 226–242.
- [23] **Kaya, B** and Poyraz, M., 2014. “Supervised link prediction in symptom networks with evolving case”, *Measurement*, **56**, pp.231-238.
- [24] **Kaya, B.** and Poyraz, M, 2014. “Link Prediction in Weighted Symptom Networks”, *15th IEEE International Symposium on Computational Intelligence and Informatics*, Budapest.
- [25] **Köksal, A.**, 1981. Bilişim Terimleri Sözlüğü, Türk Dil Kurumu Yayınları, 126.
- [26] **Karcı, A., Boy, O.**, 2011. Sosyal Ağların Web Madenciliği ile Analizi ve Ortak Atıf Analizi ile Benzerlik Tahmini, *Elektrik-Elektronik ve Bilgisayar Sempozyumu*,154.

- [27] **Jones, N., Blackey, H., Fitzgibbon, K., Chew, E.,** Get out of MySpace, *Computers & Education*, **54**, 776–782.
- [28] **Chang, W., Lin, T.,** 2010. A Cluster-Based Approach for Automatic Social Network Construction, *IEEE International Conference on Social Computing / IEEE International Conference on Privacy, Security, Risk and Trust*, 601-606.
- [29] **Bondy, J.A., Murty, U.S.R.,** 2008. *Graph Theory*, Springer.
- [30] <http://www.orgnet.com/sna.html> Social Network Analysis, A Brief Introduction, 19 Mart 2013.
- [31] **Freeman, L. ,** 1977. A set of measures of centrality based upon, *Sociometry*, **40**, 35–41.
- [32] **Girvan, M. and Newman, M. E. J.,** 2002. Community structure in social and biological Networks, *Proc. Natl. Acad. Sci.*, **99**, 7821–7826.
- [33] http://discopal.ispras.ru/SocialGraphs/Community_Detection, *SocialGraphs/Community Detection*, 21 Mart 2013.
- [34] **Newman, M. E. J. and Girvan M.,** 2004. Finding and evaluating community structure in networks, *Phys. Rev. E* , vol. **69**, no. 026113.
- [35] **Gürsakal Necmi,** 2009. *Sosyal Ağ Analizi*. Bursa, Türkiye: Dora.
- [36] <http://www.soc.ucsb.edu/faculty/friedkin/Syllabi/Soc148MA/Eigenvector%20Centrality.pdf>, *EIGENVECTOR CENTRALITY*, 21 Mart 2013.
- [37] **Menon, A.K. ve Elkan, C.,** 2011. Link Prediction via Matrix Factorization, *ECML/PKDD*, **2**, 437-452.
- [38] **Getoor, L. ve Diehl, C.P.,** 2005. “Link Mining: A Survey”, *SIGKDD Explorations*, **7(2)**, 3-12.
- [39] **Scott, J.,** 2000. “*Social Network Analysis: A Handbook*,” Sage Publications, 209s.
- [40] **Liben-Nowell, D., Kleinberg, J.,** 2003. The Link Prediction Problem for Social Networks, In *Proceeding of the 2003 International Conference on Information and Knowledge Management*, pp: 556-559.
- [41] **Dhote, Y., Mishra, N., Sharma, S.,** 2013. Survey and analysis of temporal link prediction in online social networks, *International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp: 1178-1183.
- [42] **Newman, Mark E. J.,** 2001. Scientific collaboration networks. i.network construction and fundamental results, *Physical Review E*, **64**:016131.

- [43] **Newman, M. E. J., Watts J. D., and Strogatz, S. H.,** 2002. Random graph models of social networks. *Proceedings of National Academy of Sciences of the United States of America*, **99**, 2566–2572.
- [44] **Newman, M. E. J.,** 2001. Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Physical review E*, 64(1):016132.
- [45] **Barabási, Albert-László,** 2009. From networks to human behavior. In *Proceedings of Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 435.
- [46] **Clauset A., Shalizi, C.R. and Newman, M. E. J.,** Power-law distributions in empirical data. *SIAM Review*, **51**, 661–703.
- [47] **Wang, C., Satuluri V. and Parthasarathy, S.,** 2007. Local probabilistic models for link prediction. In *Proceedings of 7th IEEE International Conference on Data Mining*, pages 322–331.
- [48] **Tylenda, T., Angelova R., and Bedathur. S.,** 2009. Towards time-aware link prediction in evolving social networks, In *Proceedings of 3rd Workshop on Social Network Mining and Analysis*, page 9.
- [49] **Kashima, H. and Abe N.,** 2006, A parameterized probabilistic model of network evolution for supervised link prediction, In *Proceedings of 6th IEEE International Conference on Data Mining*, pages 340–349.
- [50] **Leroy, V., Cambazoglu, B. and Bonchi, F.,** 2010. Cold start link prediction. In *Proceedings of 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 393–402.
- [51] **Popescul, A., Ungar L.H., Lawrence, S. and Pennock. D.,** 2003. Statistical relational learning for document mining. In *Proceedings of 3rd IEEE International Conference on Data Mining*, pages 275–282.
- [52] **Juszczyszyn, K., Gonczarek, A., Tomczak, J.M., Musial, K. and Budka, M.,** 2012. A probabilistic approach to structural change prediction in evolving social networks. In *Proceedings of International Conference on Advances in Social Networks Analysis and Mining*, pages 996–1001.
- [53] **Leskovec, J., Backstrom, L., Kumar, R. and Tomkins, A.,** 2008. Microscopic evolution of social networks. In *Proceedings of 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 462–470.

- [54] **Xiang E. W.**, 2008. A survey on link prediction models for social network data, *PhD thesis*, Department of Computer Science and Engineering, The Hong Kong University of Science and Technology.
- [55] **Soares, P.R.S. and Prudencio, R.B.C.** , 2013. Proximity measures for link prediction based on temporal events, *Expert Systems with Applications*, vol. **40**, no. 16, pp. 6652-6660.
- [56] **Hasan, M.A. and Zaki, M.J.** , 2011. A survey of link prediction in social networks, *Social network data analytics*, pp. 243-275.
- [57] **Liben-Nowell, D. and Kleinberg, J.** , 2007. The link prediction problem for social networks, *Journal of the American society for information science and technology*, vol. **58**, no. 7, pp. 1019-1031.
- [58] **Newman M.E.J** , 2001. Clustering and preferential attachment in growing networks, *Physical Review E*, vol. **64**, no. 2, p. 025102.
- [59] **Jaccard P.** , 1901. Etude comparative de la distribution florale dans une portion des Alpes et du Jura.
- [60] **Adamic L. A. and Adar E.** , 2003. Friends and neighbors on the web, *Social networks*, vol. **25**, no. 3, pp. 211-230.
- [61] **Barabasi A.L. et al.**, 2002. Evolution of the social network of scientific collaborations, *Physica A: Statistical Mechanics and its Applications*, vol. **311**, pp. 590-614.
- [62] **Moody J.**, 2001. Race, School Integration, and Friendship Segregation in America, *American Journal of Sociology* **107**, 679–716.
- [63] **Barabasi, A. L. and Bonabeau, E.**, 2003. Scale-Free Networks. *Scientific American* **5**, 60-69.
- [64] **Katz, L.**, 1953. A new status index derived from sociometric analysis, *Psychometrika*, **18**, pp. 39-43.
- [65] **Brin, S. and Page, L.** , 1998. The anatomy of a large-scale hypertextual Web search engine, *Computer networks and ISDN systems*, **30**, no. 1, pp. 107-117.
- [66] **Jeh, G. and Widom J.** , 2002. SimRank: a measure of structural-context similarity, *ACM*, pp. 538-543.
- [67] **Lichtenwalter R.N. , Lussier J.T. , and Chawla N.V.** , 2010. New perspectives and methods in link prediction, *Proceedings of the 16th ACM SIGKDD*

- international conference on Knowledge discovery and data mining*, pp. 243-252.
- [68] **Kashima, H., Kato T., Yamanishi, Y., Sugiyama, M. and Tsuda, K.**, 2009. Link propagation: A fast semi-supervised learning algorithm for link prediction. *In Proceedings of SIAM International Conference on Data Mining*, pages 1099–1110.
 - [69] **Cha, M., Mislove A. and Gummadi, P.K.**, 2009. A measurement driven analysis of information propagation in the flickr social network. *In Proceedings of 18th International Conference on World Wide Web*, pages 721-730.
 - [70] **Backstrom, L. and Leskovec J.**, 2011. Supervised random walks: predicting and recommending links in social networks. *In Proceedings of Forth International Conference on Web Search and Web Data Mining*, pages 635–644.
 - [71] **Scripps, J., Tan P.N., Chen, and Abdol-Hossein Esfahanian**, 2008. A matrix alignment approach for link prediction. *In 19th International Conference on Pattern Recognition*, pages 1–4.
 - [72] **Kahanda, I. and Neville, J.**, 2009. Using transactional information to predict link strength in online social networks. *In Proceedings of 3rd International Conference on Weblogs and Social Media*.
 - [73] **Huang, Z., Li, X. and Chen, H.**, 2005. Link prediction approach to collaborative filtering. *In Proceedings of ACM/IEEE Joint Conference on Digital Libraries*, pages 141–142.
 - [74] **Witten, I.H. and Frank, E.**, 2005. Data Mining: Practical machine learning tools and techniques.
 - [75] **Quinlan, J. R.**, 1993. C4.5: programs for machine learning.
 - [76] **Breiman, L.**, 1996. Bagging predictors, *Machine learning*, vol. 24, no. 2, pp. 123-140.
 - [77] **Breiman L.**, 2001. Random forests, *Machine learning*, vol. 45, no. 1, pp. 5-32.
 - [78] **Al Hasan M. , Chaoji V. , Salem S. , and Zaki M.**, 2006. Link prediction using supervised learning, *SDM'06: Workshop on Link Analysis, Counter-terrorism and Security*.
 - [79] **Homans G.C.**, 1951 *The human group*, London: Routledge and Kegan.

- [80] **Newman M.E.J.**, The structure of scientific collaboration networks, In Proc. of the national academy of sciences of the United States of America, vol.98, pp.404-409.
- [81] **Ou Q, Jin Y.-D., Zhou T., Wang B.-H. and Yin B.-Q**, 2007. Power-law strength-degree correlation from resource-allocation dynamics on weighted network, Phys. Rev. E, vol.75, no.2, pt 1, 021102.

ÖZGEÇMİŞ

Buket KAYA

Fırat Üniversitesi
Elazığ Organize Sanayi Bölgesi Maden MYO

e-posta: bkaya@firat.edu.tr

- | | |
|-------------------|---|
| 1986 | Elazığ'da doğdu. |
| 2003-2007 | Fırat Üniversitesi, Mühendislik Fakültesi, Elektrik-Elektronik Mühendisliğini okudu. |
| 2007-2010 | Fırat Üniversitesi, Tunceli MYO'da Uzman Öğretim Elemanı olarak görev yaptı. |
| 2008-2010 | Fırat Üniversitesi, Fen Bilimleri Enstitüsü, Biyomühendislik Bölümü'nde yüksek lisans eğitimini tamamladı. |
| 2010-devam | Fırat Üniversitesi, Elazığ Organize Sanayi Bölgesi Maden MYO'da Öğretim Görevlisi olarak çalışmaktadır. |
| 2011- | Fırat Üniversitesi, Fen Bilimleri Enstitüsü, Elektrik-Elektronik Mühendisliği Anabilim dalında Doktora çalışmasına başladı. |