T.C.
HACETTEPE ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

SOSYAL AĞ ANALİZİNİN HASTALIK BİYOBELİRTEÇLERİNİN BELİRLENMESİNDE KULLANIMI

Hatice Yağmur ZENGİN

Biyoistatistik Programı
DOKTORA TEZİ

ANKARA
2018

T.C.
HACETTEPE ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

SOSYAL AĞ ANALİZİNİN HASTALIK BİYOBELİRTEÇLERİNİN BELİRLENMESİNDE KULLANIMI

Hatice Yağmur ZENGİN

Biyoistatistik Programı
DOKTORA TEZİ

TEZ DANIŞMANI
Prof. Dr. Erdem KARABULUT

ANKARA
2018

SOSYAL AĞ ANALİZİNİN HASTALIK BİYOBELİRTEÇLERİNİN BELİRLENMESİNDE KULLANIMI

Hatice Yağmur ZENGİN

Danışman: Prof. Dr. Erdem KARABULUT

Bu tez çalışması 27/06/2018 tarihinde jürimiz tarafından "Biyoistatistik Programı" nda doktora tezi olarak kabul edilmiştir.

Jüri Başkanı:	*Prof. Dr. C. Reha ALPAR* *(Hacettepe Üniversitesi)*
Üye:	*Prof. Dr. Ersin ÖĞÜŞ* *(Başkent Üniversitesi)*
Üye:	*Prof. Dr. Ergun KARAAĞAOĞLU* *(Hacettepe Üniversitesi)*
Üye:	*Prof. Dr. Atilla Halil ELHAN* *(Ankara Üniversitesi)*
Üye:	*Dr. Öğr. Üyesi Sevilay KARAHAN* *(Hacettepe Üniversitesi)*

Bu tez, Hacettepe Üniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliğinin ilgili maddeleri uyarınca yukarıdaki jüri tarafından uygun bulunmuştur.

24 Temmuz 2018

Prof. Dr. Diclehan ORHAN

Enstitü Müdürü

YAYIMLAMA VE FİKRİ MÜLKİYET HAKLARI BEYANI

Enstitü tarafından onaylanan lisansüstü tezimin/raporumun tamamını veya herhangi bir kısmını, basılı (kağıt) ve elektronik formatta arşivleme ve aşağıda verilen koşullarla kullanıma açma iznini Hacettepe Üniversitesine verdiğimi bildiririm. Bu izinle Üniversiteye verilen kullanım hakları dışındaki tüm fikri mülkiyet haklarım bende kalacak, tezimin tamamının ya da bir bölümünün gelecekteki çalışmalarda (makale, kitap, lisans ve patent vb.) kullanım hakları bana ait olacaktır.

Tezin kendi orijinal çalışmam olduğunu, başkalarının haklarını ihlal etmediğimi ve tezimin tek yetkili sahibi olduğumu beyan ve taahhüt ederim. Tezimde yer alan telif hakkı bulunan ve sahiplerinden yazılı izin alınarak kullanılması zorunlu metinlerin yazılı izin alınarak kullandığımı ve istenildiğinde suretlerini Üniversiteye teslim etmeyi taahhüt ederim.

Yükseköğretim Kurulu tarafından yayınlanan ***"Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge"*** kapsamında tezim aşağıda belirtilen koşullar haricince YÖK Ulusal Tez Merkezi / H.Ü. Kütüphaneleri Açık Erişim Sisteminde erişime açılır.

- Enstitü / Fakülte yönetim kurulu kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren 2 yıl ertelenmiştir. [1]
- Enstitü / Fakülte yönetim kurulunun gerekçeli kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren ... ay ertelenmiştir. [2]
- Tezimle ilgili gizlilik kararı verilmiştir. [3]

26./.07./.2018

Hatice Yağmur ZENGİN

[i]

---

[i] *"Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge"*

*(1) Madde 6. 1. Lisansüstü tezle ilgili patent başvurusu yapılması veya patent alma sürecinin devam etmesi durumunda, tez **danışmanının** önerisi ve **enstitü anabilim dalının** uygun görüşü üzerine **enstitü** veya **fakülte yönetim kurulu** iki yıl süre ile tezin erişime açılmasının ertelenmesine karar verebilir.*

*(2) Madde 6. 2. Yeni teknik, materyal ve metotların kullanıldığı, henüz makaleye dönüşmemiş veya patent gibi yöntemlerle korunmamış ve internetten paylaşılması durumunda 3. şahıslara veya kurumlara haksız kazanç imkanı oluşturabilecek bilgi ve bulguları içeren tezler hakkında tez **danışmanının** önerisi ve **enstitü anabilim dalının** uygun görüşü üzerine **enstitü** veya **fakülte yönetim kurulunun** gerekçeli kararı ile altı ayı aşmamak üzere tezin erişime açılması engellenebilir.*

*(3) Madde 7. 1. Ulusal çıkarları veya güvenliği ilgilendiren, emniyet, istihbarat, savunma ve güvenlik, sağlık vb. konulara ilişkin lisansüstü tezlerle ilgili gizlilik kararı, **tezin yapıldığı kurum** tarafından verilir *. Kurum ve kuruluşlarla yapılan işbirliği protokolü çerçevesinde hazırlanan lisansüstü tezlere ilişkin gizlilik kararı ise, **ilgili kurum ve kuruluşun önerisi** ile **enstitü** veya **fakültenin** uygun görüşü üzerine **üniversite yönetim kurulu** tarafından verilir. Gizlilik kararı verilen tezler Yükseköğretim Kuruluna bildirilir.*
Madde 7.2. Gizlilik kararı verilen tezler gizlilik süresince enstitü veya fakülte tarafından gizlilik kuralları çerçevesinde muhafaza edilir, gizlilik kararının kaldırılması halinde Tez Otomasyon Sistemine yüklenir

** Tez **danışmanının** önerisi ve **enstitü anabilim dalının** uygun görüşü üzerine **enstitü** veya **fakülte yönetim kurulu tarafından karar verilir.***

ETİK BEYAN

Bu çalışmadaki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi, görsel, işitsel ve yazılı tüm bilgi ve sonuçları bilimsel ahlak kurallarına uygun olarak sunduğumu, kullandığım verilerde herhangi bir tahrifat yapmadığımı, yararlandığım kaynaklara bilimsel normlara uygun olarak atıfta bulunduğumu, tezimin kaynak gösterilen durumlar dışında özgün olduğunu, Prof. Dr. Erdem KARABULUT danışmanlığında tarafımdan üretildiğini ve Hacettepe Üniversitesi Sağlık Bilimleri Enstitüsü Tez Yazım Yönergesine göre yazıldığını beyan ederim.

Hatice Yağmur ZENGİN

TEŞEKKÜR

Öncelikle, doktora eğitimim boyunca ve tez çalışmamın her aşamasında büyük bir özveri, sabır ve içtenlikle bana destek olan; engin bilgisini, deneyimlerini ve yol göstericiliğini benden esirgemeyen tez danışmanım ve saygı değer hocam sayın Prof. Dr. Erdem KARABULUT'a, annem gibi bana kol kanat geren ve doktora eğitimim ve tez yazım sürecimde değerli bilgi ve destekleriyle yanımda olan saygı değer hocam sayın Prof. Dr. Ersin ÖĞÜŞ'e, lisans, yüksek lisans ve doktora eğitimim süresince benden sevgi, ilgi ve bilgisini esirgemeyen kıymetli hocam sayın Prof. Dr. Gül ERGÜN'e, lisans, yüksek lisans ve doktora eğitimim boyunca beni yetiştirmiş olan değerli hocalarıma,

Son olarak, bana inanan ve maddi, manevi bana yardımcı olan sevgili annem, babam ve eşime teşekkürlerimi sunarım.

ÖZET

Zengin, H.Y., Sosyal Ağ Analizinin Hastalık Biyobelirteçlerinin Belirlenmesinde Kullanımı, Hacettepe Üniversitesi Sağlık Bilimleri Enstitüsü Biyoistatistik Programı Doktora Tezi, Ankara, 2018. Özellikle son yıllarda, hastalığa özgü biyobelirteçlerin belirlenmesi amacıyla yapılan çalışmalarda sosyal ağ analizinin kullanımı ilgi uyandırmaktadır. Özellik seçim (*feature selection*) sürecinin bir adımı olarak sosyal ağ analizinin yer aldığı melez (*hybrid*) yöntemler ile hastalığa özgü biyobelirteçlerin belirlenmesi problemine farklı bir bakış açısı getirilmektedir. Bu tez çalışmasında, "Sosyal Ağ Özellik Seçimi (*SocialNetworkFeature Selection*, SNFS)" olarak adlandırılan melez yöntemin farklı aşamalarında kullanılan boyut indirgeme, kümeleme ve topluluk belirleme yöntemleri kısaca incelenmiş; erişime açık genomik mikrodizi veri setleri kullanılarak, SNFS'nin adımlarında yer alan bu yöntemlerin farklı kombinasyonları, Destek Vektör Makinesi (DVM) sınıflayıcısının sınıflama başarımına etkileri açısından karşılaştırılmıştır. Aynı zamanda, SNFS kullanılarak DVM sınıflayıcısından elde edilen sınıflama başarımındaki değişimlerin incelenmesi amacıyla bir benzetim çalışması yapılmıştır.

Sonuç olarak, R'da uygulanan SNFS yönteminin DVM sınıflayıcısının sınıflama başarımını ciddi oranda iyileştirdiği ve yüksek boyutlu veriler söz konusu olduğunda SNFS ile boyut indirgemenin sınıflama başarımı üzerinde olumlu etkisi olduğu görülmüştür.

Anahtar Kelimeler: Sosyal Ağ Analizi, Melez Özellik Seçimi, SNFS, Sınıflama, Biyobelirteçler

ABSTRACT

Zengin, H.Y., The Use of Social Network Analysis in Disease Biomarker Detection, Hacettepe University Institute of Health Sciences, Ph.D. Thesis in Biostatistics, Ankara, 2018. Especially, in recent years, the use of social network analysis has gained interest in biomarker discovery studies. Hybrid approaches involve social network analysis as a step of the feature selection process bring a different perspective to identify disease-specific biomarkers. In this thesis, dimension reduction, clustering and community detection methods used in the different steps of the hybrid approach called "SocialNetworkFeature Selection (SNFS)" were briefly reviewed; the different combinations of these methods in the steps of SNFS were compared by using open access genomic microarray data sets in terms of the effects on classification performance of Support Vector Machine (SVM) classifier. In addition, a simulation study was conducted to examine the changes in classification performance obtained from SVM classifier with the use of SNFS.

In conclusion, it had been seen that SNFS approach applied in R improves the classification performance of SVM classifier tremendously and dimension reduction with SNFS has positive effects on classification performance in case of high dimensional data.

Key Words: Social Network Analysis, Hybrid Feature Selection, SNFS, Classification, Biomarkers

İÇİNDEKİLER

SİMGELER ve KISALTMALAR

ALL	Akut Liphokistik Lösemi
AML	Akut Myeloid Lösemi
AUC	Eğri Altında Kalan Alan (*Area Under Curve*)
cDNA	Komplementer DNA
CBFS	Korelasyon temelli özellik seçimi (*Correlation based feature selection*)
COV	Kapsama (*Coverage*, Cov)
CS	Ki-Kare (*Chi-Square*)
DD	Ayrımsal Görüntü (*Differential Display*)
DNA	Deoksiribo Nükleik Asit
DSO	Doğru Sınıflama Oranı (*Accuracy*)
DVM	Destek Vektör Makinesi (*Support Vector Machine*, SVM)
EFA	Entropik Filtreleme Algoritması (*Entropic Filtering Algorithm*)
ESD	*"Equalization of cDNAs"* ve *"Subtractive hybridization and Differential display"* kelimelerinin baş harfleri
EST	İfadelenmiş Dizi Etiketleri (*Expressed Sequence Tags*)
FCBF	Hızlı korelasyon temelli filtre (*Fast Correlation Based Filter*)
FRFS	*Fuzzy Rough Set Feature Selection*
FS-P	Özellik Seçimi – Perseptron (*Feature Selection – Perceptron*)
ID	*Identity*
IFP	*Iterative Perturbation Method*
IG	Bilgi Kazancı (*Information Gain*)
K-YÖE	Karekök-Özyinelemeli Özellik Eleme (*Square Root-Recursive Feature Elimination*, SQRT-RFE)
M_FS	Çoklu-görev Filtresi (*Multi-task Filter*)
MPSS	Büyük Çaplı Paralel DNA Dizileme (*Massive Parallel Signature Sequencing*)
mRMR	Minimum gereksizlik maksimum ilgililik (*minimum Redundancy Maximum Relevance*)
mRNA	Haberci (*Messenger*) RNA

MWMR Maksimum ağırlık minimum gereksizlik (*Maximum weight minimum redundancy*)

NP-hard *Non-deterministic polynomial-time hard*

PAC *Probably Approximately Correct*

PLS Kısmi En Küçük Kareler (*Partial Least Squares*)

PZR Polimeraz Zincir Reaksiyonu (*Polymerase Chain Reaction*)

RDA *Representational Difference Analysis*

RF *Random Forest*

RFS Robust Özellik Seçimi (*Robust Feature Selection*)

RNA Ribo Nükleik Asit

RF-YÖE *Random Forest*-Özyinelemeli Özellik Eleme (*Random Forest-Recursive Feature Elimination*, RF-RFE)

RRFS İlgililik Gereksizlik Özellik Seçimi (*Relevance Redundancy Feature Selection*)

SAGE Gen İfadesinin Seri Analizi (*Serial Analysis of Gene Expression*)

SHL Çıkartılmış Melezleme Kütüphanesi (*Subtractive Hybridization Library*)

SNFS Sosyal Ağ Özellik Seçimi (*SocialNetworkFeature Selection*)

ŞEKİLLER

TABLOLAR

1. GİRİŞ

1.1. Probleme Genel Bakış

Günümüzde pek çok pozitif bilim dalı, teknolojinin de gelişmesiyle küçük boyutlu veri kümeleri ile uğraşan disiplinlerden, çok sayıda ve yüksek boyutlu verilerin analizinin söz konusu olduğu alanlara evrimleşmektedir.

Yüksek boyutlu verilerin varlığı, özellikle genetik alanında, veri madenciliği yöntemlerinin kullanımını araştırmacılar için giderek vazgeçilmez kılmakta ve çok disiplinli bir çalışma alanı oluşturmaktadır. Yaşanan bu büyük değişimin, İnsan Genom Projesi sonucunda insan genomuna ilişkin sonuçların yayınlanması ve eş zamanlı olarak binlerce gen ifadesi ölçümünün hızlı ve ekonomik şekilde elde edilmesini sağlayan genomik mikrodizi teknolojisinin ortaya çıkışı olmak üzere iki temel nedeni vardır (1, 3).

İnsan DNA'sında bulunan genlerinin yapısını, organizasyonunu ve fonksiyonunu kapsamlı şekilde açıklamayı amaçlayan uluslararası araştırma programı İnsan Genom Projesi kapsamında, 2001 Şubat'ında insan genomunun yani 3 milyar baz çiftinin %90'ının, 2003 Nisan'ında ise tamamının dizilenmesine ilişkin sonuçlar yayınlanmıştır. Bu proje, insana –türün kendi içinde ve türler arasında söz konusu olan farklılıklara– ilişkin temel soruların yanıtlanabilmesi için araştırmacılara yeni bir kapı açmıştır. Bu nedenle projenin Gregor Mendel ile başlayan süreçte tarihi bir dönüm noktası olduğu söylenebilir (2, 3). Böylece son 15 yıl içerisinde, İnsan Genom Projesi ve tetiklediği araştırmalar ile birlikte teknolojideki gelişmeler sonucunda, organizmaların dizilenmiş genomlarına ilişkin bilgi içeren büyük veri yığınları ortaya çıkmıştır (1).

Bu gelişmelerden önce araştırmacılar genellikle örnek sayısı değişken sayısından daha yüksek olan ve yüzlerce biyolojik örnek içeren veri kümeleri ile uğraşmakta; verilerin analizinde ise çoğunlukla geleneksel istatistiksel yöntemler kullanmakta ve sezgisel yaklaşımlara çok fazla ilgi göstermemekteydi. Ancak, güncel mikrodizi yöntemlerinin kullanılmasıyla elde edilen veri setleri çoğunlukla binlerce değişken ve buna karşılık, düzineler ile sınırlı olmak üzere az sayıda biyolojik örnek içermektedir. Hatta protein çip teknolojilerinin ya da ekson dizilerinin kullanımı ile milyonlarca değişken içeren veri kümelerinin analizi söz konusu olabilmektedir. Bu

nedenle, özellikle eş zamanlı olarak birçok genin incelenmesinin amaçlandığı mikrodizi gen ifade çalışmalarına, yüksek boyutlu verilerin depolanmasında, düzenlenmesinde, veriye erişimin kolaylaştırılmasında, uygun istatistiksel analizler için araç ve yöntemlerin geliştirilmesinde görev alacak veri madenciliği yöntemleri konusunda uzmanlaşmış araştırmacıların katılımı yaşamsal önem taşımaktadır (1, 4). Verilerin mikrodizi yöntemleri ile elde edildiği kanser araştırmalarında ise bu gereksinim daha da büyük bir önem kazanmaktadır (5). Çünkü hastalıklar üzerinde etkili olan "biyobelirteç" genlerin kullanılmasıyla bireylerin tanı ile tedavisinde doğrudan başarı sağlanabilmekte ve hastalık ilerlemeden kişiye özgü önleyici tedaviler uygulanabilmektedir (6). Bu nedenle, bu tür kanser araştırmalarının önemli bir bölümünü kanser alt sınıflarının keşfi ve biyobelirteç olabilecek önemli genlerin seçimi oluşturmaktadır. Ancak, kanser alt sınıflarının keşfi problemini ele alan çalışmalarda, bireylerin ait oldukları sınıflar genellikle bilinmediğinden, alanyazında sıklıkla danışmansız öğrenme veya kümeleme yöntemleri kullanılagelmiştir (5). Danışmanlı öğrenme kapsamında ele alınan biyobelirteç belirleme, özellik (*feature*) seçimi ve sınıflama yöntemleri üzerine ise ancak son yıllarda kapsamlı araştırmalar söz konusu olmuştur (1).

Veri madenciliği yöntemleriyle biyobelirteç keşfinin temel amaçları, yeni örneklerin doğru sınıflandırılmasında kullanılabilecek az sayıda değişken yani potansiyel biyobelirteçleri içeren küçük bir alt grubun belirlenmesi yoluyla klinik uygulamalara kolayca uyarlanacak hızlı ve maliyet etkinlikli sınıflayıcıların elde edilmesi, elde edilen bu biyobelirteçlerle, ilgili biyolojik süreçlerin ilişkilendirilmesi ve sonuçların görselleştirilmesi olarak sıralanabilir (1).

Son yıllarda biyobelirteç keşfinde kullanımı ilgi gören yöntemlerden biri olan "Sosyal Ağ Analizi", düğümler ile temsil edilen varlıkların kendilerine özgü özelliklerinden daha çok, varlıklar arasındaki bağların özelliklerine odaklanarak ağ ve ağ içindeki varlık yapısını tanımlamak, varlıklar arası kolayca gözlenemeyen ilişki ve etkileşimleri ayrıntılı olarak incelemek ve elde edilen bulguları değerlendirmek için kullanılan bilimsel yöntemler bütünüdür (7). Böylece hem varlıkların işlevleri hem de bağlanış şekilleri dikkate alınmaktadır (8). Bir sosyal ağ analizinde karmaşık algoritmalar, ileri istatistiksel yöntemler ve veri madenciliği yöntemleri çoğu zaman bir arada kullanılmakta ve bu nedenle çok disiplinli bir yapıya sahip olan sosyal ağ

analizi çoğu zaman farklı bilim dallarından uzmanların birlikte çalışmasını gerektirmektedir (9). Özellikle büyük veri tabanlarında gizli bilgi ve örüntüleri bulma süreci olarak tanımlanabilecek olan veri madenciliğinin sosyal ağ analizi ile birlikte kullanılmasıyla bir artı güç (sinerji) yaratılmaktadır (10).

Son yıllarda, sosyal ağ analizinin hasta izlemi, biyolojik ve genetik ağlar, hastalık biyobelirteç belirleme gibi sağlık alanındaki çeşitli uygulamalarda kullanıldığı görülmektedir (11). Gen ifade düzeylerine ilişkin verilerin analizinde kullanılan bir sosyal ağ analizi genlerin birbirlerine nasıl bağlandığını ve genler arası etkileşimleri ölçen bir araç olarak tanımlanabilir (12). Alanyazında sosyal ağ analizinin genetik alanındaki kullanımına ilişkin az sayıda kaynak olsa da, son yıllarda bu konu birçok araştırmacının ilgisini çekmekte ve bu alanda farklı bilim dallarındaki araştırmacılar dikkat çekici çalışmalar ortaya koymaktadır.

1.2. Genetik Verilerde Sosyal Ağ Analizinin Kullanımı

İlişkisel veri tabanlarından veya büyük veri yığınlarından elde edilen karmaşık sonuçları görselleştirme olanağı nedeniyle sosyal ağ analizinin tıp ve genetik alanında da kullanımı giderek yaygınlaşmaktadır. Özellikle, biyolojik sistemler gibi çok bileşenli etkileşimlerin söz konusu olduğu karmaşık sistemlerin görsel olarak modellenmesi için araştırmacılara farklı bir yol sunmaktadır. Örneğin, metabolik ağların, hücre-hücre ya da protein-protein etkileşimlerinin, genomik ortak-ifade ve gen düzenlenme ağlarının analizi için kullanılabilmektedir (13-15).

Aslında, bir genomik ortak-ifade ağını sosyal ağ analizi kapsamında oluşturmak oldukça basittir. Böyle bir ağda, düğümler genleri temsil eder ve çoğunlukla, bir gen çifti arasındaki ortak ifadelenmenin derecesi iki gen arasındaki etkileşimi tanımlar. Bu tür ağlarda, genler arası etkileşimi tanımlamak için kullanılan ölçüler çalışmadan çalışmaya farklılık gösterse de (11, 14, 16, 17) bu amaçla sıklıkla benzerlik ölçülerinden yararlanılır (18).

Genomik ortak-ifade ağları gibi karmaşık ağları görselleştirmedeki başarısı nedeniyle kullanımı daha yaygın olsa da, yüksek boyutlu verilerde boyut indirgeme amacıyla kullanıldığı çalışmalar da alanyazında yer almaktadır (15). Özellikle son yıllarda, hastalığa özgü biyobelirteçlerin belirlenmesi amacıyla yapılan çalışmalarda sosyal ağ analizinin kullanımı ilgi uyandırmakta ve hastalığa özgü biyobelirteçlerin

belirlenmesi problemine sosyal ağ analizinin de yer aldığı melez yaklaşımlarla farklı bir bakış açısı getirilmektedir (14). Bu yaklaşımlar ile sadece sosyal ağ analizi değil veri madenciliği ve makine öğrenmesi kapsamında yer alan birden fazla yöntem bütünleşik olarak çalıştırılarak biyobelirteçlerin belirlenmesi ve sonuç olarak sınıflamada başarım arttırımı amaçlanmaktadır.

1.3. Alanyazına Katkı

Son on yılda, bir sınıflayıcının sınıflama başarımını arttırabilen hastalığa özgü biyobelirteçlerin belirlenmesi için birçok melez yöntem geliştirilmiştir (1, 14, 19). Özellikle, küresel çapta pek çok kişinin ölümüne neden olan kanser türleri de dahil olmak üzere çeşitli hastalıklara özgü biyobelirteçlerin belirlenmesi problemine sosyal ağ analizinin de yer aldığı melez yaklaşımlarla daha iyi çözümler sunma arayışı devam etmektedir (14).

Ancak, melez yöntemler kapsamında kullanılabilecek birden çok makine öğrenmesi, veri madenciliği ve sosyal ağ analizi yönteminin söz konusu olması nedeniyle problemin çözümü için bu yöntemlerin uygun biçimde birleştirilmesinin ve probleme özgü en uygun (optimal) çözüm arayışının önemi ortaya çıkmıştır (14, 19).

Bu tez çalışmasında, genomik bir ağın analizi için Özyer ve diğ. (14)'nin Java programlama dili kullanarak geliştirdiği ve temelde melez bir özellik seçim yöntemi olan SNFS melez biyobelirteç belirleme yönteminin boyut indirgeme, kümeleme, topluluk belirleme gibi farklı aşamalarında kullanılan makine öğrenmesi, veri madenciliği ve sosyal ağ analizi yöntemlerinin kısaca incelenmesi; melez yöntemin her adımının tez çalışması kapsamında kullanılması planlanan erişime açık genomik mikrodizi veri setleri kullanılarak R yazılımında (20) uygulanması ve SNFS'nin adımlarında tez kapsamında kullanılan yöntemlerin farklı kombinasyonlarının biyobelirteç belirleme ile sınıflama problemi açısından karşılaştırılması amaçlanmıştır. Aynı zamanda, bir benzetim çalışması yapılarak SNFS'nin adımlarında tez kapsamında kullanılan yöntemlerin hangisinin belirli özelliklere sahip (ilişkili, gürültülü, az sayıda ilgili özellik içeren, yüksek boyutlu, vb.) mikrodiziler için en uygun çözümü sağladığı araştırılacaktır.

SNFS'nin ikiden fazla sınıf içeren genomik mikrodizi verileri için de genelleştirilmesi olanaklıdır. Ancak, bu tez çalışması kapsamında yalnızca iki sınıf içeren veriler için değerlendirilme yapılması planlanmıştır.

1.4. Tezin Planı

Tez çalışmasının amacı doğrultusunda, tezin ilk bölümünde kısaca genomik mikrodizi verilerinin analizinde klasik yöntemler dışında kullanılan yöntemlere ilişkin bilgi verilmiş; daha sonra sırasıyla biyobelirteç belirleme probleminde alanyazında yer alan bazı farklı yöntemlere değinilmiş; ardından bu amaçla sosyal ağ analizinin kullanımı ile ilgili alanyazında yer alan çalışmalar ve sosyal ağ analizinin genomik mikrodizi verilerinde kullanımı hakkında özet bilgi verilerek bu tez çalışmasının alanyazına yapacağı olası katkıdan söz edilmiştir.

Tezin ikinci bölümünde; SNFS melez özellik seçim yönteminin anlaşılabilmesi ve uygulanabilmesi için gereksinim duyulan genel bilgiler kısaca aktarılmış, alanyazında kansere özgü biyobelirteç belirleme problemini inceleyen bazı çalışmalara ilişkin kısa bilgiler verilerek konuya ilişkin bilinirlik sağlanması amaçlanmıştır.

Üçüncü bölümde; tez çalışması kapsamında kullanılan veri madenciliği, makine öğrenmesi ve sosyal ağ analizi yöntemlerine ilişkin özet bilgiler verilmiştir. Melez yöntemin uygulanması için kullanılan alanyazından elde edilmiş gerçek ve yapay veri setleri ile birlikte tezin benzetim çalışması aşamasında üretilmiş olan veri setlerine ilişkin detaylı açıklamalar da bu bölümde yer almıştır.

Dördüncü bölümde; uygulama ve benzetim çalışmasından elde edilen sonuçlar paylaşılarak ayrıntılı şekilde yorumlanmış, beşinci bölümde ise elde edilen bulgular alanyazındaki bilgilerle karşılaştırmalı olarak irdelenmiştir.

Altıncı ve son bölümde ise tez çalışmasının ulaştığı sonuçtan kısaca söz edilerek araştırmacılara melez yaklaşım ile ilgili önerilerde bulunulmuş ve aynı zamanda çalışmanın sınırlılıkları ile genişletilebilirliği üzerinde durulmuştur.

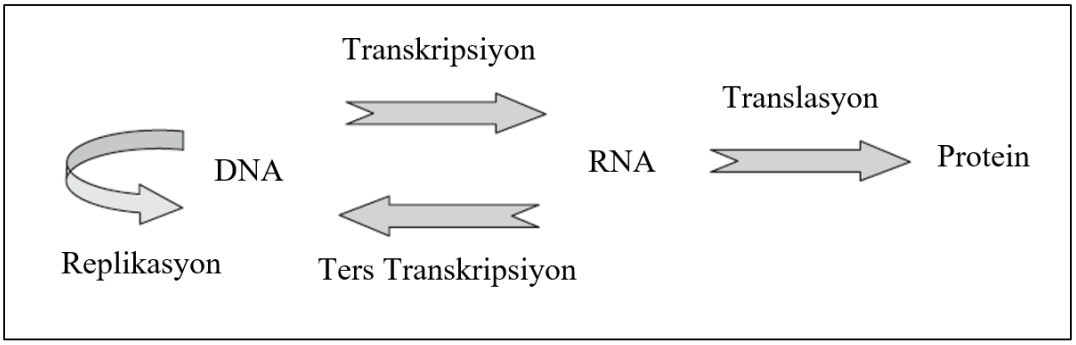
2. GENEL BİLGİLER

2.1. Gen, Genetik Bilgi, Gen İfadesi ve Gen İfade Analiz Yöntemleri

Fonksiyonlarını sürdürmek için gerekli olan tüm genetik bilgiyi depolamak, kullanmak ve bir sonraki kuşağa aktarmak canlı organizmaların ortak özelliğidir (21). Dolayısıyla "Canlı"; Deoksiribo Nükleik Asit (DNA) ve/veya Ribo Nükleik Asit (RNA) içeren, belirli koşullarda benzerini üreterek çoğalan, enerji tüketen ve üreten varlık olarak tanımlanabilir (22).

DNA'nın histon proteinleri etrafına sarılmasıyla yoğunlaşarak oluşturduğu kromozomlar üzerinde taşınan ve bir organizmanın özelliklerini belirleyen genler, kalıtımın en temel birimleridir ve hücre çekirdeğindeki kromozomlarda intron ile ekzonlardan (kodlama yapan ve yapmayan nükleotid dizilimi) oluşan DNA parçası olarak tanımlanmaktadırlar (6, 23).

Genetik bilginin akışı DNA'dan RNA'ya (Transkripsiyon); RNA'dan proteine (Translasyon) doğru gerçekleşir (Şekil 2.1.). Canlı doku ve hücrelerinin yapısal ve yaşamsal bileşeni olan proteinlerin üretimi DNA'da kodlu genler tarafından kontrol edilir. Genlerden protein üretiminin ilk aşaması olan transkripsiyon sırasında, mRNA iplikçiği gen kodlayan DNA segmentinden kopyalanır. Translasyon aşamasında ise kopyalanmış olan bu mRNA protein üretimi için amino asit zincirinin bir araya getirilmesinde taslak olarak kullanılır (6, 24). Retrovirüsler dışında tüm canlılar için santral dogma olarak adlandırılan bu mekanizma söz konusudur (21).



Şekil 2.1. Ardışık bilgi akışı: DNA'dan DNA'ya (replikasyon), DNA'dan RNA'ya (transkripsiyon), RNA'dan proteine (translasyon) ve RNA'dan DNA'ya (ters transkripsiyon) (1).

Gen ifadesi, genlerin fonksiyonel protein veya RNA yapılarına dönüşmesi sürecidir (6). Yüksek ökaryotlarda biyolojik gelişim, hücresel büyüme ve organogenez farklı gen ifadeleri ile meydana gelmektedir (25).

Günümüzde işlevsel genomik alanında RNA düzeyinde yapılan çalışmalarda, gen ifade analizi yöntemleri; proteom araştırmalarında ise kütle spektrometri dizilemesi, iki boyutlu jel elektroforezi ve protein dizisi gibi yöntemler kullanılmaktadır. RNA düzeyindeki gen ifade analizi yöntemleri genel olarak dört gruba ayrılabilir (Tablo 2.1.). Bu yöntemler kodlanan proteinin fonksiyonu hakkında bilgi sahibi olmak için tek başlarına yeterli olmasalar da farklılıkların belirlenmesini sağlayarak biyolojik süreç hakkında bilgi verici olmaktadırlar (25).

Tablo 2.1. Gen İfade Analizi Yöntemleri (25).

1- Dizileme temelli yöntemler • EST (İfadelenmiş Dizi Etiketleri - *Expressed Sequence Tags*) dizileme • SAGE (Gen İfadesinin Seri Analizi - *Serial Analysis of Gene Expression*) • MPSS (Büyük Çaplı Paralel DNA Dizileme - *Massively Parallel Signature Sequencing*)
2- Polimeraz Zincir Reaksiyonu (PZR) temelli yöntemler • Gerçek-zamanlı kantitatif PZR • DD (Ayrımsal Görüntü - *Differential Display*)
3- Melezleme (Hybridization) temelli yöntemler • DNA mikrodizileri (cDNA ya da oligonükleotid dizi) • Dot/Northern blot tekniği • Nükleaz koruma analizi • Ayırıcı plak melezlemesi • In situ melezleme
4- Melezleme ve PZR birleşiminden oluşan yöntemler • RDA (*Representational Difference Analysis*) • ESD (*"Equalization of cDNAs"* ve *"Subtractive hybridization and Differential display"* yöntemlerinin baş harfleri) • SHL (Çıkartılmış Melezleme Kütüphanesi - *Subtractive Hybridization Library*)

Farklı dokulara ait hücrelerin, hücreye özgü proteinler tarafından belirlenen farklı fonksiyonları vardır. Hangi proteinin sentezleneceği ise gen ifadesine dayalı olduğundan, genin ifadelenme örüntüsü dolaylı olarak hücre fonksiyonu hakkında bilgi sağlamaktadır (26). Çünkü gen ifadesi bir genin ilgilenilen durum için ne düzeyde aktif olduğunu gösterir. Gen ifade düzeyi ne kadar yüksekse o kadar fazla protein üretilir (6). Böylece araştırmacılar mikrodizi deneylerini çeşitli dokularda hangi

genlerin ifadelendiğini belirlemede kullanarak hücre ve genlerin fonksiyonuyla kontrol edilen mekanizmalara ilişkin değerli bilgiler edinebilmektedirler (26).

"Mikrodizi" ifadesi biyomedikal örneklerin çalışılması için özelleşmiş çip tabanlı, yüksek çıktılı teknolojiler için yaygın olarak kullanılmaktadır (4). Bir mikrodizi, genelde cam, plastik ya da silikondan oluşan bir lam ya da matriks yüzeyine DNA moleküllerinin ilgilenilen lokasyon veya spotlara tutturularak eş zamanlı on binlerce spotu incelenmesi yoluyla gen ifade düzeylerinin ölçümünü sağlayan bir çiptir (6). Basitleştirirsek, bir mikrodizi aslında işlenmiş en küçük DNA dizisini içeren spotlardan büyük sayılarda içeren küçük bir lamdır. Lamdaki spotlar dikdörtgen ızgara şeklinde düzenlenmiş olup ızgaranın her bir hücresi, bir geni temsil eden DNA materyalini (DNA, cDNA ya da oligonükleotid) içerir ve genellikle prob adını alır (1). Cam ya da silikon gibi katı bir yüzey üzerine ince uçlu iğnelerle baskı, ink-jet baskı vb. farklı yöntemler kullanılarak, çeşitli firmalar tarafından üretilen ticari mikrodiziler ve üniversitelerin kendi laboratuvarlarında ürettikleri çeşitli mikrodiziler bulunmaktadır (4).

Genel olarak, mikrodizi hazırlandığında hedef mRNA (araştırılmak istenen biyolojik örnekten elde edilmiş mRNA) floresan boya ile etiketlenir. Daha sonra hedef mRNA'nın solüsyonuyla mikrodizi yıkanır. mRNA molekülleri melezlenir ve mikrodiziye yapışır. Melezlenmemiş solüsyon yıkanarak uzaklaştırılır. Hedef boya ile etiketlenen her bir spot tarafından yayılan floresan sinyalini ölçmek için lazer tarayıcı kullanılır. Bir spotun sinyal yoğunluğu ilgili gene karşılık gelen mRNA'nın çokluğu ile ilişkilidir. Bu yolla, tek bir DNA mikrodizisi binlerce genin ifade düzeyine ilişkin eş zamanlı bilgi sağlayabilmektedir (1).

Çok sayıda gen ifadesinin eş zamanlı incelenebilmesi biyolojik çalışmalar için hayati bir öneme sahiptir. Çünkü çok sayıda gen ifadesinin eş zamanlı ölçülebilir olması genler arası etkileşimin kapsamlı olarak incelenebilmesi, fonksiyonu yeterince anlaşılamamış pek çok genin organizmadaki rolünün keşfi ve metabolik yolakların çeşitli koşullar altında nasıl değiştiğinin belirlenebilmesi için bir kapı açmıştır (24).

Birkaç farklı mikrodizi teknolojisi söz konusu olsa da günümüzde, cDNA ve oligonükleotid dizilerinin kullanımı daha yaygındır (24, 25). Bu çipler farklı tasarlanmalarına karşın, ikisi de melezleme temelli yöntemler olup (Bkz. Tablo 2.1.) yaygın olarak tüm-genom düzeyindeki gen ifade analizlerinde kullanılmaktadır.

Ayrıca, bu mikrodizilerden genetik bağlantı veya ilişkilendirme çalışmalarında da yararlanılmaktadır (25, 27).

2.2. Gen İfade Verileri ile Danışmanlı Makine Öğrenmesi

Makine öğrenmesi, veri madenciliği ya da ileri istatistiksel analizlerin uygulanması için gereken verilerin elde edilebilmesi amacıyla ilk olarak mikrodizilerden elde edilen görüntülerin işlenmesi gerekmektedir. Gen ifade düzeylerini etkileyecek birçok sistematik varyasyon kaynağı bulunduğundan, işlenen bu görüntülerden elde edilen verilerin normalizasyonunun yapılması gerekir. Verilere yapılan ön işleme sonrası (normalizasyon, arka plan düzeltmesi, vs.) örnekleri temsil eden n sütun ve genleri temsil eden p satırdan oluşan nxp boyutlu gen ifade veri matrisi elde edilir. Bu veri matrisi artık ileri istatistiksel analizler, veri madenciliği ya da makine öğrenmesi yöntemlerini uygulamak için hazırdır (6).

Bir makine öğrenme algoritması, bir başarım ölçüsüne göre gözlenen örneklerden (deneyimden) insan müdahalesi olmaksızın öğrenebilen algoritma olarak adlandırılmaktadır. Makine öğreniminde temel olarak iki tür öğrenme yaklaşımı söz konusudur. Bunlardan biri veri ya da çıktı hakkında öğreniciye, herhangi bir önsel bilginin verilmediği danışmansız öğrenme, diğeri ise çıktının algoritmaya önsel olarak verildiği danışmanlı öğrenmedir (28).

Alanyazında yaygın olarak yer alan regresyon ve sınıflama problemleri temelde birer danışmanlı öğrenme problemidir. Böyle problemlerde girdinin yanı sıra öğreniciyi besleyen çıktı da söz konusu olup öğrenme algoritmasının görevi girdiden çıktıya dönüşüm yapmaktır (29).

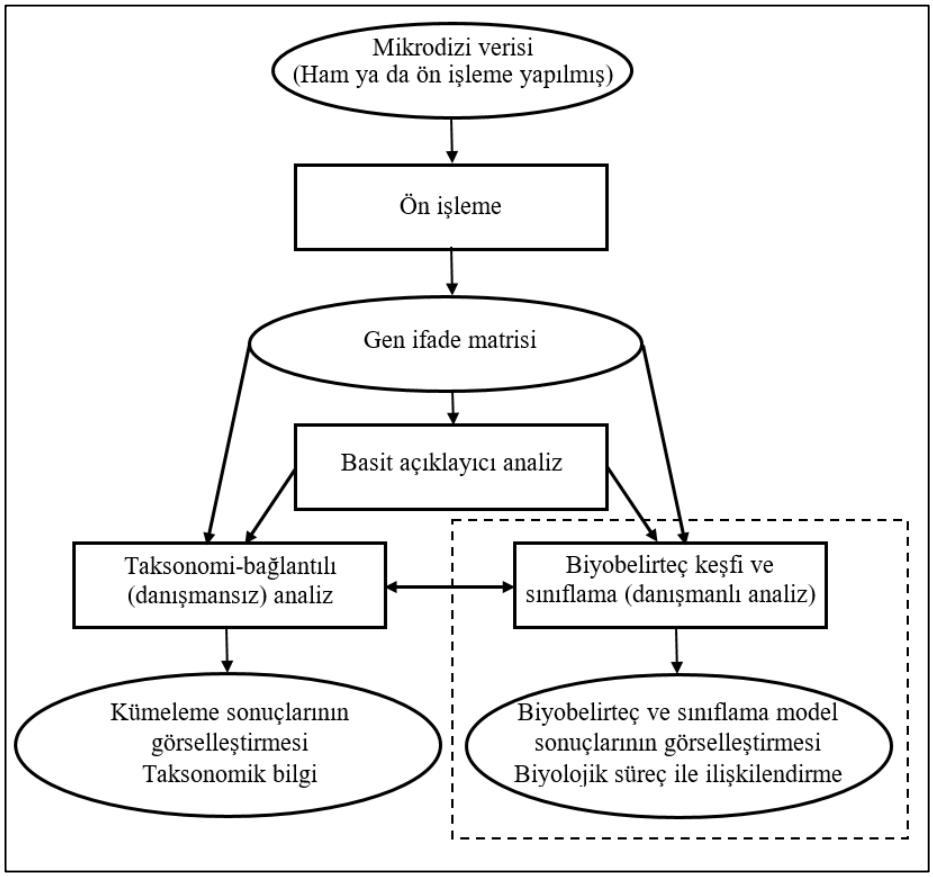
“Biyobelirteç belirleme”, bir sınıflama probleminde kullanılan öğrenme algoritmasının sınıflama başarımını arttırmak amacıyla, sınıfları anlamlı derecede ayırabilen ve sınıf üyeliklerini doğru şekilde kestirmek amacıyla kullanılabilecek en uygun değişken alt kümesini seçmek olarak tanımlanabilir. Değişkenlerin bu en uygun alt kümesine götüren algoritma ya da sezgisel sürece ise “değişken seçimi” adı verilmektedir. Biyobelirteç belirlemenin temel adımları:

1) Değişken seçimi,
2) Sınıflama modelinin oluşturulması (değişken seçimi adımına bağlı olarak ilk iki adım ayrı ya da birleşik olabilir),

3) Tercihen bağımsız bir test seti üzerinde model geçerliğinin incelenmesidir (Şekil 2.2.).

Özellikle ikinci adımı yani sınıflama modelinin oluşturulmasını içerdiğinden, biyobelirteç belirleme problemi danışmanlı öğrenme kapsamında değerlendirilir (1).

Alanyazında, moleküler kanser sınıflandırması probleminin çözümüne ilişkin gen ifade düzeyi verileri kullanılarak yapılan birçok çalışma söz konusudur (6). Biyobelirteç keşfi kapsamında da değerlendirilebilecek bu çalışmaların çoğunda, değişken seçimi için kullanılabilecek bilimsel yaklaşımlar göz ardı edilmiş ve değişkenlerin keyfi seçimi sonrası sınıflama modelleri oluşturulmuştur (1). Ancak, biyobelirteç belirleme probleminde önemli bir adım olan değişken seçimi, genellikle yüksek boyutlu verilerde boyut indirgeme ve sınıflayıcının sınıflama başarımını arttırmak amacıyla uygulandığından özellikle son on yılda yüksek boyutlu genetik verilerin söz konusu olduğu çalışmalarda ilgi görmüştür.



Şekil 2.2. Gen ifade analizi kapsamında biyobelirteç belirleme ve sınıflama (1).

Bu tez çalışması kapsamında da kullanılan Lösemi veri seti ilk olarak Golub ve diğ. (30) tarafından sınıflama amacıyla kullanılmış ve akut myeloid lösemi (AML) ve akut liphokistik lösemi (ALL) sınıflarının ayrılmasında mikrodizi verilerinin kullanılabileceği gösterilmiştir. O zamandan günümüze, mikrodizi verilerinin kullanılmasıyla doğru sınıflama üzerine odaklanarak geleneksel patolojik yaklaşımların sunabileceğinden daha doğru tanısal araçların geliştirilmesini amaçlayan çalışmalar alanyazında geniş yer bulmuştur. Bu tür mikrodizi verilerinin kullanıldığı çalışmalarda, bilinen sınıflara bireylerin atanması problemine en uygun çözümün elde edilmesi için farklı yöntemler kullanılmıştır (31).

Golub ve diğ. (30)'nin çalışmasında yer alan Lösemi veri seti, tanı konduğunda akut lösemi hastalarından alınmış 38 (27 ALL, 11 AML) kemik iliği örneğinden oluşmaktadır. RNA, kemik iliği mononükleer hücrelerinin yüksek yoğunluklu oligonükleotid mikrodizisine melezlenmesi ile Affymetrix Hu6800 Chip tarafından üretilmiştir ve 6817 insan geni probu içermektedir. Daha sonra, 24 kemik iliği ve 10 periferik kan örneğinden elde edilen 34 (20 ALL, 14 AML) örnekten oluşan bağımsız bir veri de bu veri setine eklenmiştir (30). Bu çalışmada Golub ve diğ. (30) moleküler kanser sınıflandırması problemini, kanser alt sınıflarının keşfi (danışmansız makine öğrenmesi kapsamında kümelemenin kullanılması yoluyla) ve bilinen sınıflara örneklerin atanması (danışmanlı makine öğrenmesi kapsamında sınıflama yoluyla) olmak üzere iki aşamalı olarak ele almışlardır (6). Golub ve diğ. (30)'nin alanyazına kazandırdığı Lösemi veri seti, sınıflama problemi kapsamında yaygın olarak kullanılan bir veri setidir. Örneğin, bu veri setinin doğru sınıflandırılması amacıyla alanyazında hem lojistik regresyon (31), Fisher doğrusal diskriminant analizi (32, 33), adımsal çapraz geçerlikli diskriminant analizi (34), köşegen doğrusal diskriminant analizi (35), kısmi en küçük kareler (36) gibi klasik istatistiksel yöntemler hem de destek vektör makineleri (DVM) (37, 38), *random forest* (RF) (39), Bayes ağları (40) gibi makine öğrenmesi yöntemleri kullanılmıştır.

Yüksek çıktılı gen ifade verisinin söz konusu olduğu sınıflama problemlerinin çözümü için geçmişte karar-ağaçları, doğrusal ayırma analizi, Bayes ağı, ağırlıklı oylama, vb. pek çok yöntem önerilmiştir. Ancak, yüksek boyutlu verilerin yol açtığı sorunlar nedeniyle son yıllarda araştırmacılar en yakın komşu, DVM, RF, yapay sinir ağları gibi makine öğrenmesi yöntemlerine ve özellikle, bu yöntemlerin yer aldığı

melez yaklaşımlara yönelmektedirler. Alanyazında, her ne kadar tek bir sınıflama yönteminin diğer tümü üzerindeki kesin üstünlüğü kanıtlanamamışsa da, DVM'nin ya da DVM'nin de yer aldığı gömülü veya melez yaklaşımların gen ifade verilerine uygulandığı birçok çalışmalar vardır (6).

Örneğin, Guyon ve diğ. (41) DVM ile Özyinelemeli Özellik Eleme (*Recursive Feature Elimination*, RFE) yöntemlerini birleştirdikleri Özyinelemeli Özellik Eleme-Destek Vektör Makinesi (DVM-YÖE) gömülü yöntemini Lösemi veri setine uygulayarak sınıflama başarımını arttırmayı amaçlamışlardır. DVM-YÖE ile farklı yöntemlerin karşılaştırıldığı çalışmalar da (42) alanyazında yer almaktadır. Aynı zamanda, *Random Forest*-Özyinelemeli Özellik Eleme (RF-YÖE) (43) gibi farklı makine öğrenme yöntemleriyle özellik seçim yöntemlerinin birleştirilmesi yoluyla elde edilen yöntemler de söz konusudur.

Ancak özellikle son yıllarda, hem mikrodizi verilerinin kendine özgü özellikleri nedeniyle boyut indirgemenin sınıflamanın önemli bir aşaması olarak görülmeye başlaması (44) hem de sosyal ağ analizinin karmaşık ağları görselleştirme olanağı sağlaması sebebiyle, sosyal ağ analizinin de yer aldığı çok aşamalı melez yaklaşımlar ortaya çıkmıştır. Örneğin, Özyer ve diğ. (14) sosyal ağ analizi ile değişken seçimi ve kümeleme yöntemlerinin birleştirildiği SNFS adlı melez bir yöntem önermişlerdir. Bu yöntemle Lösemi veri setindeki sınıflama başarımını arttırırken seçilen gen sayısını da azaltmayı başarmışlardır. Özyer ve diğ. (14)'nin bu çalışmasında, sosyal ağ analizini de kapsayan bu melez yaklaşım ile hastalık biyobelirteçlerinin belirlenmesi amaçlanmıştır. Bu amaçla, gerçek veri setleri üzerinde SNFS'nin DVM ve J48 (C4.5 karar ağacının Java sürümü) sınıflayıcılarının sınıflama başarımına katkısı karşılaştırılmış ancak, bir benzetim çalışması yapılmamıştır (14). Aynı zamanda, alanyazında gen ifade verileri kullanarak sosyal ağ analizi temelli geliştirilen bir sınıflayıcı da yerini almıştır (15).

Yapılan alanyazın taramasında görülmüştür ki teknolojinin gelişmesi ile birlikte araştırmacılar, özellikle sağlık alanındaki yüksek boyutlu verilerin analizi için melez yaklaşımları giderek daha çok tercih etmekte ve bu kapsamda sosyal ağ analizine de ilgi duymaktadırlar.

3. GEREÇ ve YÖNTEM

3.1. SNFS Melez Biyobelirteç Belirleme Yöntemi

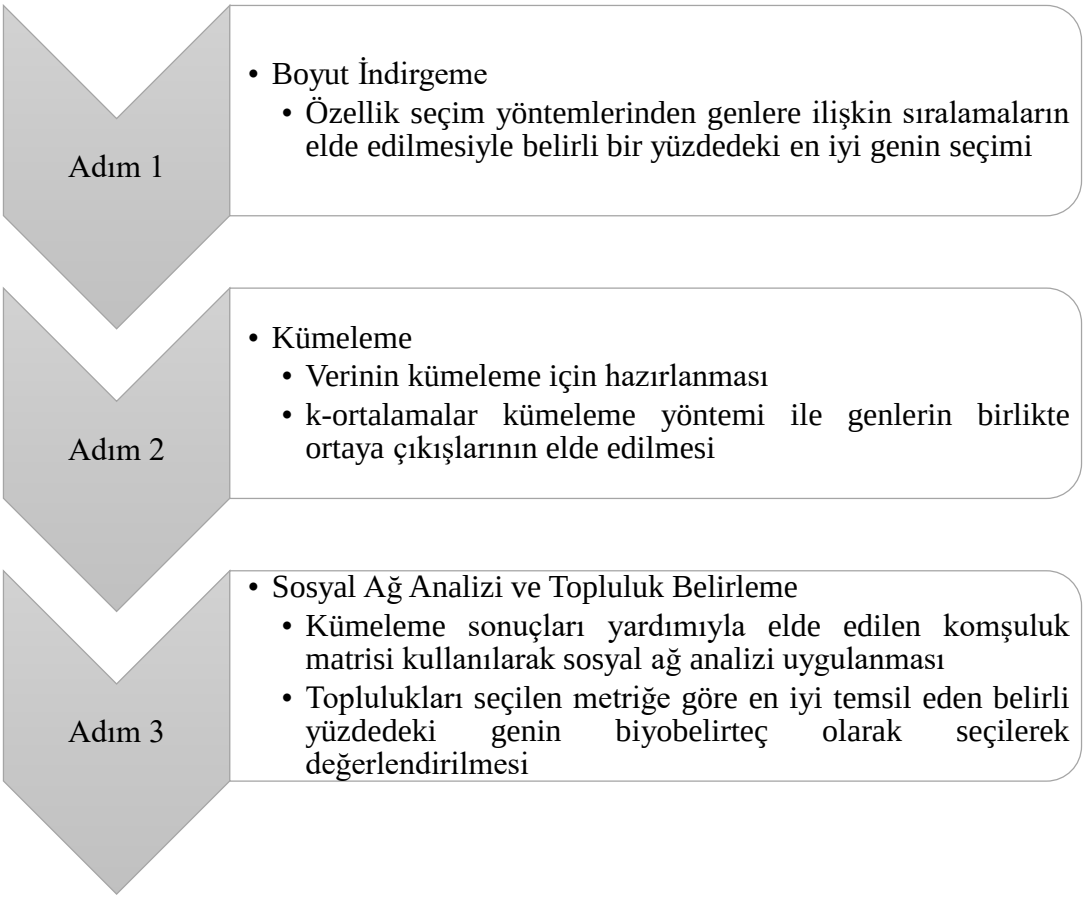
SNFS (14), yüksek boyutlu mikrodizi verilerinde gen sayısının etkin şekilde azaltılması amacıyla kullanılan ve bazı veri madenciliği, makine öğrenmesi ve sosyal ağ analizi yöntemlerinin birleşiminden oluşan melez bir özellik seçim yöntemidir. SNFS yöntemi kullanılarak örneklerin doğru sınıflandırılmasında etkili genlerin, başka bir deyişle hastalığa özgü biyobelirteçlerin seçimi sağlanarak sınıflama model başarımı arttırılabilir.

Melez biyobelirteç belirleme yöntemi SNFS;

1) Boyut indirgeme kapsamında özellik seçim yöntemlerinden elde edilen sıralamaların birleştirilmesi ve bu birleşik gen sıralamasına göre belirli bir yüzdedeki en iyi genin seçilmesi,
2) Birinci adımda elde edilen indirgenmiş gen kümesindeki her bir gen için, sınıflara göre gen ifade düzeylerinin ortalamalarının elde edilmesi ve genlerin tekrarlı olarak kümelenmesiyle aynı kümede birlikte ortaya çıkış sayılarının hesaplanması,
3) İkinci adımda elde edilen ağırlıklı komşuluk matrisi kullanılarak sosyal ağ analizinin uygulanması ve topluluk belirleme yöntemi ile ağdaki toplulukların elde edilmesi; ağa özel metriklere göre her topluluğu en iyi şekilde temsil edeceği düşünülen belirli bir yüzdedeki genin biyobelirteç olarak adlandırılması; bağımsız bir test seti üzerinde bu biyobelirteçler kullanılarak uygulanan sınıflayıcıdan elde edilen sınıflama başarımının değerlendirilmesi,

olmak üzere üç ana aşamadan oluşmaktadır (Şekil 3.1.). Üçüncü adımın sonunda yöntemin geçerliği ve genlerin biyobelirteç olarak potansiyelleri ortaya konulabilir.

SNFS melez biyobelirteç belirleme yöntemi birden fazla yöntemin birleşiminden oluştuğundan, SNFS'nin farklı adımlarında seçilebilecek değişik yöntemler aynı veri seti için sınıflama başarımını değiştirebilir.



Şekil 3.1. SNFS analiz adımları.

Bu nedenle, SNFS'nin adımlarında kullanılabilecek farklı yöntemlerin incelenmesi ve farklı kombinasyonların denenerek en uygun çözümün aranması önem kazanmaktadır.

3.1.1. Boyut İndirgeme Aşaması ve Özellik Seçimi

Yüksek boyutluluk terimi;

a) örnek sayısı çok yüksek,

b) özellik sayısı çok yüksek,

c) hem örnek hem özellik sayısı çok yüksek olan veriler için kullanılan bir terim olmakla birlikte, alanyazında çoğunlukla özellik sayısı çok yüksek olan verilerin tanımlanmasında kullanılmaktadır (45). DNA mikrodizileri ile elde edilen gen ifade verileri, yüksek boyutlu verilerdir ve genellikle çok sayıda özellik ile oldukça az sayıda örnek içermektedirler (30).

Bu tür veriler için sınıflama probleminin, yüksek boyutluluğun neden olduğu sıkıntılar nedeniyle, çözümü zor bir makine öğrenmesi problemi haline geldiği alanyazında ifade edilmiştir. Ayrıca, gen ifadesi ölçülmüş genlerin aslında pek çoğunun da doğru sınıflama yapmak için ilgili/gerekli özellikler olmadığı pek çok çalışmada gösterilmiştir (30).

Gen ifade verileri gibi yüksek boyutlu verilerin makine öğrenmesi yöntemleriyle analizi sırasında ortaya çıkan boyutsallık sorunu ve aşırı uyum gibi problemleri aşmak için alanyazında farklı boyut indirgeme yöntemleri geliştirilmiş ve öğrenme başarımının arttırılması sağlanmıştır (19).

Boyut indirgeme yöntemleri, özellik seçimi ve özellik çıkarımı (*feature extraction*) olmak üzere iki ana sınıfta toplanmaktadır. Her iki yöntem sınıfının da kendine özgü avantajları söz konusu olsa da özellik seçim yöntemleri, orijinal özellikleri koruyarak ilgisiz (*irrelevant*) ya da gereksiz (*redundant*) özellikleri kaldırma yoluyla boyut indirgediklerinden, orijinal özelliklerin modelin yorumlanması ve bilgi çıkarımı için önemli olduğu uygulamalarda daha fazla kullanılmaktadır (19). Özellik çıkarma yöntemlerinde ise girdi olarak veride var olan özellikler kullanılarak boyut indirgeme amacıyla yeni özellikler oluşturulmaktadır. Dolayısıyla, biyobelirteç belirleme problemi söz konusu olduğunda alanyazında özellik çıkarma yöntemlerinin kullanımı tercih edilen bir yaklaşım olmamış ve yaygın olarak özellik seçim yöntemleri üzerine çalışmalar yürütülmüştür (19).

Tipik bir DNA mikrodizi çalışmasında, gen ifade düzeylerine ilişkin (kalite kontrol ve ön işleme yapılmış) veri seti 5000 ile 20000 arasında gen (özellik) içerebilmektedir. Bu kadar fazla sayıdaki gen arasından, en uygun gen alt kümesinin bulunmasını sağlayacak kapsamlı bir arama yapmak olanaklı olmayabilir. Alanyazında özellik seçimi ile bağlantılı problemlerin genel olarak NP-hard (*Non-deterministic polynomial-time hard*) olduğu gösterilmiştir (46). Başka bir deyişle, özellik seçimi ile bağlantılı problemlere ait olası tüm örneklerin, polinomiyal zamanda en uygun çözümünün elde edilmesini sağlayan bilinen genel bir algoritma olmadığı kabul edilmektedir. Bu nedenle, Bellman'dan (47) bu yana alanyazındaki pek çok çalışmada, özellik seçim adımı göz ardı edilerek keyfi özellik seçimi sonrasında sınıflama modelinin oluşturulması tercih edilmiştir (1). Oysa bu adım, öğrenme algoritmasının başarımı için büyük önem taşımaktadır. Çünkü uygulamada karşılaşılan

yüksek boyutsallığın varlığı, verideki ilişkili/ilgili özelliklerin uygun şekilde tanımlanmasını bir zorunluluk haline getirmiştir. Yüksek boyutlu verilerin yol açtığı sorunlardan kaçınmak için özellik seçimi, özellikle DNA mikrodizi analizinde hayati bir rol oynamaktadır (48).

Yüksek boyutlu veriler söz konusu olduğunda makine öğrenmesi problemi karmaşık hale gelmektedir. Örneğin, öğrenme algoritmasının başarımı aşırı uyum nedeniyle bozulabilir, hesaplama hızı ve etkinliği zayıflayabilir, öğrenilen model daha karmaşık olacağından yorumlanması zorlaşabilir. Yüksek boyutsallığın yarattığı bu gibi sorunların üstesinden boyut indirgeme yoluyla gelmek mümkün olabilmektedir. Bu bakımdan, özellik seçimi en az bozulmayla ilgili problemi tanımlayan ve hatta başarımı iyileştiren ilişkili/ilgili özelliklerin belirlenmesi ve ilişkisiz/ilgisiz olanların veriden atılması süreci olarak tanımlanabilir. Veri ön işlemedeki önemli bir adım olarak özellik seçimi, son yıllarda özellikle sınıflama problemleri ile ilişkili uygulamalarda önemli ve aktif bir araştırma alanı haline gelmiştir. Ancak, alanyazında belirtilen bazı sıkıntılar nedeniyle tüm genomik mikrodizi verileri için bir en uygun özellik seçim yönteminin söz konusu olmadığı, sadece probleme özgü yaklaşımların söz konusu olabileceği ifade edilmektedir. Bu sıkıntılardan biri, son gelişmeler ışığındaki en uygun sonuçların (*state-of-the-art results*) eksikliğidir. Bir diğeri, alanyazında oldukça geniş bir açık erişimli mikrodizi veri seti grubunun söz konusu olması ve hatta bazılarının aynı isimle isimlendirildiği halde örnek sayılarının ya da özelliklerinin çalışmadan çalışmaya farklılaşmasıdır. Bu tür sıkıntılar, özellik seçim yöntemleri ile ilgili karşılaştırmalı olarak karar vermeyi daha da karışık bir duruma getirmektedir. Alanyazındaki probleme özgü yaklaşımların ötesinde bir en uygun yöntemin olmadığı baskın kanısına rağmen, genomik mikrodizi verilerinin yapay olarak üretilmesi yoluyla çeşitli benzetim çalışmaları yapılarak belirli özellikteki veri setleri için en uygun yöntem arayışları devam etmektedir (19).

Özellik seçim yöntemleri genellikle "Filtreler (*Filters*)", "Sarmal Yöntemler (*Wrapper Methods*)" ve "Gömülü Yöntemler (*Embedded Methods*)" olmak üzere üç alt gruba ayrılmaktadır (49). Filtreler, bir tür veri ön işleme adımı gibi eğitim setinin genel özelliklerine dayalı olarak tümevarım (öğrenme) algoritmasından bağımsız şekilde özellik seçimini gerçekleştirir. Buna karşın, sarmal yöntemler seçim sürecinin bir parçası olarak tahmin ediciyi en iyilerler (optimize ederler). Bu iki yöntemin

arasında ise gömülü yöntemler genellikle verilen öğrenme algoritmalarına özgündürler ve öğrenme sürecinin içinde özellik seçimini gerçekleştirirler (45).

Küçük ölçekli öğrenme problemleri yakınsama sorunu ile ilgiliyken büyük ölçekli öğrenme problemleri daha karmaşıktır. Çünkü yalnızca özellik seçiminin doğruluğunu değil stabilite (başka bir deyişle eğitim seti değişimlerine hassas sonuçlar) ya da ölçeklenebilirlik gibi diğer açılarını da kapsar. Bu nedenle alanyazında var olan özellik seçim yöntemlerinin içinden doğru yöntemin uygulanabilmesi için araştırmacı hem özellik seçimi konusunda bilgili olmalı hem de kullanılabilir yöntemlerin teknik detaylarını anlayabilmelidir. Bunun ötesinde pek çok yöntem küçük ölçekli öğrenme problemleri için geliştirilmiştir ancak günümüzde büyük ölçekli öğrenme problemlerinin söz konusu olduğu durumlar için de farklı yaklaşımlar gerekmektedir (19).

Filtreler: Hesaplamada kolaylık ve iyi genelleştirme özelliği nedeniyle avantajlı olan filtreler, özelliklerin indirgenmesi sonrası kullanılacak olan tahmin edicilerden geri bildirim almaksızın doğrudan veriden başarım değerlendirme metriklerinin hesaplanmasına dayanmaktadırlar. Bunlar, yalnızca verinin yapısal özelliklerinin (örneğin istatistiksel ölçütler) gözlenmesi yoluyla gen alt kümelerinin iyiliğini değerlendirirler ki genellikle tek bir gen ya da bir gen kümesini sınıf etiketine karşı değerlendirirler (45).

Mikrodizi verilerine de uygulanan korelasyon temelli özellik seçimi (CBFS), hızlı korelasyon temelli filtre (FCBF), ReliefF, tutarlılık temelli filtre, minimum gereksizlik maksimum ilgililik (*minimum Redundancy Maximum Relevance*, mRMR), gibi klasik filtreler (Tablo 3.1.) dışında da pek çok filtre vardır (Tablo 3.2.) (19).

Tablo 3.1. Klasik filtreler ve özellikleri (19).

Filtre	Tek/Çok Değişkenli	Sıralayan/Alt Küme Seçici
Ki-Kare (*Chi-Square*, CS)	Tek Değişkenli	Sıralayan
Bilgi Kazancı (*Information Gain*, IG)	Tek Değişkenli	Sıralayan
ReliefF	Çok Değişkenli	Sıralayan
mRMR	Çok Değişkenli	Sıralayan
$\mathcal{M}_d$	Çok Değişkenli	Sıralayan
Korelasyon Temelli Özellik Seçimi	Çok Değişkenli	Alt Küme Seçici
FCBF	Çok Değişkenli	Alt Küme Seçici
INTERACT	Çok Değişkenli	Alt Küme Seçici
Tutarlılık Temelli Filtre	Çok Değişkenli	Alt Küme Seçici

Filtreleri, sıralayan (*ranker*) ve alt küme seçici (*subset*); tek ve çok değişkenli olmak üzere farklı alt gruplara ayırmak söz konusu olabilir (45). Sıralayan filtreler, her özelliği skorlamak için bir ölçüt kullanan ve sonucunda bir sıralama üreten özellik seçim yöntemleridir. Alt küme seçici filtreler, sadece özelliklerin bir alt kümesini herhangi bir sıralama vermeden seçen yöntemlerdir (50). Tek değişkenli filtreler her bir özelliği diğer özelliklerden bağımsız değerlendirir. Bu dezavantajın üstesinden çok değişkenli yöntemlerle gelinmektedir. Ancak, çok değişkenli filtreler için daha fazla hesaplama süresi gerekmektedir.

Tablo 3.2. Mikrodizi verilerinde kullanılan filtreler ve özellikleri (19).

Filtre	**Tek/Çok Değişkenli**	**Sıralayan/Alt Küme Seçici**
BAHSIC	Çok Değişkenli	Sıralayan
Discretizer+filter	Çok Değişkenli	Alt Küme Seçici
Entropik Filtreleme Algoritması (*Entropic Filtering Algorithm*, EFA)	Tek Değişkenli	Alt Küme Seçici
$\mathcal{M}_d$	Çok Değişkenli	Sıralayan
Çoklu-görev Filtresi (*Multi-task Filter*, M_FS)	Çok Değişkenli	Sıralayan
MASSIVE	Çok Değişkenli	Sıralayan
Maksimum ağırlık minimum gereksizlik (*Maximum weight minimum redundancy*, MWMR)	Tek Değişkenli	Alt Küme Seçici
Kısmi En Küçük Kareler (*Partial Least Squares*, PLS)	Çok Değişkenli	Sıralayan
Robust Özellik Seçimi (*Robust Feature Selection*, RFS)	Çok Değişkenli	Sıralayan
İlgililik Gereksizlik Özellik Seçimi (*Relevance Redundancy Feature Selection*, RRFS)	Çok Değişkenli	Sıralayan

SNFS yönteminin ilk adımında, klasik filtrelerden Ki-Kare (CS) ve Bilgi Kazancı (IG) filtreleri kullanılmıştır (14). Uygulama kolaylığı dışında bu filtrelerin kullanılmasındaki temel nedenlerden bir diğeri, bu aşamada genlerin indirgenebilmesi için genlere ilişkin objektif bir sıralama elde etme gerekliliğidir. Dolayısıyla bu filtreler, özellikleri sıralayarak sonucunda genlere ilişkin bir önem sıralaması vermeleri nedeniyle seçilmişlerdir. Bu adımda, sıralayan filtrelerin herhangi biri de SNFS yöntemine uyarlanabilme olasılığına sahiptir. Ancak, araştırmacılar kolay uygulanabilir olması nedeniyle CS ve IG filtrelerini yöntem kapsamına almışlardır (14).

CS filtresi (51), sınıflara göre her bir özelliği bağımsız olarak değerlendiren, χ^2 istatistiğine dayalı tek değişkenli bir özellik seçim yöntemidir. Bu filtreye göre, bir özelliğe ilişkin χ^2 değeri arttıkça özelliğin ilgililik düzeyi de artmaktadır.

CS filtresi temelde iki aşamalı uygulanır. İlk aşamada, tüm sayısal özelliklerin (*attributes*) ayrıklaştırılması için 0,5 gibi yüksek bir önemlilik düzeyi belirlenir. Her bir özellik kendi değerine göre sıralanır. Daha sonra,

1) Eşitlik 3.1.'de tanımlanan χ^2 değeri her bir komşu aralık çifti için hesaplanır.

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{(A_{ij} - E_{ij})^2}{E_{ij}} \qquad (3.1.)$$

Burada, k sınıf sayısı, A_{ij} $i.$ aralık $j.$ sınıftaki örnek sayısı, $R_i = \sum_{j=1}^{k} A_{ij}$ $i.$ aralıktaki örnek sayısı, $C_j = \sum_{i=1}^{2} A_{ij}$ $j.$ sınıftaki örüntü sayısı, N= $\sum_{i=1}^{2} R_i$ toplam örnek sayısı, E_{ij}=$R_i * C_j / N$ olmak üzere A_{ij}'nin beklenen sıklığı ve serbestlik derecesi sınıf sayısının bir eksiği olarak tanımlanabilir. Eğer R_i ya da C_j 0 ise E_{ij} 0,1 olarak ele alınır.

2) Bu adımda, en düşük χ^2 değerine sahip komşu aralık çifti birleştirilir. Birleştirme işlemi, önemlilik düzeyi ile belirlenmiş χ^2 değeri tüm aralık çiftleri tarafından aşılıncaya kadar devam eder. Bu süreç, ayrıklaştırılmış veride δ tutarsızlık oranı aşılıncaya kadar azalan önemlilik düzeyleri ile tekrar edilir. Burada "tutarsızlık" kavramı, aynı olan iki örneğin farklı sınıflara sınıflanmış olması olarak tanımlanabilir.

İlk aşamada belirlenen önemlilik düzeyi$_0$ ile başlamak üzere her i özelliği, önemlilik düzeyi$_i$ dikkate alınarak birleştirme işlemi için ele alınır. Eğer tutarsızlık oranı aşılmazsa önemlilik düzeyi$_i$, i özelliğinin bir sonraki birleştirme sırası için azaltılır; aksi durumda i özelliği ileride yapılacak olan birleştirme için değerlendirmeye alınmaz. Bu işlem, hiçbir özelliğin değeri birleştirilemeyinceye dek devam eder. İkinci aşamanın sonunda eğer ilgili özellik sadece tek bir değere birleştirilirse, basitçe bu özelliğin orijinal veri setini temsil etmede ilgili olmadığı söylenebilir. Sonuç olarak ayrıklaştırma işlemi bittiğinde özellik seçimi de başarılmış olur (51).

IG filtresi (52), yaygın olarak kullanılan entropi temelli bir tek değişkenli filtredir ve özellikleri tek tek bilgi kazançlarına göre değerlendirir. Bilgi kazancı ilgili sınıf ve özellik entropilerinin toplamından ilgili sınıf ile özelliğin bileşik entropisinin

çıkartılması yoluyla hesaplanır. Böylece filtre, tüm özellikler için entropi temelli bir sıralamanın elde edilmesini sağlar. Daha sonra, elde edilmiş olan bu sıralamalara göre belli sayıda örnek seçebilmek için bir kesim noktası belirlenir ve özellik seçimi gerçekleştirilir.

Sarmal Yöntemler: Kendi iç yapısında barındırdığı öğrenme algoritmasının kestirim başarımını özellik alt kümelerini değerlendirmek için kullanan sarmal yöntemler, temelde her bir aday özellik alt kümesini değerlendirmek için öğrenme algoritmasını çağıran; bir başka deyişle öğrenme algoritmasını bir altyordam olarak çalıştıran özellik seçim yöntemi alt sınıfıdır.

Bu yinelemeli yaklaşım, filtrelere oranla daha iyi bir sınıflama başarımı sağlama eğiliminde olsa da her aday özellik alt kümesi değerlendirildiğinden oldukça zaman alıcı olabilir. Özellikle, özellik sayısı arttıkça özellik alt küme uzayı üstel olarak büyüdüğünden on binlerce özelliğe sahip verilerin değerlendirilmesinde bu durum oldukça kritik bir hal alır. Aynı zamanda, mikrodizi verilerindeki küçük örnek genişliklerine bağlı olarak aşırı uyum riski de söz konusudur. Bunun sonucu olarak sarmal yöntemler, filtreler gibi alanyazında geniş ilgi uyandıramamış ve yüksek boyutlu verilerin söz konusu olduğu çalışmalarda sıklıkla göz ardı edilmiştir. Bu tez çalışmasında, hem sarmal yöntemlerin dezavantajları hem de alanyazında ilgili melez yöntem kapsamında kullanılmamış bir özellik seçim alt sınıfı olması nedeniyle, sarmal yöntemler tez çalışması dışında bırakılmıştır.

Gömülü Yöntemler: Daha az zaman tüketimine rağmen filtrelemenin ana dezavantajı sınıflayıcı ile iletişimde olmamasıdır. Bu durum genellikle sarmal yöntemlerden daha kötü başarım göstermelerine neden olur. Ancak, mikrodizi verilerinin yüksek boyutluluğu nedeniyle sarmal yöntemlerde ortaya çıkan sorunlar da bu yöntemleri ideal olmaktan uzaklaştırmaktadır. Bu nedenle, araştırmacılar için çözüm, gömülü yöntemlerin kullanılmasıdır (41, 53).

Özellik seçimini öğrenme algoritmasının eğitimi sürecinde gerçekleştiren ve genellikle, belirli öğrenme algoritmalarına özgü olan gömülü yöntemlerde, en uygun özellik alt kümesinin aranması sınıflayıcının içinde yapılandırılır. Filtreler ve sarmal yöntemlerin aksine, gömülü yöntemler öğrenme sürecini özellik seçim sürecinden ayırmaz (53).

Alanyazında farklı gömülü yöntemler olmasına (Tablo 3.3.) rağmen belki de en ünlü gömülü yöntem Guyon ve diğ. (41) tarafından, kanser sınıflaması probleminde gen seçimi için geliştirilmiş olan DVM-YÖE'dir. DVM-YÖE özellik seçimini, güncel özellik kümesi ile bir DVM sınıflayıcısını yinelemeli olarak eğitip her tekrarda DVM ile belirlenen en önemsiz özelliği eleme yoluyla gerçekleştirmektedir (41). Bu yöntem gen seçiminde standart yöntemler arasında yerini almış, bu nedenle çeşitli eklenti ve modifikasyonları önerilmiştir (41, 53).

Alanyazından bilindiği üzere DVM, hem örnek sayısı hem de özellik sayısı açısından çok büyük ölçekli sınıflama problemlerinin çözümünde diğer yöntemlere göre genellikle daha üstün bir sınıflama başarımına ulaşan etkin bir makine öğrenmesi yöntemidir. DVM'ler iki adımda oluşturulur. İlk adımda, veri vektörleri çok boyutlu uzayda haritalanır. İkinci adımda, DVM bu uzayda sınıfları ayıran en büyük marjlı bir hiper-düzlem bulmaya çalışır. Burada "marj (*margin*)" kavramı, sınırdan özellik uzayındaki en yakın veri noktasına dek olan uzaklığı tanımlamak için kullanılmaktadır. Bazen çok yüksek boyutlu uzayda bile ayırım yapabilen bir hiper-düzlem bulmak mümkün olmaz. Bu durumda, marj içerisinde olmak şartıyla her bir vektör için marj ve ceza (*penalty*) arasında bir denge durumu (*trade-off*) tanımlanır. En basit haliyle yani doğrusal formda bir DVM, en büyük marj ile örnekleri iki sınıfa ayıran bir hiper-düzlemdir. DVM, bu işlem sonucunda w ağırlık vektörünü üretir (53).

DVM-YÖE gömülü metodunda, DVM'den elde edilen bu ağırlık vektörü en önemsiz özelliğin belirlenmesinde kullanılır. Temel olarak bir geriye doğru eleme yöntemi olarak tanımlanabilecek bu yaklaşımda, güncel özellik listesi kullanılarak DVM tekrarlı olarak eğitilir ve en az önemli olan özellik, ilgili tekrarda elde edilen ağırlık vektörüne göre belirlenerek, veri setinden kaldırılır. Bir başka deyişle, ağırlık vektörünün karesi cinsinden en küçük değere sahip özellik sonuç hiper-düzlemine en az katkı sağlayan özelliktir ve veri setinden atılır (53).

Alanyazında gömülü yöntemlerin sarmal yöntemlere oranla daha az hesaplama süresi gerektirdiğinden bahsedilmiştir (19). Ancak, yüksek boyutlu verilerin analizinde DVM-YÖE yöntemi fazla hesaplama zamanı isteyen bir yapıya sahiptir. Bu nedenle, yöntemin başarımını geliştirmek amacıyla en az önemli tek bir özellik yerine; her tekrarda belirli sabit ya da değişken sayıda en az önemli özelliğin elenmesi (41,

53) veya özellik sıralamasının farklı ölçütlere göre yapılması (54) gibi özyinelemeli özellik eleme adımı için farklı varyasyonlar önerilmiştir.

Tablo 3.3. Gömülü yöntemler ve özellikleri (19).

Gömülü Yöntem	Tek/Çok Değişkenli	Sıralayan/Alt Küme Seçici
DVM-YÖE	Çok Değişkenli	Alt Küme Seçici
RF	Çok Değişkenli	Alt Küme Seçici
FRFS (*Fuzzy Rough Set Feature Selection*)	Çok Değişkenli	Alt Küme Seçici
IFP (*Iterative Perturbation Method*)	Tek Değişkenli	Alt Küme Seçici
Çekirdek Cezalandırılmış DVM	Çok Değişkenli	Alt Küme Seçici
Özellik Seçimi – Perseptron (*Feature Selection – Perceptron*, FS-P)	Çok Değişkenli	Sıralayan
PAC-Bayes (*Probably Approximately Correct*-Bayes)	Tek Değişkenli	Sıralayan

Özyer ve diğ. (14) SNFS yöntemi kapsamında CS ve IG filtrelerinin yanı sıra DVM-YÖE gömülü yöntemini de kullanmışlardır. Ancak, bu yöntem sıralayan bir özellik seçim yöntemi değildir. Bu nedenle, DVM'den elde edilen ağırlıklar genlerin sıralanması amacıyla kullanılmış ve IG ile DVM-YÖE yöntemlerinden elde edilen sıralamalar birleştirerek her bir özellik yani gen için tümel bir sıralama elde edilmiştir (14). Bu tez kapsamında, DVM-YÖE'nin yanı sıra DVM-YÖE'nin hesaplama zamanını azaltmak amacıyla önerilmiş modifikasyonlardan biri olan ve özyinelemeli özellik eleme adımında gereksiz özelliklerin teker teker değil, toplam özellik sayısının karekökü kadar sayıda elenmesini içeren Karekök-Yinelemeli Özellik Eleme (K-YÖE) (41) yöntemi de uygulanarak SNFS'nin ilk adımındaki hesaplama zamanında kayda değer bir azalma sağlanması amaçlanmıştır.

Filtreler, gömülü ve sarmal yöntemler dışında alanyazında özellik seçim yöntemleri kapsamında yer alan alt sınıflardan olan melez yöntemler ile topluluk (*ensemble)* yöntemleri de son yıllarda oldukça fazla ilgi uyandırmaktadır. Alanyazında yer alan pek çok çalışmada (41, 55, 56) özellik seçim sürecinin yararlarına değinilmişse de çoğu araştırmacı "en iyi yöntem" olarak adlandırılabilecek bir yöntemin söz konusu olmadığı, aksine ancak belirli bir problem için en iyi yöntem seçiminin söz konusu olduğunda hemfikirdirler. Bu nedenle, değişik stratejiler uygulayan farklı özellik seçim yöntemleri sürekli geliştirilmektedir. Alanyazında,

a) Aynı ya da farklı özellik seçim yöntemi alt sınıfında yer alan yöntemlerin, örneğin iki filtrenin (57) ya da bir filtre ile sarmal yöntemin (58, 59) veya gömülü yöntemin (60) birleştirilmesi;
b) Özellik seçim yöntemlerinin özellik çıkarımı (61), karar ağaçları (62), olabilirlik yaklaşımı (63) veya sosyal ağ analizi (14) gibi farklı yöntemlerle birleştirilmesi;
c) Var olan yöntemlerin yeniden yorumlanması (64) ya da belli problemlere uyarlanması (65);
d) Hala çözüm bulunamamış durumlar için yeni yöntemlerin geliştirilmesi (66);
e) Bir grup özellik seçim yönteminin daha iyi bir başarım için kullanımı (67, 68);
f) Yarı-danışmanlı olarak yeniden yorumlanma (69) gibi melez yaklaşımlar bulunmaktadır.

Görüldüğü gibi alanyazında oldukça fazla özellik seçim yöntemi bulunduğundan, SNFS kapsamındaki özellik seçimi aşamasında CS, IG ve DVM-YÖE'nin yanı sıra başka özellik seçim yöntemlerinin kullanılması da söz konusu olabilir. Ancak, burada önemli olan tüm özellik seçim yöntemlerinin temelde probleme özgü başarım gösterdiklerinin ifade edilmiş olmasıdır (63). Bu nedenle, benzetim çalışmaları ile karşılaştırmalı olarak başarımın değerlendirilmesi oldukça önemlidir.

SNFS'nin ilk adımı olan özellik seçim aşamasının önemli bir parçası, farklı özellik seçim yöntemlerinden elde edilen indirgenmiş özellik listelerinin birleştirilmesi ve birleşik listeden belirli sayıda özelliğin seçilmesidir. SNFS'nin bu aşamasında, bu birleşik liste kullanılarak seçilecek en önemli gen yüzdesi kullanıcı tarafından belirlenmektedir. Bu tez çalışması kapsamında, bu yüzdenin verideki özellik sayısına göre en uygun şekilde belirlenmesi ile ilgili değerlendirme yapılmamıştır.

3.1.2. Genlerin Kümelenmesi

Kümeleme, özellik ya da örneklerin birden fazla grup ya da kümeye, küme içi özellik ya da örneklerin oldukça yüksek benzerliğe; bu özellik ya da örneklerin diğer

kümelerdeki özellik ya da örnekler ile oldukça yüksek benzemezliğe sahip olacağı şekilde, gruplandırılması sürecidir. Özellik ya da örneğin benzerlik ve benzemezlik durumu sıklıkla uzaklık ya da benzerlik ölçüsü kullanılarak elde edilen değerlerine göre değerlendirilir.

Eğitim setindeki sınıf etiketinin analizinin söz konusu olduğu sınıflama ve regresyonun aksine, kümeleme özellik ya da örnekleri sınıf etiketlerine danışmadan küme içi benzerliğin maksimize edilmesi ve kümeler arası benzerliğin minimize edilmesi prensibine dayalı olarak gruplandırır. Bu nedenle, kümeleme aslında danışmansız öğrenme ile eşanlamlıdır. Öğrenme süreci danışmansızdır çünkü girdiler yani özellik ya da örnekler sınıf etiketine sahip değildir.

Alanyazında pek çok kümeleme yöntemi söz konusudur. Mikrodizilerden elde edilen gen ifade düzeylerinin analizinde kullanılmak üzere "bölünmeli (*partitioning*) kümeleme (*k-means, k-medoids*), hiyerarşik (*hierarchical*) kümeleme (AGNES, DIANA, Chameleon, ROCK, BIRCH, vb.), yoğunluk temelli (*density-based*) kümeleme (DBSCAN, OPTICS, DENCLUE), ızgara tabanlı (*grid-based*) kümeleme (STING, *Wavecluster*) ya da çok boyutlu veriler için geliştirilmiş ileri düzey olasılıksal model tabanlı (*probability model-based*) kümeleme (Bulanık (*fuzzy*) kümeler, EM, kavramsal (*conceptual*) kümeleme, yapay sinir ağları ile kümeleme kapsamında öz-düzenlemeli ağ (*self-organizing map*)), alt uzay (*sub-space*) kümeleme (aynı zamanda ızgara tabanlı olan CLIQUE, PROCLUS ya da temel bileşenler analizi gibi korelasyon temelli kümeleme, iki-yönlü/boyutlu kümeleme, vb.) ve spektral (*spectral*) kümeleme sınıflarında yer alan çok sayıda kümeleme yöntemlerinden veri setine en uygun olanı seçilebilir (70).

Bir sosyal ağ analizinde veya sosyal ağ analizini içeren melez yaklaşımlarda, sosyal ağ analizinin uygulanabilmesi için bir düğümün (genin) bir diğerine nasıl benzediği ya da benzemediğinin belirlenmesi gerekmektedir. Başka bir deyişle, düğümler (genler) arası etkileşim değerlendirilmelidir. SNFS yönteminde genler arası etkileşimlerin tanımlanması için gereken benzerlik ölçütü, tekrarlı olarak genlerin kümelenmesi yoluyla elde edilmektedir.

Daha önce bahsedildiği üzere, SNFS yönteminin ilk aşamasında özellik seçimi gerçekleştirilir. Özellik seçimi sonrası uygulanacak olan ikinci ana adım kümelemedir. Özyer ve diğ. (14) sosyal ağ analizi için gereken komşuluk matrisinin oluşturulması

amacıyla kümeleme yöntemlerinden k-ortalamalar kümeleme yönteminin tekrarlı kullanımını içeren bir yaklaşım izlemişlerdir. Bu yaklaşıma "iki-yönlü/boyutlu kümeleme yöntemi" demişlerdir. Bu adımda amaç benzer genleri gruplayabilmektir.

Ancak kümeleme öncesi, devriği alınmış gen ifade verisindeki örnekleri sınıflara göre ayırmak ve gen ifade düzeylerinin sınıflara göre ortalamalarını almak önemlidir. Elde edilen bu ortalamalar kullanıldığında, k-ortalamalar kümeleme yöntemi ile hem örnek hem de özellik boyutunun dikkate alındığı bir kümeleme yapılabilmektedir. Kümeleme yönteminin tekrarlı kullanımı sayesinde de her genin bir diğer genle ortak ortaya çıkış sayıları elde edilir. Bu şekilde hesaplanan genlerin birbirleri ile olan etkileşim düzeyleri yani birlikte ortaya çıkışları, ağırlıklı ağ analizi için kullanılmaktadır. Tez çalışmasında da SNFS yöntemine uygun olarak k-ortalamalar kümeleme yöntemi kullanılmıştır.

K-ortalamalar kümeleme yöntemi (71-73), verilen *k* küme sayısı için en uygun veri parçalanmasını bulmak amacıyla çalışan bir açgözlü (*greedy*) algoritma olmanın ötesinde en basit ve en popüler kümeleme yöntemlerinden biridir (74). Yöntem, *k* tane küme merkezinin rastgele seçimi ile başlar. Burada *k* bir girdi parametresidir. Bu merkezler *v* değişkenin *v* boyutlu uzayındaki rastgele noktalar olabileceği gibi, verideki rastgele seçilmiş örnekler (nesneler) de olabilir. Veri yapısına göre ön tanımlı olan örnekler arası benzerlik ya da uzaklık/yakınlık ölçülerinin ve küme merkezlerinin kullanılmasıyla her örnek kendine en yakın merkezin olduğu kümeye atanır (1, 75). Daha sonra, küme merkezleri yeniden tanımlanır (1). Küme merkezlerinin yeniden tanımlanması adımında, yeniden belirlenecek olan bu küme merkezleri amaç fonksiyonu ve yakınlık/uzaklık ölçümüne bağlı olarak değişecektir (75). Ardından örnekler, bu yeni merkezlere olan uzaklıklarına bağlı olarak yeniden kümelere atanır. Bu yinelemeli süreç örneklerin kümelere atanmasında herhangi bir değişiklik olmayıncaya kadar devam eder.

Farklı başlangıç noktaları farklı kümelerin ortaya çıkmasına neden olabileceğinden yöntemin tekrarlı olarak çalıştırılması ve küme içi değişkenliğin en küçük olduğu çözümün seçilmesi önerilmektedir. Aynı zamanda, k-ortalamalar yönteminin temel dezavantajı küme sayısının kullanıcı tarafından belirlenmesi gerekliliğidir. Her ne kadar *k* parametresinin farklı değerleri için ek tekrarlar eklenebilse de (örneğin en küçük küme içi değişkenliği sağlayan *k* değerinin seçimi

söz konusu olabilse de) yöntem sonunda, örnekler arası ya da kümeler arası ilişkiler hakkında ön bilgi sahibi olmaksızın bir k kümeler topluluğu elde edilir (1). Bunun dışında aykırı değerlerden etkilenme, her veri türüne uygun olmama, küresel olmayan ve farklı nokta sayısı ve yoğunluğuna sahip kümeler ile başa çıkamama gibi çeşitli ek dezavantaj ve kısıtlılıklara da sahiptir (75).

İki-yönlü/boyutlu kümeleme bakış açısına göre, gen ifade düzeyi veri seti örnek ve gen boyutları olmak üzere iki boyutta değerlendirilir. SNFS kapsamında bunun başarılabilmesi için aslında tek boyutta kümeleme yapan k-ortalamalar kümeleme yönteminin ikinci boyutu değerlendirebilmesi amacıyla önce veri matrisinin devriği alınır. Böylece örnekler özellik, özellikler örnek olarak girdi biçimine dönüşür. Verinin kümeleme adımı öncesi ön hazırlık yapılarak sınıflara *(attribute*) göre değerlerin birleştirilmesi gerekir. Diğer bir deyişle, sınıflara göre gen ifade düzeylerinin ortalamalarının elde edilmesi gerekir. Ardından metrik olarak "Öklid" uzaklığının seçilmesiyle k-ortalamalar kümeleme yöntemi farklı sayıda *k* parametresi ile tekrarlı şekilde uygulanır. Eğer iki gen aynı kümede birden fazla tekrarda ortaya çıkıyor ise benzer fonksiyonlara sahip oldukları ve etkileşimde oldukları varsayımı altında, birlikte ortaya çıkışları yani genlerin etkileşimine ilişkin bu şekilde bilgi elde edilir. SNFS yönteminde *k* küme sayısı kullanıcı tarafından belirlenir (*k*=3, 4 ya da 5). Başka bir deyişle, yöntemin her bir *k* küme sayısı için farklı bir alt versiyonu söz konusudur. Ancak, alanyazında SNFS kapsamında belirlenen *k* küme sayısı için k-ortalamalar kümeleme yönteminin kaç kez tekrar edilmesi gerektiğine ilişkin net bir bilgi bulunamamıştır (14). Aynı zamanda, SNFS'nin yöntem bölümünde anlatılandan farklı olarak, Lösemi veri seti üzerinde *k* küme sayısı 2, 3 ve 4 olmak üzere her bir *k* küme sayısı için tek tekrar uygulanarak birlikte ortaya çıkışların elde edildiğinden de bahsedilmiştir (14).

Bu nedenle bu tez çalışmasında kümeleme adımında değişikliğe gidilerek *k* küme sayısı 3, 4 ve 5 için birer tekrar söz konusu olmak üzere k-ortalamalar kümeleme yöntemi 3 kez çalıştırılarak işlem yapılmış ve genlerin birlikte ortaya çıkışlarına ilişkin bilgi edinilmiştir.

Bu bilgi kullanılarak üçüncü adımda yapılacak olan sosyal ağ analizinde kullanılacak olan ağırlıklı komşuluk (*adjacency*) matrisi kolayca yapılandırılabilir. Başka bir deyişle, k-ortalamalar kümeleme yöntemi birden fazla kez tekrarlanır. Bu

tekrarlarda aynı kümede ortaya çıkan genlerin birlikte ortaya çıkışlarının saydırılır. Bu yolla, elemanlarının aldığı en küçük değerin 0 ve en büyük değerin "kümeleme yönteminin tekrar sayısı" olduğu bir pxp kare matris oluşturulmaktadır (14). Böylece ağırlıklı ağ yapısı için kullanılacak temel veri matrisi elde edilmiş olur.

Ancak, gen ifade verileri söz konusu olduğunda uygulanacak bir sosyal ağ analizi için komşuluk matrisinin oluşturulmasında kullanılabilecek yöntem yalnızca bu yaklaşımla sınırlı değildir. Genler arası etkileşimin belirlenmesinde ortak-ifade ağı ya da korelasyon ağı olarak da tanımlanan ağırlıklı ağ yapılarının oluşturulması, SNFS analizinin bu adımında harcanan hesaplama zamanını azaltmak ve doğrudan gen ifade düzeylerinden elde edilen bilgide kayıp yaşanmadan analiz yapmak için başvurulabilecek farklı bir yaklaşım olarak düşünülebilir (13, 17). Çünkü korelasyon ağının kullanılmasıyla, verinin kümeleme için hazırlanması, k-ortalamalar kümeleme yönteminin tekrarlı olarak uygulanması, birlikte ortaya çıkışların saydırılması süreci atlanarak doğrudan sosyal ağ analizine geçiş bu şekilde olanaklı hale gelmektedir. Bu nedenle, tez çalışması kapsamında SNFS'nin ilk adımında indirgenmiş genlere ilişkin korelasyonların kullanılması yoluyla sosyal ağ analizi için komşuluk matrisinin oluşturulması yaklaşımı da dikkate alınmıştır.

Biyolojik ağlar da dahil olmak üzere gerçek hayatta karşılaşılan pek çok ağda, ağın elemanlarının yani düğümlerin arasındaki bağların gücü eşit olmayabilir. Bazı bağların diğerlerine göre daha güçlü ya da zayıf olması oldukça olasıdır. Bu nedenle, bağlara ilişkin ağırlıklara ulaşılabilir ise bu bilginin göz ardı edilmemesi, en azından teoride, ilgili sistemi daha iyi anlamada araştırmacıya yardımcı olacaktır (13).

Örneğin, böyle bir ağda iki genin ifade düzeyleri arasındaki korelasyonun işareti bir genin diğer bir gen nedeniyle yukarı ya da aşağı düzenlenmesini ifade edebilir (13, 18).

Korelasyon ağı gibi ağırlıklı bir ağda, komşuluk matrisi elemanları sadece sıfır ve bir değerlerini almaz. Böyle bir ağda komşuluk matrisi elemanları düğümler arası bağların ağırlıklarından oluşur. Örneğin, genlerin söz konusu olduğu bir korelasyon ağına ilişkin komşuluk matrisi A_{ij} genler arası ikili korelasyonlar temelinde oluşturulur. Her ne kadar ağırlıklı bir ağda, ağırlıkların negatif olması olası olsa da, sosyal ağ analizinde kullanılan topluluk belirleme yöntemleri negatif ağırlıklar ile çalışamazlar. Bu nedenle, alanyazında korelasyon ağına ilişkin komşuluk matrisinin

oluşturulmasında farklı yaklaşımlar (13, 17, 76, 77) söz konusudur. Ancak, yaygın olarak ağırlıklı komşuluk matrisi aşağıdaki gibi tanımlanabilir.

$$A_{ij} = \left|\mathrm{cor}(x_i, x_j)\right|^{\beta} \tag{3.2.}$$

Burada üs parametresi $\beta \geq 1$ koşulunu sağlamak zorundadır. SNFS'nin ilk adımında filtreleme kullanıldığından ölçekten-bağımsız ağ yapısı özellikleri bozulmaya uğramakta ve yumuşak eşik (*soft-threshold*) değeri olarak da ifade edilen üs parametresinin belirlenmesinde alanyazında önerilen yaklaşımlar uygun olmamaktadır (13). Bu nedenle, tez kapsamında üs parametresi $\beta=1$ olarak belirlenmiştir (17).

Üçüncü adımda ağa özel metriklerin hesaplanmasında ağırlıksız komşuluk matrisine de gereksinim duyulduğundan ağırlıksız korelasyon ağı, korelasyon matrisinin mutlak değeri belirlenen bir eşik değer kullanılması yoluyla (*hard-thresholding*) aşağıdaki gibi yapılandırılır.

$$A_{ij} = \begin{cases} \left|\mathrm{cor}(x_i, x_j)\right| \geq \tau \text{ ise } 1 \\ \left|\mathrm{cor}(x_i, x_j)\right| < \tau \text{ ise } 0 \end{cases} \tag{3.3.}$$

Burada sert eşik (*hard-threshold*) değeri olarak da adlandırılan τ parametresinin seçimi için farklı yaklaşımlar söz konusudur (13, 76, 77). Tez çalışmasında, uygulama kolaylığı açısından korelasyonların mutlak değerinin sıralanması sonrası 99. yüzdeliğe karşılık gelen korelasyon değerinin sert eşik değerinin kestirimi olarak kullanılması yaklaşımı benimsenmiştir (77).

3.1.3. Sosyal Ağ Analizi ve Toplulukların Belirlenmesi

Topluluk belirleme, temel olarak ağdaki kümelerin belirlenmesi işlemi olarak tanımlanabilir. En çok bilinen topluluk belirleme yöntemleri Girvan-Newman; Clauset, Newman, ve Moore; Pons ve Latapy; Watika ve Tsurumi; Louvain, vb. olmakla birlikte bir sosyal ağdaki toplulukların belirlenmesinde alanyazında farklı algoritma ve yöntemlerin olduğu bilinmektedir (78).

SNFS kapsamında topluluk belirlenme yöntemlerinden sadece Louvain yöntemi (79) kullanılmıştır (14). Ancak alanyazında Louvain yönteminden daha iyi

başarım gösteren Infomap yöntemi (80), Lancichinetti ve Fortunato'nun (81) karşılaştırmalı çalışmasında incelenmiştir. Bu nedenle, Infomap yönteminin de SNFS kapsamında incelenmesi mutlaka düşünülmelidir. Aynı zamanda Walktrap yönteminin de (82) iyi bir başarım gösterdiği alanyazında gösterilmiştir (83). Tez çalışması kapsamında Louvain dışında bu iki farklı yöntemin de SNFS'nin topluluk belirleme aşamasında incelenmesi söz konusudur.

Louvain Yöntemi: Modülarite (*modularity*), topluluklar arası bağlantılarla karşılaştırıldığında topluluk içi bağlantıların yoğunluğunu ölçen ve -1 ile +1 arasında değerler alabilen bir ölçüttür. Farklı yöntemler tarafından elde edilen bölüntülerin (*partitions*) kalitesini karşılaştırmak için kullanılan modülarite, aynı zamanda en iyilenebilecek bir görev fonksiyonudur. Modülarite, ağırlıklı ağlarda aşağıdaki gibi tanımlanır:

$$Q = \frac{1}{2m}\sum_{i,j}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j) \qquad (3.4.)$$

Burada A_{ij} i düğümü ile j düğümü arasındaki bağlantının ağırlığı, $k_i = \sum_j A_{ij}$ olmak üzere i düğümüne bağlı bağların ağırlıklarının toplamı, c_i i düğümünün atandığı topluluk, δ fonksiyonu $\delta(u, v)$ eğer $u = v$ ise 1 ve aksi durumda 0, $m = \frac{1}{2}\sum_{ij} A_{ij}$ olarak tanımlanır (79).

Louvain yöntemi, modülarite en iyilemesine dayalı olan hiyerarşik bir topluluk belirleme yöntemidir. Yöntemin yinelemeli olarak tekrar eden iki aşaması vardır. Sürece N düğüm içeren bir ağırlıklı ağ ile başlandığı varsayılırsa, ilk olarak ağdaki her bir düğüm sadece kendisinin olduğu farklı bir topluluğa atanır. Böylece, başlangıçta ağdaki düğüm sayısı kadar başlangıç bölüntüleri oluşturulmuş olur. Daha sonra, her i düğümü için bir komşu j düğümü ele alınır ve i düğümünün kendi topluluğundan kaldırılarak j düğümünün olduğu topluluğa yerleştirilmesi ile elde edilecek olan modülaritedeki kazanç değerlendirilir. Kazancın pozitif olması durumunda, kazancın en büyük olduğu topluluğa i düğümü yerleştirilir. Eğer pozitif kazanç söz konusu değilse, i düğümü orijinal topluluğunda kalır. Bu süreç, yinelemeli ve sıra gözeterek tüm düğümler için (genellikle bir düğüm birkaç kez değerlendirilir) hiçbir iyileşme sağlanamayana kadar devam eder. Başka bir deyişle, modülarite için yerel maksimuma

erişilince yani hiçbir düğümün hareketi modülaritede iyileşme sağlamayınca bu adım tamamlanır (79).

Louvain yönteminin etkin olmasını sağlayan kısımlardan biri, bir *i* düğümünü *C* topluluğuna geçirerek elde edilen modülaritedeki kazanç ΔQ'dur ve aşağıdaki biçimde kolaylıkla hesaplanabilir.

$$\Delta Q = \left[\frac{\Sigma_{in} + 2k_{i,in}}{2m} - \left(\frac{\Sigma_{tot} + k_i}{2m}\right)^2\right] - \left[\frac{\Sigma_{in}}{2m} - \left(\frac{\Sigma_{tot}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right] \qquad (3.5.)$$

Burada, Σ_{in} *C* topluluğu içindeki bağların ağırlıklarının toplamı, Σ_{tot} *C* topluluğundaki düğümlere özgü bağlantıların ağırlıklarının toplamı, k_i *i* düğümüne özgü bağların ağırlıklarının toplamı, $k_{i,in}$ *C* topluluğundaki düğümlere *i* düğümünden giden bağların ağırlıklarının toplamı ve m ağdaki tüm bağların ağırlıklarının toplamı olarak tanımlanır. *i* düğümünün kendi topluluğundan kaldırılması sonucunda modülaritede ortaya çıkan değişimin değerlendirilmesi için de benzer bir ölçü kullanılır (79).

İkinci aşamada, ilk aşamada elde edilen toplulukların ağdaki düğümler olduğu yeni bir ağ oluşturulur. Bu yeni düğümler arası bağların ağırlıkları ise karşılık gelen iki topluluk arasındaki bağların ağırlıklarının toplamıyla elde edilir. Aynı topluluk içinde yer alan düğümler arası bağlar, bu yeni ağda iç döngü (*self-loop*) olarak değerlendirilir. İkinci aşama tamamlandığında, elde edilen bu ağırlıklı yeni ağ ile ilk aşama yeniden uygulanır. Bu şekilde, yöntem yinelemeli olarak gerçekleşir (79).

Sonuç olarak, başlangıçta her düğüm sadece kendisinin yer aldığı bir topluluğa atanır. Adım adım düğümler başka bir topluluğa taşınarak modülaritedeki kazanç hesaplanır ve her bir düğüm modülariteye en büyük katkının sağlandığı topluluğa atanır. Daha sonra, elde edilen her topluluk ağdaki birer düğüm gibi düşünülür ve süreç birleştirilmiş topluluklar ile yeniden başlar. Ağda tek bir düğüm kalıncaya ya da modülarite hiçbir şekilde arttırılamayıncaya dek süreç devam eder (79).

Walktrap Yöntemi: Sonuçların güvenirliğinden ödün vermeden yüksek hesaplama etkinliği sağlayan bir topluluk belirleme yöntemi olan Walktrap yöntemi, rasgele yürüyüşe dayalı bir yaklaşımdır. Yöntemin uygulanması, her bir düğüm çifti için geçiş olasılıklarının yer aldığı geçiş matrisinin oluşturulması ile başlar. Matrisin her bir elemanı, bir rasgele yürüyüşte bir düğümden bir düğüme gitmenin geçiş

olasılığı olarak tanımlanabilir. Ağa özel metriklere dayalı olan geçiş olasılıkları, her bir düğüm çifti için uzaklık ölçüsü hesaplama amacıyla kullanılır. Daha sonra, geleneksel hiyerarşik kümeleme yöntemi bu uzaklık matrisine uygulanır. Başka bir deyişle, her bir düğümün kendi topluluğu içindeki diğer düğümler ile arasındaki uzaklığın kareler toplamının en küçüklenmesi yoluyla topluluklar oluşturulur (17, 82).

Temelde bu yaklaşım, topluluk dışına oranla topluluk içi geçişlerin daha fazla olduğu toplulukların elde edilmesini amaçlar. Bu yığınsal yaklaşım en uygun bölüntülemeyi seçmek için modülariteden yararlanır (17, 82)

Infomap Yöntemi: Walktrap yöntemine benzer şekilde rasgele yürüyüşten yararlanan Infomap yaklaşımında da topluluk içi geçişlerin topluluklar arası geçişlere oranla daha düşük olduğu toplulukların elde edilmesi amaçlanır.

Ancak, bu yaklaşım rasgele yürüyüşten elde edilen bilginin sıkıştırılmasını içerir. Bu açıdan Walktrap yönteminden ayrılan Infomap yöntemi kapsamında yapılan bu sıkıştırma ile oldukça büyük boyutlu ağlar için hesaplama etkinliği sağlanabilmektedir. Infomap yöntemiyle, ağdaki bilgi dağılımını tanımlamak için gereken bilgiyi içeren en uygun sıkıştırma elde edilir. Yöntem kapsamında en küçükleme, Louvain yöntemindeki açgözlü aramanın bir tür adaptasyonu ve benzetimli tavlama (*simulated annealing*) gibi birkaç yöntemin birleştirilmesi ile gerçekleştirilir. Louvain yöntemindekine benzer şekilde, gizli süper-düğümlere (*supernodes*) ulaşılabilmesi amacıyla düğümler yinelemeli olarak birleştirilirler. Daha sonra, birleştirilen bu düğümler arasındaki komşuluklar ele alınır. Bu yaklaşım ile ağdaki bilgi akışı sadece ikili ilişkilere odaklanan modülariteden daha iyi şekilde yakalanabilmektedir (17, 80).

SNFS kapsamında uygulanan topluluk belirleme sonrası, elde edilen topluluklarda yer alan genlerin değerlendirilmesi adımına geçilir. Bu adımda ağa özel metrikler hesaplanır ve bir ya da birkaçının bir arada değerlendirilmesi ile biyobelirteçler belirlenir.

SNFS kapsamında hesaplanan ağa özel metrikler sırasıyla "Kapsama (*Coverage*)", "Bağlılık derecesi (*Degree of domesticity*)", "Topluluk-içi ağırlıksız derece merkeziliği (*Intra-community unweighted degree centrality*)", "Topluluk-dışı ağırlıksız derece merkeziliği (*Out-of-community unweighted degree centrality*)",

"Ağırlıksız derece merkeziliği (*Unweighted degree centrality*)" ve "Ağırlıklı derece merkeziliği (*Weighted degree centrality*)" olarak sıralanabilir (14).

Ağırlıksız derece merkeziliği (z):

Derece (*degree*) ya da derece merkeziliği (*degree centrality*) ilk kez 1978 yılında Freeman (84) tarafından, fokal düğümün ağdaki komşuluk sayısı, yani bir başka deyişle fokal düğümün bağlı olduğu düğüm sayısı olarak tanımlanmıştır. Derece, ağların analiz edilmesinin ilk aşaması olarak ele alınan basit bir göstergedir ve Eşitlik 3.6.'daki gibi tanımlanır.

$$z(i)=\sum_{j=1}^{n} x_{ij} \qquad (3.6.)$$

Burada *i* indisi fokal düğümü ve *j* indisi diğer tüm düğümleri temsil etmektedir. *n* toplam düğüm sayısını ve *x*, x_{ij} hücresi eğer *i* düğümü *j* düğümü ile bağlı ise 1; değilse 0 olarak tanımlanan ağırlıksız komşuluk matrisini temsil etmektedir (85).

SNFS yöntemi kapsamında alanyazında derece olarak bilinen ağa özel metrik, ağırlıksız derece merkeziliği olarak adlandırılarak kullanılmıştır.

Ağırlıklı derece merkeziliği (D):

Derece, ağırlıklı ağlar söz konusu olduğunda ağırlıkların toplamı olarak genişletilebilir ve düğüm gücü (*node strenght*) olarak adlandırılır (84). Ağırlıklı derece merkeziliği ise bir düğümün gücünün ilgili düğümün bulunduğu topluluk genişliğine bölünmesi yoluyla aşağıdaki gibi elde edilir.

$$D_i=\frac{\sum_{j=1}^{n} w_{ij}}{n_c} \qquad (3.7.)$$

Burada, n_c *i* düğümünün bulunduğu topluluğun örnek genişliği (topluluktaki özellik sayısı), w_{ij} *i*. düğüm ile *j*. düğüm arası bağlantının ağırlığı/gücü olarak tanımlanabilir (14).

Topluluk-içi ağırlıksız derece merkeziliği (z_{in}):

i düğümünün yani n_i geninin kendi topluluğunda kurduğu bağ sayısıdır. Pozitif tam sayı değerleri alır ve $z_{in}(i)$ ile ifade edilir (14).

Topluluk-dışı ağırlıksız derece merkeziliği (z_{out}):

i düğümünün yani n_i geninin kendi topluluğu dışında kurduğu bağ sayısıdır. Negatif olmayan tam sayı değerleri alır ve $z_{out}(i)$ ile ifade edilir. Dolayısıyla *i.* genin derecesi ilgili genin topluluk-içi ve topluluk-dışı derece merkeziliğinin toplamı ile bulunabilir. Başka bir deyişle,

$$z(i)=z_{in}(i)+z_{out}(i) \tag{3.8.}$$

olarak yazılabilir (14).

Bağlılık derecesi (z_d):

i. düğümün kendi topluluk elemanları ile kurduğu bağ sayısının kendi topluluğu dışındaki düğümler ile kurduğu bağ sayısına oranı olarak tanımlanabilir. Buna göre,

$$z_d=z_{in}(i)/z_{out}(i) \tag{3.9.}$$

olarak hesaplanır (14). Eşitlik 3.9.'dan da görülebileceği gibi, ilgilenilen düğümün kendi topluluğu dışında hiç bağı yoksa bağlılık derecesi tanımsız olmaktadır. Bu durumda, kendi topluluğu dışında hiç bağı olmayan düğümler bağlılık derecesi açısından karşılaştırılamamaktadır. Başka bir deyişle, ilgili düğümün kendi topluluğunda kurduğu bağ sayısının etkisi görülememektedir.

Bu nedenle tez kapsamında, bağlılık derecesinin hem pay hem de paydasına oldukça küçük bir epsilon değeri eklenerek bu sorunun üstesinden gelinmiştir. Böylece, kendi topluluğu dışında hiç bağı olmayan düğümlerin kendi toplulukları içinde kurdukları bağ sayısının etkisi görülebilmiştir.

Kapsama (Cov):

Kapsama, n_i genine ilişkin genin kendi topluluğu içindeki bağlantılarına bağlı bir metriktir. Diğer topluluk üyelerine göre ilgili genin kapsam düzeyi olarak adlandırılabilir ve pozitif gerçel bir sayıdır (14).

$$Cov(i)=z_{in}(i)/n_c \tag{3.10.}$$

Bu ağa özel metrikler dışında, SNFS'nin ilk adımında DVM-YÖE'den ilgili gen için elde edilen sıralama olarak tanımlanmış "*Identity* (ID)" ölçüsü de genleri değerlendirmek amacıyla kullanılmıştır (14). Ancak bu tez çalışmasında ID, SNFS'nin ilk adımında yalnız tek bir özellik seçim yöntemi kullanılmadığından bu adımda kullanılan özellik seçim yöntemlerinden elde edilen tümel sıralama olarak ele alındı.

Ağa özel metriklere göre değerlendirme yapılırken eğer bir genin kendi topluluğunun üyeleri ile etkileşim düzeyi yüksek, kendi topluluğu dışındaki üyelerle etkileşim düzeyi düşük ise o genin topluluğu temsil yeteneğinin iyi olduğu düşünülmektedir. Bu aşamda, bir ya da birden fazla metrik kullanılarak genler sıralanabilir. Tez çalışması kapsamında ise tek bir metriğe (bağlılık derecesi) göre genlerin sıralanması tercih edilmiştir. İlgili metriğe göre sıralamaların elde edilmesi sonrasında her toplulukta yer alan genlerden, en iyi belirli yüzdedeki gen biyobelirteç olarak belirlenir. Tez çalışmasında, bu yüzde uygulamada %10, benzetim çalışmasında ise %5 olarak alınmıştır.

Seçilen bu biyobelirteçlerin sınıflama başarımı, başka bir deyişle elde edilen sınıflama modelinin geçerliği ise *holdout* yöntemi ile DVM kullanılarak test edilmiştir.

3.2. Çalışmada Kullanılan Veri Setleri

3.2.1. Gerçek Veri Setleri

Lösemi Veri Seti: Golub ve diğ. (30)'nin çalışmasında yer alan veri seti, tanı konduğunda akut lösemi hastalarından alınmış 38 (27 ALL, 11 AML) kemik iliği örneğinden oluşmaktadır. İlgili veri seti kemik iliği mononükleer hücrelerinin hazırlanmasıyla elde edilen RNA'nın yüksek yoğunluklu oligonükleotid mikrodizisine melezlenmesi ile Affymetrix Hgu6800 çip tarafından üretilmiştir ve 6817 insan geni probu içermektedir. Aslında, veriye ilişkin 7129 probe seti (kontrol ve gereksiz genler ile birlikte) söz konusudur. Daha sonra, 24 kemik iliği ve 10 periferik kan örneğinden elde edilen toplam 34 (20 ALL, 14 AML) örneğe ilişkin veri de bu veri setine eklenmiştir (30).

Tez çalışmasında kullanılan ve Tablo 3.4.'te yer alan özelliklere sahip Lösemi veri seti, R Bioconductor golubEsets paketinde (86) yer almaktadır. Pakette exprSet

sınıfında bir nesne olarak tanımlı olan veri setinin ExpressionSet sınıfına dönüştürülmüş olan güncellemesi de daha sonra pakete konulmuştur. Bu iki nesne sınıfının temel farkı, R yazılımında farklı nesne sınıflarını tanımlıyor olmalarıdır. Tez çalışması kapsamında, kolaylık olması açısından bir exprSet sınıfından nesneyi ExpressionSet sınıfına dönüştüren bir fonksiyon da hazırlanmıştır. Böylece, yalnızca Lösemi veri setinin değil herhangi bir exprSet sınıfından mikrodizi verisinin de kullanımı mümkün olmaktadır. Bu paketteki tanımına göre veri setinde 7129 prob yani gene ilişkin gen ifade düzeyi yer almaktadır.

Tablo 3.4. Lösemi veri setine ilişkin bilgiler (30, 86).

Lösemi Veri Seti	
Gen Sayısı	**:** 7129
Eğitim Seti Örnek Sayısı	**:** 38 (27 ALL, 11 AML)
Test Seti Örnek Sayısı	**:** 34 (20 ALL, 14 AML)
Toplam Örnek Sayısı	**:** 72
Sınıf Sayısı	**:** 2 (AML, ALL)

Veri Ön İşleme: Çapraz geçerliğin en basit hali olan "*Holdout* yöntemi" SNFS yönteminin sınıflama başarımını inceleyebilmek amacıyla kullanılacağından mikrodizi veri setinin eğitim ve test olmak üzere iki parçaya ayrılması ya da önceden tanımlanmış eğitim ve test setinin kullanılması gerekmektedir. Bu nedenle, önceden var olan eğitim ve test setinin kullanılmasının yanı sıra veri setinin istenen oranda rastgele olarak eğitim ve test setine ayrılmasını sağlayan bir R fonksiyonu da yazılmıştır. Ancak, sonuçların alanyazın ile karşılaştırılabilir olması açısından rastgele ayrılan eğitim ve test seti üzerinde SNFS yöntemi uygulanmamıştır.

Aynı zamanda, alanyazında yer alan çalışmaların birçoğunda Lösemi veri seti Dudoit ve diğ. (32)'nin çalışmasına göre ön işleme (ölçümlerin dağılım aralığını değiştirecek dönüşüm, logaritmik dönüşüm) ve filtreleme sürecinden geçirilmektedir. Çünkü bu ön işleme ve filtreleme süreci Dudoit ve diğ. (32)'ne kişisel yazışma yoluyla Lösemi veri setinin yer aldığı orijinal çalışmanın (30) yazarları tarafından önerilmiştir. Aynı zamanda, logaritmik dönüşümün bu veri seti için sınıflama başarımı üzerinde olumlu etkiye sahip olduğu da belirtilmiştir (32). Bu nedenle, tez çalışması kapsamında da Dudoit ve diğ. (32)'nin çalışmasına göre veri seti ön işlemeden geçirilmiş ancak filtrelenmemiştir. Filtreleme adımının göz ardı edilmesinin nedeni

SNFS kapsamında ilk adımda yapılacak olan özellik seçim adımının etkisini gözlemleyebilmektir.

Kolon Kanseri Veri Seti: Tez kapsamında, SNFS yönteminin R yazılımında uygulanması için hazırlanan fonksiyonların Lösemi veri seti dışında başka veri setleri üzerinde de kullanılabilirliğinin test edilmesi amacıyla, erişime açık bir başka veri seti olan Kolon kanseri veri seti (87, 88) incelenmiştir. Kolon kanseri veri seti (87), Affymetrix Hum6000 oligonükleik dizi teknolojisi ile elde edilen bir veri setidir ve kolon kanseri hastalarından alınan 62 örneğin (tümör dokusundan alınan 40 örnek, normal dokudan alınan 22 örnek) analizi sonucu elde edilmiştir. Bu veri seti R yazılımında colonCA paketi (88) kapsamında yer almaktadır. Veri setine ilişkin temel özellikler ise Tablo 3.5.'te görülebilir.

Tablo 3.5. Kolon kanseri veri setine ilişkin bilgiler (87, 88).

Kolon Kanseri Veri Seti	
Gen Sayısı	**:** 2000
Toplam Örnek Sayısı	**:** 62 (40 Tümör, 22 Normal)
Sınıf Sayısı	**:** 2 (Normal doku, Tümör dokusu)

Kolon kanseri veri seti colonCA paketinde (88) hazır şekilde eğitim ve test seti olarak ayrılmamış olduğundan, eğitim ve test setlerinin kullanıcı tarafından belirlenen oranlarda (%60 eğitim, %40 test) ayrılmasını ve SNFS yönteminin uygulanmasını sağlayan R fonksiyonları tez kapsamında hazırlanarak kullanılmıştır. Böylece, tümörlü doku örnek sayısının normal doku örnek sayısına oranı da dikkate alınarak veri setinin rastgele şekilde eğitim ve test seti olarak ayrılması söz konusu olmuştur.

Ayrıca bu veri setine colonCA paketi (88) kapsamında gen indirgeme haricinde herhangi bir ön işleme (normalizasyon, vs.) uygulanmamış olduğundan Alon ve diğ. (87)'nin çalışmasında uyguladığı gibi standartlaştırma ile gen ifade düzeyleri üzerinde ön işleme yapılmıştır.

Özyer ve diğ. (14)'nin çalışmasında da Kolon kanseri veri seti kullanılmıştır. Ancak, verideki temel farklılıklar (farklı eğitim ve test setleri, farklı ön işleme yöntemleri, vb.) ve yönteminin sınıflama başarımının başarım ölçülerine göre değerlendirilmemesi nedeniyle, tez çalışmasında bu veri seti için elde edilen bulguların alanyazındaki sonuçlarla (14) doğrudan karşılaştırılabilir olduğu düşünülmemektedir.

Bu nedenle, tez kapsamında bu veri setine ilişkin sadece R yazılımında elde edilen bulgulara yer verilmiştir.

3.2.2. Sentetik Veriler ve Benzetim Çalışması

Alanyazında, melez yöntemlerin başarımının değerlendirilebilmesi amacıyla yapılacak bir benzetim çalışması için gereken verinin üretim sürecinde kullanılabilecek farklı yaklaşımlar olduğu görülmüştür. Bunlardan bazıları,

a) Özellik seçimi için belirlenmiş sentetik/yapay veri setlerinin yeniden örnekleme yoluyla kullanımı;
b) Özellik seçimi için kullanılan gerçek veri setlerinden yeniden örnekleme ile veri setlerinin elde edilmesi;
c) Özellik seçimi, sınıflama ya da kümeleme yöntemlerinin karşılaştırıldığı benzetim çalışmalarında kullanılan benzer yaklaşımlarla çok değişkenli normal dağılımdan veri üretimi;
d) İki boyutlu kümeleme için belirlenmiş yaklaşımlarla veri üretimi;
e) Topluluk belirleme yöntemlerinin karşılaştırılması ve sonuçların geçerliğinin incelenmesinde kullanılan *Benchmark* veri setlerinin kullanımı olarak sıralanabilir.

Bu tez çalışması kapsamında yapılan benzetim çalışması için yukarıda belirtilen yaklaşımlardan (a) sentetik veri seti kullanımı ile (c) çok değişkenli normal dağılımdan veri üretimi yaklaşımları benimsenmiştir. Bu kapsamda, özellik seçim yöntemlerinin başarımlarının değerlendirilmesinde kullanılan sentetik veri setlerinden Madelon'un (89) kullanılması ve Guo ve diğ. (42)'nin veri üretim yaklaşımı kullanılarak veri üretilmesi tercih edilmiştir.

Veri üretim sürecinde Guo ve diğ. (42) esas alındığından, "iki grup bağımlılık yapısı" adı verilen ve gerçek mikrodizi verilerinin yapısını daha iyi temsil ettiği belirtilen bir senaryo söz konusudur. Benzetim aşamasında veri, iki gruplu sınıflama problemi temel alınarak aşağıdaki tanımlamaya uygun olarak üretilmiştir.

Her bir sınıf eşit sayıda örnek içerecek şekilde toplam 200 eğitim örneği ile p=10000 gen üretilmiştir. 10000 gen, her biri 100 gen içeren k=100 bloğa bölünmüştür. Farklı bloklardaki genlerin birbirlerinden bağımsız, aynı bloktaki genlerin ise otoregresif yapı olarak adlandırılan ve aşağıda gösterilen kovaryans yapısına uygun

şekilde ilişkili olduğu varsayılmıştır. Aynı blok içinde |ρ|=0,9 (eğitim setindeki 50 blok için pozitif, 50 blok için ise negatif) olarak alınmıştır.

$$\Sigma = \begin{pmatrix} 1 & \rho & \cdots & \rho^{98} & \rho^{99} \\ \rho & 1 & \ddots & \ddots & \rho^{98} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \rho^{98} & \ddots & \ddots & \ddots & \rho \\ \rho^{99} & \rho^{98} & \cdots & \rho & 1 \end{pmatrix} \qquad (3.11.)$$

Buna göre, ilk olarak her genin ifade düzeyi standart normal dağımdan üretilmiştir. Daha sonra, Eşitlik 3.11.'de tanımlanan kovaryans matrisinin kare kökü ile çarpılan ifade düzeyleri böylece dönüşüme uğramış ve temelde MVN(0,Σ) ile çok değişkenli normal dağılıma uymuştur. Bundan sonra ise sadece ikinci sınıftaki ilk 200 genin ifade düzeyine 0,5 sabiti eklenmiştir. Eğitim verisi bu şekilde üretildikten sonra ise 600 örnek içeren test verisi de benzer şekilde üretilmiştir.

Yukarıda bahsedilen yapı ile R yazılımında veri üretimi için sortinghat paketinde (90) yer alan "simdata_guo" fonksiyonu kullanılmıştır. Ancak bu fonksiyon ile üretilen verilerde ikinci sınıfta yer alan belli sayıdaki gen ifade düzeyine 0,5 sabiti eklenmemiş olduğundan, bu işlem fonksiyondan elde edilen veriye sonradan uygulanmıştır.

4. BULGULAR

4.1. Lösemi Veri Setinden Elde Edilen Uygulama Sonuçları

İlk olarak, platform ve/veya kullanılan veri setinden kaynaklı farklılıkların etkisini gözlemleyebilmek amacıyla SNFS yöntemi sonucunda Özyer ve diğ. (14)'nin Lösemi veri seti için elde ettikleri biyobelirteç genlere ("M23197_at", "X95735_at", "X59417_at", "Y12670_at", "X04085_rna1_at", "U22376_cds2_s_at", "M81933_at", "M84526_at", "U05259_rna1_at") ilişkin gen ifade düzeyleri doğrudan kullanılmıştır. e1071 paketinde (91) yer alan "svm" fonksiyonu ile uygulanan DVM'den bu genlere ilişkin gen ifade düzeyleri doğrudan kullanıldığında elde edilen sınıflama başarımı incelenmiştir.

Buna göre, DVM'de Özyer ve diğ. (14) ile aynı çekirdek fonksiyonun (polinomiyal) kullanılması ile eğitim setinde %100 doğru sınıflama oranına ulaşılırken test setinde doğru sınıflama oranı, gerçekte AML sınıfına ait 6 örnek ALL sınıfına yanlış olarak sınıflandığından, yaklaşık %82 olarak elde edilmiştir. Özyer ve diğ. (14) ise eğitim setinde yaklaşık %92, test setinde ise %97 doğru sınıflama oranına ulaşmışlardır. DVM'nin sınıflama başarımında ortaya çıkan bu ciddi farklılık, platform farklılığından ya da kullanılan fonksiyonların çalıştırdığı algoritmalardaki küçük değişikliklerden kaynaklanabileceği gibi kullanılan veri setlerinin aslında birebir örtüşmemesinden de ileri geliyor olabilir.

Daha sonra, R yazılımında DVM'nin sınıflama başarımını iyileştirmek amacıyla modele ilişkin en uygun parametreler yine aynı pakette yer alan "tune.svm" fonksiyonu ile elde edilmiştir. Aynı zamanda, bu sınıflama probleminin çözümü için, "svm" fonksiyonu içerisinde yer alan diğer çekirdek fonksiyon seçenekleri de (doğrusal, radyal tabanlı, sigmoid) denenmiştir. Ancak, tez kapsamında yalnızca en iyi sınıflama başarımının elde edilmesini sağlayan çekirdek fonksiyon ile uygulanan en iyilenmiş DVM sonuçları tablolar halinde paylaşılmıştır. Buna göre, aynı genlere ilişkin gen ifade düzeylerinin kullanılmasıyla radyal tabanlı çekirdek fonksiyonun söz konusu olduğu en iyilenmiş DVM'den elde edilen sınıflama başarımı Tablo 4.1. ve Tablo 4.2.'de verilmiştir.

Tablo 4.1. Aynı biyobelirteç genler kullanıldığında Lösemi eğitim setinde radyal tabanlı DVM'nin sınıflama başarımı.

	Gerçek Durum		
Kestirim	ALL	AML	Toplam
ALL	27	0	27
AML	0	11	11
Toplam	27	11	38
Doğru Sınıflama Oranı=1;Duyarlılık=1;Seçicilik=1; AUC=1			

Eğitim setinde hiçbir örnek hatalı sınıflanmamıştır (Bkz. Tablo 4.1.). Ancak, burada asıl önemli olan test setinde elde edilecek başarımdır. Test setinde de yalnız tek bir örnek yanlış sınıflanmıştır (Tablo 4.2.).

Tablo 4.2. Aynı biyobelirteç genler kullanıldığında Lösemi test setinde radyal tabanlı DVM'nin sınıflama başarımı.

	Gerçek Durum		
Kestirim	ALL	AML	Toplam
ALL	20	1	21
AML	0	13	13
Toplam	20	14	34
Doğru Sınıflama Oranı=0,97;Duyarlılık=1;Seçicilik=0,93; AUC=0,99			

Eğitim setindeki doğru sınıflama oranı R yazılımında daha yüksek bulunurken (R yazılımında %100; Java'da yaklaşık %92), test setindeki doğru sınıflama oranı aynı elde edilmiştir (R yazılımında yaklaşık %97; Java'da yaklaşık %97). Böylece, aynı 9 gen kullanılarak karşılaştırma yapıldığında R'da radyal tabanlı çekirdek fonksiyon kullanılarak uygulanan DVM ile Java platformunda Özyer ve diğ. (14)'nin elde ettiği sonuçlardan tümel olarak daha iyi bir sınıflama başarımı elde edildiği görülmüştür.

Aynı veri seti için, dönüştürülmüş gen ifade düzeyleri kullanıldığında ise yine Özyer ve diğ. (14)'nin elde ettiğinden farklı bir sonuçla karşılaşılmıştır. Logaritmik normalizasyonun olumsuz değil özellikle polinomiyal, doğrusal ve sigmoid çekirdek fonksiyon kullanılarak elde edilen sınıflama başarımları üzerinde olumlu bir etki gösterdiği görülmüştür. Örneğin, veri dönüştürülmeden kullanıldığında test setinde

doğrusal çekirdek fonksiyon ile DVM'de %85 doğru sınıflama oranı elde edilirken bu oran, dönüştürülmüş veri kullanıldığında %97'ye yükselmiştir. Benzer şekilde sigmoid çekirdek fonksiyon için dönüşüm doğru sınıflama oranında %94'ten %97'ye yükselme sağlamıştır. Radyal tabanlı çekirdek fonksiyon için eğitim setinde herhangi bir değişim gözlenmemiş (Tablo 4.3.), test setinde doğru sınıflama oranında değişim olmamış (Bkz. Tablo 4.2.) ancak eğri altında kalan alanda yükselme gözlenmiştir (Tablo 4.4.). Polinomiyal çekirdek fonksiyon söz konusu olduğunda ise doğru sınıflama oranı yaklaşık %82'den yaklaşık %79'a gerilemiş ancak eğri altında kalan alanda bir artış söz konusu olmuştur.

Tablo 4.3. Aynı biyobelirteç genler kullanıldığında dönüşüm uygulanmış Lösemi eğitim setinde radyal tabanlı DVM'nin sınıflama başarımı.

	Gerçek Durum		
Kestirim	ALL	AML	Toplam
ALL	27	0	27
AML	0	11	11
Toplam	27	11	38
Doğru Sınıflama Oranı=1;Duyarlılık=1;Seçicilik=1; AUC=1			

Tablo 4.4. Aynı biyobelirteç genler kullanıldığında dönüşüm uygulanmış Lösemi test setinde radyal tabanlı DVM'nin sınıflama başarımı.

	Gerçek Durum		
Kestirim	ALL	AML	Toplam
ALL	19	0	19
AML	1	14	15
Toplam	20	14	34
Doğru Sınıflama Oranı=0,97;Duyarlılık=0,95;Seçicilik=1; AUC=1			

Dönüşüm uygulanmış ve uygulanmamış gen ifade düzeylerinin kullanılması ile elde edilen sınıflama tablosu benzer elde edilmiş olsa da, test setinde eğri altında kalan alanda bir artış söz konusu olduğundan dönüşümün radyal tabanlı çekirdek fonksiyon ile DVM'den elde edilen tümel sınıflama başarımı üzerinde olumlu etkisi olduğu söylenebilir.

Sonuç olarak, R yazılımında uygulanacak olan SNFS yönteminin Java uygulamasından (dolayısıyla alanyazında yer alan orijinalinden) görece olarak farklı sonuçlar üreteceği görülmüştür.

SNFS yöntemi R yazılımı kullanılarak adım adım aşağıdaki gibi Lösemi veri setine uygulanmış ve her adımda elde edilen sonuçlara ilişkin örnek çıktılar tablolar halinde verilmiştir. Böylece, her adımda gerçekleştirilen işlemler daha rahat takip edilerek yöntemin adım adım işleyişi daha rahat görülebilir (Bkz. Şekil 3.1.).

Adım 1: SNFS yönteminin ilk adımında, Lösemi eğitim veri seti (Bkz. Tablo 3.4.) üzerinde özellik seçim yöntemlerinden FSelector paketinde (92) yer alan IG filtresi ve internette yer alan R kodu (93) sayesinde gömülü yöntemlerden DVM-YÖE birlikte uygulanmıştır. Özellik seçim yöntemlerinin uygulanması sonrasında, ayrı ayrı her bir özellik seçim yönteminden genlere ilişkin sıralamalar elde edilebilmektedir. Bu sıralamalar birleştirilerek de genlere ilişkin tümel bir sıralama elde edilmiş olur.

Adım 1'deki sürecin daha iyi anlaşılabilmesi amacıyla, bu adımda elde edilen tümel sıralamaya göre en önemli ilk 6 gene ilişkin sonuçlar aşağıda verilmiştir (Tablo 4.5.). Burada dikkat edilmesi gereken önemli nokta, Tablo 4.5.'te yer alan genlerin tümel sıralamaya göre sıralanmış olduğudur. Başka bir deyişle, tümel sıralamaya göre elde edilen en küçük gen sıra numarası 8 olmuştur.

Tablo 4.5. Lösemi eğitim veri setinden SNFS'nin ilk adımında elde edilen gen sıralamalarına ilişkin örnek tablo.

Gen Adı	DVM-YÖE sıralaması	IG Sıralaması	Tümel Sıralama
M27891_at	5	3	8
U22376_cds2_s_at	37	14	51
M96326_rna1_at	10	45	55
U46751_at	45	19	64
M92287_at	54	12	66
X59417_at	41	30	71

İlk sütunda isimleri yer alan genlere ilişkin DVM-YÖE gömülü yönteminden elde edilmiş sıra numaraları ikinci sütunda, IG filtresinden elde edilmiş sıra numaraları üçüncü sütunda, hem DVM-YÖE gömülü yönteminden hem de IG filtresinden elde

edilmiş sıra numaralarının toplamı ise küçükten büyüğe sıralanmış şekilde son sütunda yer almaktadır. SNFS'ye göre en iyi belirli yüzdedeki (%1) genin seçimi için son sütunda yer alan bu sıra numaraları kullanılmaktadır. Örneğin, "M27891_at" geni bu tümel sıralamaya göre en iyi gen olarak belirlenmiştir (Bkz. Tablo 4.5.). Bu şekilde, toplam 7129 genin %1'i yani yaklaşık 72 gen bu son sıra numaralarına göre seçilmiştir.

Adım 2: Bu adımın başında, indirgenmiş genlere ilişkin gen ifade düzeyleri ön işlemden geçirilir. Bu kapsamda, ilk olarak gen ifade düzeylerinin transpozu alınır. Sonra, sınıflara (ALL, AML) göre gen ifade düzeylerinin ortalamaları elde edilir. Böylece, kümeleme için hazırlanmış 72x2 boyutlu bir veri matrisi elde edilir. Örnek olması açısından elde edilen bu matrisin ilk 6 gene ilişkin olan bölümü Tablo 4.6.'da verilmiştir.

Tablo 4.6. Lösemi eğitim veri setinden SNFS'nin ikinci adımında ön işlem yapılarak elde edilen veri matrisine ilişkin örnek tablo.

Gen Adı	ALL	AML
M27891_at	144,444	7423,545
U22376_cds2_s_at	3863,296	702,634
M96326_rna1_at	571,111	7117,545
U46751_at	1512,148	5291,273
M92287_at	4029,481	1073,636
X59417_at	4361,370	1138,182

Örneğin, "M27891_at" geninin ALL sınıfında gen ifade düzeyi ortalaması 144,444; AML sınıfında ise 7423,545 olarak hesaplanmıştır (Bkz. Tablo 4.6.).

Ardından stats paketinde (20) yer alan "kmeans" fonksiyonu yardımıyla k-ortalamalar kümeleme yöntemi (72), *k* parametresi 3, 4 ve 5 olmak üzere her bir *k* küme sayısı için tek tekrarlı olarak uygulanmıştır.

Kümeleme sonunda, genlerin hangi kümelere atandığına ilişkin detaylı bilgi elde edilmektedir. Örnek olması açısından ilk 6 gene ilişkin sonuçlar Tablo 4.7.'de verilmiştir. Burada sırasıyla *k* parametresi 3, 4 ve 5 iken ilk 6 genin isimleri ile birlikte hangi kümeye düştükleri yani küme üyelikleri görülmektedir.

Tablo 4.7. Lösemi eğitim veri setinden SNFS'nin ikinci adımında k-ortalamalar kümeleme yöntemi ile elde edilen sonuçlara ilişkin örnek tablo.

Gen Adı	Küme Üyeliği (*k*=3 iken)	Küme Üyeliği (*k*=4 iken)	Küme Üyeliği (*k*=5 iken)
M27891_at	2	1	5
U22376_cds2_s_at	1	4	1
M96326_rna1_at	2	1	5
U46751_at	3	2	5
M92287_at	1	4	1
X59417_at	1	4	1

Daha sonra, genlerin birlikte ortaya çıkışlarına ilişkin bu bilgi kullanılarak ağırlıklı komşuluk matrisi oluşturulmuştur. Başka bir deyişle, k-ortalamalar kümeleme yönteminin 3 kez tekrarlanması ve bu tekrarlarda, aynı kümede ortaya çıkan genlerin birlikte ortaya çıkışlarının saydırılması yoluyla elemanlarının aldığı en küçük değerin 0 ve en büyük değerin 3 olduğu bir 72x72 boyutlu kare matris oluşturulmuştur. Dolayısıyla oluşturulan ağırlıklı komşuluk matrisinde, aynı kümede birlikte hiç görülmeyen (*i*,*j*) gen çifti için matris elemanının değeri ilgili satır ve sütunda 0, üç defa da birlikte görülen (*i*,*j*) gen çifti için matris elemanının değeri ilgili satır ve sütunda 3 olmaktadır (Tablo 4.8.).

Ağırlıklı komşuluk matrisi oluşturulurken orijinal SNFS yönteminden (14) farklı olarak, bir genin kendi ile olan etkileşimi yani köşegen değerleri 0 olarak ele alınmıştır. Bu şekilde değişikliğe gidilmesinin nedeni ise yapay olarak oluşturulan ağırlıklı komşuluk matrisinde genin kendi ile aynı kümede ortaya çıkışının biyolojik açıdan herhangi bir anlam ifade etmemesi, başka bir deyişle genin kendi ile etkileşiminin söz konusu olmamasıdır.

Ağırlıklı komşuluk matrisine ek olarak bir sonraki bölümde ağa özel metriklerin hesaplanmasında kullanılacağından ağırlıksız komşuluk matrisi de hesaplanmıştır. Ağırlıksız komşuluk matrisi hesaplanırken, ağırlıklı komşuluk matrisinde 0'dan farklı değere sahip tüm (*i*,*j*) gen çiftleri için ağırlıksız komşuluk matrisinin ilgili satır ve sütununa karşılık gelen matris elemanı 1 olarak yazılır. Ağırlıksız komşuluk matrisindeki diğer tüm matris elemanları ise 0 değerini alır (Tablo 4.8.).

Tablo 4.8. Lösemi eğitim veri setinden SNFS'nin ikinci adımında elde edilen ağırlıklı ve ağırlıksız komşuluk matrislerine ilişkin örnek tablo.

Ağırlıklı Komşuluk Matrisinin İlk 6 Gene İlişkin Bölümü

	M27891_at	U22376_cds2_s_at	M96326_rna1_at	U46751_at	M92287_at	X59417_at
M27891_at	0	0	3	1	0	0
U22376_cds2_s_at	0	0	0	0	3	3
M96326_rna1_at	3	0	0	1	0	0
U46751_at	1	0	1	0	0	0
M92287_at	0	3	0	0	0	3
X59417_at	0	3	0	0	3	0

Ağırlıksız Komşuluk Matrisinin İlk 6 Gene İlişkin Bölümü

	M27891_at	U22376_cds2_s_at	M96326_rna1_at	U46751_at	M92287_at	X59417_at
M27891_at	0	0	1	1	0	0
U22376_cds2_s_at	0	0	0	0	1	1
M96326_rna1_at	1	0	0	1	0	0
U46751_at	1	0	1	0	0	0
M92287_at	0	1	0	0	0	1
X59417_at	0	1	0	0	1	0

Tez kapsamında kullanılan Lösemi veri seti için adım adım elde edilen uygulama sonuçlarının Özyer ve diğ. (14)'nin çalışmasındaki orijinal SNFS sonuçları ile karşılaştırılabilir olması istendiğinden analizin ikinci adımının yani ön işleme ve kümelemenin atlanarak doğrudan korelasyon ağının oluşturulması ile elde edilen sonuçlar adım adım örneklendirilmemiştir. Ancak, SNFS'nin ikinci adımda ön işleme ve kümelemenin atlanarak doğrudan korelasyon ağının oluşturulması yoluyla elde edilen sınıflama başarımına ilişkin sonuçlar da uygulama sonunda paylaşılacaktır.

Adım 3: İkinci adımda uygulanan ön işleme ve kümeleme sonrası elde edilen ağırlıklı komşuluk matrisi kullanılarak sosyal ağ analizi (grafik çizimi ve topluluk belirleme yöntemlerinin çalıştırılması) R yazılımında igraph paketi (94) kullanılarak uygulanmıştır.

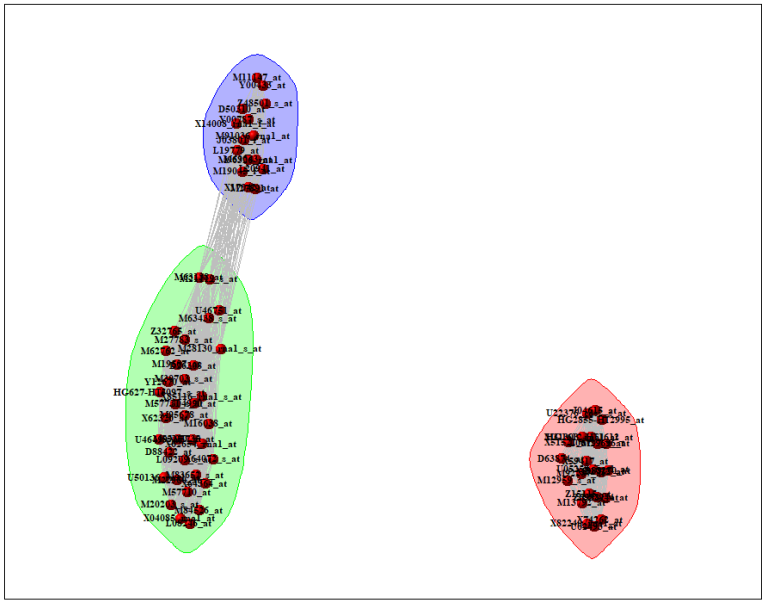
Bu adımda, ağdaki toplulukların belirlenmesi için Özyer ve diğ. (14)'nin de kullanmış olduğu Louvain yöntemi kullanılmıştır. Louvain topluluk belirleme yönteminin uygulanmasında R'da igraph paketinde yer alan "cluster_louvain" fonksiyonundan yararlanılmıştır. Ancak, bu paket kapsamında daha önce bahsedilmiş olan Walktrap ve Infomap gibi topluluk belirleme yöntemleri de yer almaktadır.

R yazılımında, Louvain topluluk belirleme yöntemi sonucunda hangi genlerin hangi topluluklarda yer aldığının listesi de elde edilebilmektedir (Tablo 4.9).

Tablo 4.9. Lösemi eğitim veri setinden SNFS'nin üçüncü adımında Louvain yöntemi ile elde edilen topluluk gen listesi.

Topluluk	Topluluktaki Genlerin Listesi			
1	U22376_cds2_s_at	M92287_at	X59417_at	M13792_at
	U05259_rna1_at	X51521_at	Z69881_at	U02493_at
	HG1612-HT1612_at	M31303_rna1_at	M12959_s_at	M11722_at
	Z15115_at	D88270_at	M29696_at	J04615_at
	HG2855-HT2995_at	L06797_s_at	X74262_at	D63874_at
	U32944_at	X82240_rna1_at		
2	U46751_at	X04085_rna1_at	M28130_rna1_s_at	X95735_at
	M63138_at	M84526_at	J04990_at	M27783_s_at
	D26308_at	L08246_at	M19507_at	U50136_rna1_at
	X85116_rna1_s_at	M62762_at	M20203_s_at	M83652_s_at
	X62654_rna1_at	Y12670_at	M55150_at	X62320_at
	HG627-HT5097_s_at	M22960_at	L09209_s_at	M16038_at
	X64072_s_at	M95678_at	M57731_s_at	M63438_s_at
	X64364_at	D88422_at	Z32765_at	M57710_at
	U46499_at	M30703_s_at	M21119_s_at	
3	M27891_at	M96326_rna1_at	X14008_rna1_f_at	Y00787_s_at
	Z48501_s_at	M19045_f_at	J03801_f_at	M69043_at
	X17042_at	L20941_at	M11147_at	Y00433_at
	D50310_at	L19779_at	M91036_rna1_at	

Topluluk belirleme sonucunda her bir topluluğun farklı renkler ile temsil edildiği sosyal ağ analizi grafiği Şekil 4.1'deki gibi çizdirilmiştir. Bu grafikte *layout* olarak Özyer ve diğ. (14)'nin kullandığı "Force Atlas 2" github'dan elde edilen kod (95) kullanılarak uygulanmıştır. Ancak, daha farklı layout seçenekleri de mevcuttur.



Şekil 4.1. Lösemi eğitim veri setinden SNFS'nin üçüncü adımında elde edilen sosyal ağ analizi sonuçlarına ilişkin ağ grafiği.

Daha sonra, ağa özel metrikler hesaplatılarak bağlılık derecesine göre her topluluğu en iyi temsil edecek olan genlerin %10'u seçilmiştir. Özyer ve diğ. (14)'nin

çalışmasından farklı olarak bağlılık derecesi yapılan epsilon düzeltmesi sayesinde topluluk-dışı ağırlıksız derece merkeziliği 0 olmasına rağmen ilgili gen için bağlılık derecesi tanımsız olarak elde edilmemiş ve özellikle genin kendi topluluğu içindeki bağ sayısının etkisi gözlemlenebilmiştir. SNFS ile Lösemi eğitim veri seti kullanılarak R'da seçilmiş olan 9 gen ve bu genlere ilişkin hesaplanmış olan ağa özel metrikler Tablo 4.10.'da görülmektedir. Alanyazın (14) ile ortak olan genler ise tabloda koyu yazılmıştır.

Tablo 4.10. SNFS ile Lösemi eğitim veri seti kullanılarak seçilmiş olan biyobelirteç genlere ilişkin ağa özel metrikler.

Biyobelirteç Gen	Topluluk Üyeliği	Topluluk Gen Sayısı	z_d	z(i)	D_i	z_{in}	z_{out}	ID	Cov
U22376_cds2_s_at	1	21	210001	21	2,454	21	0	51	0,954
M92287_at	1	21	210001	21	2,454	21	0	66	0,954
X59417_at	1	21	210001	21	2,454	21	0	71	0,954
X04085_rna1_at	2	34	340001	34	2,742	34	0	72	0,971
X95735_at	2	34	340001	34	2,747	34	0	89	0,971
M84526_at	2	34	340001	34	2,743	34	0	101	0,971
J04990_at	2	34	340001	34	2,743	34	0	113	0,971
Z48501_s_at	3	14	140001	14	1,933	14	0	111	0,933
M11147_at	3	14	140001	14	1,933	14	0	252	0,933

R yazılımı ile bu tez çalışmasında yapılan uygulamada, alanyazın ile 5 gen ortak olarak seçilmiş ancak diğer 4 genin farklı olduğu gözlenmiştir. Dolayısıyla, seçilen bu 9 gen ile yapılacak olan sınıflama başarımı hem eğitim hem test seti üzerinde değerlendirilmiş ve alanyazından farklı olarak aşağıdaki sonuçlara ulaşılmıştır.

Tablo 4.11. SNFS ile seçilmiş biyobelirteç genler ile Lösemi eğitim setinde DVM'nin sınıflama başarımı.

	Gerçek Durum		
Kestirim	ALL	AML	Toplam
ALL	27	1	28
AML	0	10	10
Toplam	27	11	38
Doğru Sınıflama Oranı=0,97;Duyarlılık=1;Seçicilik=0,91; AUC=1			

SNFS ile seçilen biyobelirteç genler kullanıldığında radyal tabanlı DVM ile eğitim setinde yalnız tek bir örnek yanlış sınıflanmıştır (Bkz. Tablo 4.11.). Ancak, önemli olan test veri setindeki başarımdır ve Tablo 4.12.'ye bakıldığında 7129 gen içerisinden yalnız SNFS ile seçilmiş olan 9 gen kullanılarak radyal tabanlı DVM'den elde edilen sınıflama başarımının oldukça yüksek olduğu söylenebilir.

Tablo 4.12. SNFS ile seçilmiş biyobelirteç genler ile Lösemi test setinde DVM'nin sınıflama başarımı.

	Gerçek Durum		
Kestirim	ALL	AML	Toplam
ALL	20	1	21
AML	0	13	13
Toplam	20	14	34
Doğru Sınıflama Oranı=0,97;Duyarlılık=1;Seçicilik=0,93; AUC=0,99			

Özyer ve diğ. (14) ile her ne kadar birbirine çok yakın sınıflama başarımları elde edilmiş olsa da R'da uygulanan SNFS yöntemi ile seçilen genler kullanıldığında, özellikle eğitim setinde DVM'den daha iyi bir sınıflama başarımı elde edilmiştir. Alanyazında doğru sınıflama oranı dışında başka bir başarım ölçüsü verilmediğinden daha detaylı bir karşılaştırma yapmak olanaklı değildir.

Aynı zamanda, 7129 genin tamamı kullanıldığında DVM'den elde edilecek sınıflama başarımı da daha önce elde edilen bulgular ile karşılaştırılmıştır.

Tablo 4.13. Tüm genler ile Lösemi eğitim setinde DVM sınıflama başarımı.

	Gerçek Durum		
Kestirim	ALL	AML	Toplam
ALL	27	0	27
AML	0	11	11
Toplam	27	11	38
Doğru Sınıflama Oranı=1;Duyarlılık=1;Seçicilik=1; AUC=1			

Böylece, yüksek boyutlu veri yapısının neden olduğu aşırı uyum gibi sorunların ne boyutta etkili olduğu gözlemlenmiştir. Örneğin, eğitim setinden her ne kadar iyi bir

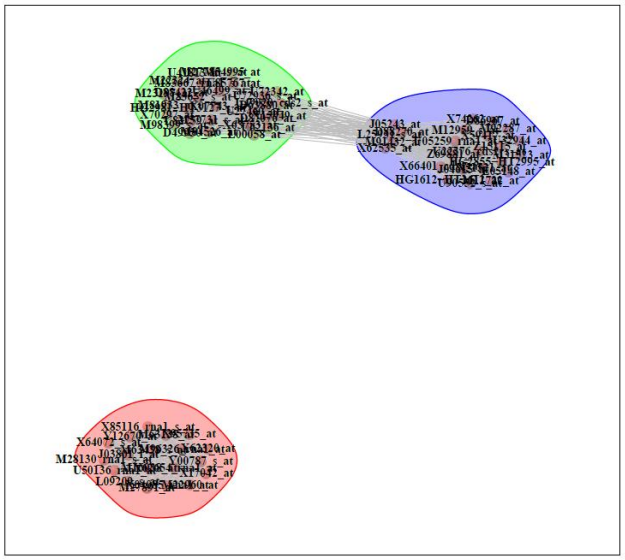
sınıflama başarımı elde edilmişse de yüksek boyutluluğun neden olduğu aşırı uyum sebebiyle test setindeki sınıflama başarımı düşmüştür (Tablo 4.14.).

Tablo 4.14. Tüm genler ile Lösemi test setinde DVM sınıflama başarımı.

	Gerçek Durum		
Kestirim	ALL	AML	Toplam
ALL	18	4	22
AML	2	10	12
Toplam	20	14	34
Doğru Sınıflama Oranı=0,82;Duyarlılık=0,90;Seçicilik=0,71; AUC=0,92			

7129 genin tamamının kullanılmasının yüksek boyut nedeniyle sınıflama başarımı üzerinde olumsuz etkisinin olduğu ve SNFS yönteminin boyut indirgeme yoluyla özellikle test setinde doğru sınıflama oranını yükselttiği söylenebilir (Bkz. Tablo 4.12., Tablo 4.14.).

Gen ifade düzeyleri dönüştürülmüş olan Lösemi veri seti kullanılarak da SNFS uygulanmıştır. Tez kapsamında adım adım elde edilen bütün sonuçlar verilmemiş ancak SNFS'nin üçüncü adımında elde edilen ağ grafiği Şekil 4.2.'de gösterilmiştir. Bu grafik için *layout* "layout_nicely" olarak seçilmiştir. Bu seçenek ile uygun layout grafiğe otomatik olarak atanmaktadır.



Şekil 4.2. Dönüşüm uygulanmış Lösemi eğitim veri setinden SNFS'nin üçüncü adımında elde edilen sosyal ağ analizi sonuçlarına ilişkin ağ grafiği.

Dönüşüm uygulanmış Lösemi veri seti için, SNFS ile elde edilen biyobelirteç genler kullanılarak DVM'den elde edilen sınıflama başarımı Tablo 4.15. ve Tablo 4.16.'daki gibi elde edilmiştir.

Tablo 4.15. SNFS ile seçilmiş biyobelirteç genler ile dönüşüm uygulanmış Lösemi eğitim setinde DVM'nin sınıflama başarımı.

	Gerçek Durum		
Kestirim	ALL	AML	Toplam
ALL	27	0	27
AML	0	11	11
Toplam	27	11	38
Doğru Sınıflama Oranı=1;Duyarlılık=1;Seçicilik=1; AUC=1			

Tablo 4.16. SNFS ile seçilmiş biyobelirteç genler ile dönüşüm uygulanmış Lösemi test setinde DVM'nin sınıflama başarımı.

	Gerçek Durum		
Kestirim	ALL	AML	Toplam
ALL	19	1	20
AML	1	13	14
Toplam	20	14	34
Doğru Sınıflama Oranı=0,94;Duyarlılık=0,95;Seçicilik=0,93; AUC=0,99			

SNFS ile Lösemi veri seti için seçilen biyobelirteç genler kullanıldığında DVM'den elde edilecek sınıflama başarımı üzerinde dönüşümün ciddi bir etkisi olmamıştır. Hem doğru sınıflama oranında hem de diğer başarım ölçülerinde dönüşüm beklenenin aksine belirgin bir yükselme sağlayamamıştır. Dolayısıyla, bu veri seti için dönüşümün sınıflama başarımını arttırdığını söylemek olanaklı değildir (Bkz. Tablo 4.11., Tablo 4.12., Tablo 4.15., Tablo 4.16.).

Her ne kadar eğitim setinde doğru sınıflama oranı yükselmişse de test setinde başarım ölçülerinin hiç birinde ciddi bir artış söz konusu olamamıştır. Bu durum dönüşüm uygulanmış veri ile birlikte tüm genlerin kullanılmasında da benzerdir. Dönüşüm burada da sınıflama başarımı üzerinde ciddi bir artış sağlayamamıştır (Tablo 4.17., Tablo 4.18.).

Tablo 4.17. Tüm genler ile dönüşüm uygulanmış Lösemi eğitim setinde DVM sınıflama başarımı.

	Gerçek Durum		
Kestirim	ALL	AML	Toplam
ALL	27	0	27
AML	0	11	11
Toplam	27	11	38
Doğru Sınıflama Oranı=1;Duyarlılık=1;Seçicilik=1; AUC=1			

Tablo 4.18. Tüm genler ile dönüşüm uygulanmış Lösemi test setinde DVM sınıflama başarımı.

	Gerçek Durum		
Kestirim	ALL	AML	Toplam
ALL	18	1	19
AML	2	13	15
Toplam	20	14	24
Doğru Sınıflama Oranı=0,91;Duyarlılık=0,90;Seçicilik=0,93; AUC=0,99			

Dönüşüm uygulanmış Lösemi veri setinde, tüm genler ile uygulanan DVM'nin sınıflama başarımında ciddi bir artış gözlemlenmemiştir (Bkz. Tablo 4.15., Tablo 4.16., Tablo 4.17., Tablo 4.18.). Dönüşümün etkisinin olumlu olmamasının nedeni büyük oranda Dudoit ve diğ. (32)'nin çalışmasında kullanılan veri ile golubEsets paketi kapsamında yer alan verinin birebir aynı olmaması olabilir. Bu probleme, yani aynı isme sahip verilerin geçirdiği ön işleme uygulamalarının, içerdiği özellik ve/veya örnek sayısının vb. farklı olması ve bu nedenle objektif karşılaştırma yapmanın zorluğuna daha önce alanyazında da değinilmiştir (19). golubEsets paketi kapsamında yer alan veriye uygulanan ön işleme süreçleriyle ilgili ayrıntılı bilgi alanyazında bulunamamıştır. Bu nedenle, tez çalışmasının ilerleyen bölümünde Dudoit ve diğ. (32)'nin önerdiği dönüşümün yapılmadığı veri seti kullanılmıştır.

Lösemi veri seti kullanılarak hazırlanan R fonksiyonu ile SNFS'nin adımlarında farklı yöntemlerin seçimi söz konusu olmuştur. Elde edilen tüm sonuçlara ilişkin bilgiler Tablo 4.19.'da yer almaktadır. Bu tabloda dönüştürülmemiş Lösemi verisi kullanılmış, birinci adımda özellik seçim yöntemlerinden elde edilen tümel

sıralamalar ile %1 en iyi gen seçilmiş, ikinci adımda ise k-ortalamalar kümeleme yönteminin *k*=3,4 ve 5 için tek tekrarı ile ağırlıklı komşuluk matrisi elde edilmiştir. Topluluk belirleme sonrasında her topluluğu temsil eden en iyi %10 gen seçilmiştir.

Sonuç olarak, gömülü yöntemlerden DVM K-YÖE ve filtrelerden IG ya da CS filtresinin özellik seçim adımında birleştirilmesi ve topluluk belirleme yöntemlerinden Walktrap yönteminin kullanılması sınıflama başarımında daha iyi sonuç elde edilmesini sağlamıştır (Tablo 4.19.).

Tablo 4.19. SNFS'nin adımlarında farklı algoritma kombinasyonları ile Lösemi veri seti için seçilmiş biyobelirteç genler kullanılarak test setinde DVM'den elde edilen sınıflama başarımları.

Özellik Seçimi	Topluluk Belirleme	Topluluk Sayısı	Doğru Sınıflama Oranı	Duyarlılık	Seçicilik	AUC
DVM K-YÖE+IG	Louvain	2	0,9412	1,00	0,8571	1,0000
DVM K-YÖE+CS		3	0,9706	1,00	0,9286	0,9964
DVM K-YÖE+RF		3	0,8823	0,90	0,8571	0,9357
IG+CS		3	0,8529	1,00	0,6429	0,9679
IG+RF		3	0,8235	0,90	0,7143	0,9018
CS+RF		3	0,9412	1,00	0,8571	0,9929
DVM K-YÖE+IG	Walktrap	3	0,9412	1,00	0,8571	1,0000
DVM K-YÖE+CS		3	0,9706	1,00	0,9286	0,9964
DVM K-YÖE+RF		3	0,8823	0,90	0,8571	0,9357
IG+CS		5	0,8235	1,00	0,5714	0,9714
IG+RF		4	0,8529	0,95	0,7143	0,9321
CS+RF		4	0,9412	0,95	0,9286	0,9964
DVM K-YÖE+IG	Infomap	2	0,9412	0,95	0,9286	0,9750
DVM K-YÖE+CS		3	0,9706	1,00	0,9286	0,9964
DVM K-YÖE+RF		3	0,8823	0,90	0,8571	0,9357
IG+CS		3	0,8529	1,00	0,6429	0,9679
IG+RF		3	0,8235	0,90	0,7143	0,9018
CS+RF		2	0,9412	1,00	0,8571	1,0000

Bu nedenle, SNFS'nin ikinci adımının atlanmasını sağlayan korelasyon ağı oluşturma yaklaşımı için bu sonuçlar dikkate alınmıştır. Böylece, DVM K-YÖE gömülü yöntemi ve CS filtresinden elde edilen sıralamalar kullanılarak genler indirgenmiştir. Sonrasında ise korelasyon ağının oluşturulmuş ve genler arası

korelasyonlardan yararlanılarak elde edilen ağırlıklı komşuluk matrisi yardımıyla uygulanan sosyal ağ analizinde Walktrap yöntemi ile topluluklar belirlenmiştir (Tablo 4.20.). Benzer şekilde burada da ağa özel metriklerden bağlılık derecesi, her topluluğu temsil edebilecek en iyi %10 genin belirlenmesinde kullanılmıştır. Buradan elde edilen sonuçlar ise Tablo 4.20.'de yer almaktadır.

Tablo 4.20. SNFS'nin ikinci adımı atlanarak Lösemi veri seti için seçilmiş biyobelirteç genler kullanılarak test setinde elde edilen DVM sınıflama başarımı.

Özellik Seçimi	Topluluk Belirleme	Topluluk Sayısı	Doğru Sınıflama Oranı	Duyarlılık	Seçicilik	AUC
DVM K-YÖE+CS	Walktrap	2	0,8235	0,85	0,7857	0,9393

Görüldüğü üzere bu veri seti için, ikinci adımda kümeleme yaklaşımının kullanılması bu adımın atlanarak korelasyon ağı oluşturulması ile elde edilenden daha iyi bir sınıflama başarımı sağlamıştır (Bkz. Tablo 4.19., Tablo 4.20.). Bunun nedeni, birinci adımda filteleme yapılması nedeniyle ölçekten-bağımsız ağ yapısının bozulmaya uğraması ve bunun sonucunda, korelasyon ağının gerçekte var olan korelasyon yapısından uzaklaşması ile topluluk belirleme sürecinde genler arası gerçekte olan etkileşimin yakalanamaması olabilir (13).

4.2. Kolon Kanseri Veri Seti Üzerinde Uygulama Sonuçları

Kolon kanseri veri setinde (87, 88) yer alan her bir gene ilişkin gen ifade düzeyleri, ortalaması 0, standart sapması 1 olacak şekilde normalleştirildikten (87) sonra verinin %60'ı eğitim, %40'ı test seti olarak ayrılmıştır.

SNFS yönteminin özellik seçim adımında farklı özellik seçim yöntemleri ile elde edilen tümel sıralamaya göre genlerin en iyi %10'u seçilmiştir. SNFS'nin ikinci adımında yine benzer şekilde k=3,4 ve 5 için tek tekrarlı olarak k-ortalamalar kümeleme yöntemi uygulanmış ve buradan genlerin birlikte ortaya çıkışları elde edilerek ağırlıklı ve ağırlıksız komşuluk matrisleri oluşturulmuştur. SNFS'nin üçüncü adımında farklı topluluk belirleme yöntemleri ile ağdaki toplulukların belirlenmesi sonrası ise her toplululuğu bağlılık derecesine göre temsil edebilecek en iyi %5 genin seçilmesi yoluyla biyobelirteç genler elde edilmiştir. Tablo 4.21.'de SNFS yönteminin

adımlarında kullanılan farklı kombinasyonları ile seçilen biyobelirteç genler kullanılarak test setinde DVM'den elde edilen sınıflama başarımına ilişkin özet bilgiler görülmektedir.

Tablo 4.21. SNFS adımlarındaki farklı algoritma kombinasyonları ile Kolon kanseri veri seti için seçilmiş biyobelirteç genler kullanılarak test setinde elde edilen DVM sınıflama başarımları.

Özellik Seçimi	Topluluk Belirleme	Topluluk Sayısı	Doğru Sınıflama Oranı	Duyarlılık	Seçicilik	AUC
DVM K-YÖE+IG	Louvain	2	0,84	0,8667	0,80	0,8733
DVM K-YÖE+CS		2	0,76	0,8667	0,60	0,9000
DVM K-YÖE+RF		3	0,76	0,8667	0,60	0,8467
IG+CS		3	0,72	0,8667	0,50	0,8400
IG+RF		3	0,80	0,8667	0,70	0,8667
CS+RF		3	0,76	0,8667	0,60	0,9067
DVM K-YÖE+IG	Walktrap	2	0,84	0,8667	0,80	0,8733
DVM K-YÖE+CS		2	0,76	0,8667	0,60	0,9000
DVM K-YÖE+RF		3	0,76	0,8667	0,60	0,8467
IG+CS		3	0,72	0,8667	0,50	0,8400
IG+RF		3	0,80	0,8667	0,70	0,8667
CS+RF		3	0,76	0,8667	0,60	0,9067
DVM K-YÖE+IG	Infomap	3	0,80	0,8667	0,70	0,7533
DVM K-YÖE+CS		3	0,76	0,8667	0,60	0,8533
DVM K-YÖE+RF		3	0,76	0,8667	0,60	0,8467
IG+CS		3	0,72	0,8667	0,50	0,8400
IG+RF		3	0,80	0,8667	0,70	0,8667
CS+RF		3	0,76	0,8667	0,60	0,9067

SNFS yönteminin ilk adımında boyut indirgeme için DVM K-YÖE ile IG kombinasyonunun, topluluk belirleme için ise Louvain ya da Walktrap topluluk belirleme yöntemlerinden herhangi birinin kullanımı ile seçilen biyobelirteç genler kullanılarak DVM ile iyi bir sınıflama başarımı elde edildiği söylenebilir. Çünkü verideki tüm genlerin kullanılması ile test setinde DVM parametrelerinin en iyilendiği ve en iyi sınıflama başarımını veren doğrusal çekirdek fonksiyonun kullanıldığı durumda bile sırasıyla doğru sınıflama oranı, duyarlılık, seçicilik ve eğri altında kalan alan %81,25, %90, %67, %86,25 olarak elde edilmiştir.

Genel olarak Kolon kanseri veri seti için, SNFS yöntemi kullanıldığında seçilen 11 gen ile DVM'den elde edilebilecek sınıflama başarımının 2000 genin tamamının kullanılmasıyla elde edilen sınıflama başarımından özellikle eğri altında kalan alan dikkate alındığında daha iyi bir sonuç elde edildiği görülmüştür.

4.3. Sentetik Veri Seti Üzerinde Uygulama Sonuçları

Madelon veri seti 5 tane ilgili özellik olmak üzere toplamda 500 özellik ve 2000 örnek içeren iki sınıflı bir veri setidir (45). Bu veri setinin ilk olarak %60'ı eğitim, %40'ı test seti olarak hazırlanan bir fonksiyon yardımıyla ayrılmıştır.

SNFS'nin ilk adımında, tez kapsamına alınan farklı filtre kombinasyonları değiştirilerek ve bu adımda tümel sıralamaya göre genlerin %10'u seçilerek uygulama yapılmıştır. İkinci adımda, k-ortalamalar kümeleme yönteminin, küme sayısı k=3, 4 ve 5 olmak üzere her k küme sayısı için tek tekrarlı olarak uygulanması ile ağırlıklı ve ağırlıksız komşuluk matrisleri elde edilmiştir. Topluluk belirleme adımında ise Walktrap topluluk belirleme yöntemi kullanılarak bağlılık derecesine göre her topluluğu temsil ettiği düşünülen %10 gen biyobelirteç olarak seçilmiştir. Elde edilen biyobelirteçlerin DVM'de sağladığı sınıflama başarımına ilişkin sonuçlar sırasıyla Tablo 4.22., Tablo 4.23., Tablo 4.24. ve Tablo 4.25.'te görüldüğü gibidir.

Tablo 4.22. SNFS'nin ilk adımında CS ve IG filtrelerinin kullanılmasıyla seçilmiş biyobelirteç genler ile Madelon test setinde DVM'nin sınıflama başarımı.

	Gerçek Durum		
Kestirim	-1	+1	Toplam
-1	338	178	516
+1	193	331	524
Toplam	531	509	1040
Doğru Sınıflama Oranı=0,64;Duyarlılık=0,64;Seçicilik=0,65; AUC=0,69			

SNFS'nin ilk adımında kullanılan özellik seçim yöntemleri arasında en hızlı sonuç veren yöntemlerden olan CS ve IG filtrelerinin SNFS'nin ilk adımında birlikte kullanılmasıyla elde edilen sınıflama tablosu yukarıdaki gibidir. Toplam 7 özellik biyobelirteç olarak seçilmiştir. Bu 7 özellik ile eğitim setinde doğru sınıflama oranı yaklaşık %68 iken test setinde bu oran az miktarda düşmüş ve %64 olarak elde edilmiştir. Duyarlılık, seçicilik ve eğri altında kalan alan da test setinde (Bkz. Tablo 4.22.) daha düşük bulunmuştur (eğitim setinde sırasıyla %68,08; %67,5; %74,5).

Lösemi veri setinde en iyi başarım gösteren üç kombinasyondan biri olan CS+RF kombinasyonuna ilişkin ayrıntılı sonuçlar ise Tablo 4.23.'te verilmiştir.

Tablo 4.23. SNFS'nin ilk adımında CS ile RF kombinasyonunun kullanılmasıyla seçilmiş biyobelirteç genler ile Madelon test setinde DVM'nin sınıflama başarımı.

	Gerçek Durum		
Kestirim	-1	+1	Toplam
-1	428	131	559
+1	103	378	481
Toplam	531	509	1040
Doğru Sınıflama Oranı=0,77;Duyarlılık=0,81;Seçicilik=0,74; AUC=0,84			

CS ve RF filtre kombinasyonuyla Walktrap sonrası elde edilen topluluk sayısı 3'tür. SNFS sonrası toplam 7 özellik seçilmiştir. Bu 7 özellik ile eğitim setinde doğru sınıflama oranı %80 iken test setinde bu oran %77'ye düşmüştür. Benzer şekilde, duyarlılık, seçicilik ve eğri altında kalan alan da test setinde (Bkz. Tablo 4.23.) daha düşük bulunmuştur (eğitim setinde sırasıyla %81,53; %78,51; %87,7).

Kombinasyonlar arasında karşılaştırmayı daha kolay yapabilmek adına Madelon veri setine ilişkin aynı eğitim ve test seti kullanılarak elde edilen bulguların tamamı Tablo 4.24.'te yer almaktadır.

Tablo 4.24.'te görüldüğü üzere tüm kombinasyonlar ile referans doğru sınıflama oranından daha iyi bir doğru sınıflama oranı elde edilmiştir. Ancak, DVM K-YÖE ile CS; CS ile IG ve RF kombinasyonları ile daha iyi bir sınıflama başarımı elde edilmiştir. En önemlisi, 500 özellikten indirgenmiş az sayıdaki özellik (6-7 arası) ile referans doğru sınıflama oranı olan %50,13'ün üzerine ciddi şekilde çıkılmıştır.

Tablo 4.24. SNFS ile Madelon veri seti için seçilmiş biyobelirteç genler kullanılarak test setinde DVM'den elde edilen sınıflama başarımları.

Özellik Seçimi	Topluluk Belirleme	Topluluk Sayısı	Biyobelirteç Sayısı	Doğru Sınıflama Oranı	Duyarlılık	Seçicilik	AUC
DVM K-YÖE+CS	Walktrap	3	6	0,667	0,727	0,605	0,72
DVM K-YÖE+IG		4	7	0,556	0,559	0,552	0,58
CS+IG		3	7	0,643	0,636	0,650	0,69
CS+RF		3	7	0,775	0,806	0,743	0,84

Yöntemin sınıflama başarımını bir başka yaklaşım ile değerlendirebilmek amacıyla, ilgili veri seti (eğitim ve test birlikte) ile yeniden örnekleme (n=25, 1000

iterasyon) yapıldığında aynı yeniden örnekleme verisi üzerinde elde edilmiş DVM'ye ilişkin sınıflama başarım sonuçları Tablo 4.25.'te verilmiştir. Tablo 4.25.'te 1000 iterasyondan elde edilen başarım ölçüsüne ilişkin ortalama ve standart hatalar (parantez içinde) yer almaktadır.

Yeniden örnekleme sonuçlarına bakıldığında sınıflama başarımının SNFS ile elde edilen biyobelirteç genler kullanıldığında kabul edilebilir düzeyde olduğu söylenebilir (Tablo 4.25.).

Tablo 4.25. Madelon veri seti için yeniden örneklemede SNFS kullanılarak seçilmiş biyobelirteç genler kullanılarak test setinden elde edilen DVM sınıflama başarımları.

Özellik Seçimi	Topluluk Belirleme	Doğru Sınıflama Oranı	Duyarlılık	Seçicilik	AUC
DVM K-YÖE+CS	Walktrap	0,9454 (0,10504)	0,9365 (0,14629)	0,9394 (0,15401)	0,9927 (0,01456)
DVM K-YÖE+IG		0,9451 (0,10507)	0,9362 (0,14311)	0,9397 (0,15166)	0,9925 (0,01495)
CS+IG		0,9140 (0,17481)	0,9020 (0,21847)	0,9060 (0,22020)	0,9920 (0,01564)
CS+RF		0,9317 (0,13837)	0,9237 (0,18625)	0,9189 (0,19033)	0,9928 (0,01429)

Alanyazında belirtilen referans doğru sınıflama oranı bu veri seti için %50,13 (45) iken denenen tüm özellik seçim yöntemi kombinasyonlarında SNFS ile bu oran ciddi şekilde yükseltilebilmiştir (Bkz. Tablo 4.24., Tablo 4.25.). Ancak, DVM K-YÖE ve CS ya da IG filtre kombinasyonlarını içeren SNFS ile seçilen biyobelirteçler kullanılarak yapılacak bir DVM sınıflama başarımının daha yüksek olduğu söylenebilir.

4.4. Benzetim Çalışmasına İlişkin Sonuçlar

Benzetim çalışması için veri üretiminde Guo ve diğ. (42) esas alındığından, iki grup bağımlılık yapısına uygun olarak veri üretilmiştir. İki sınıfın söz konusu olduğu durum için, her sınıfta eşit sayıda olmak üzere toplamda 200 eğitim ve 600 test örneği ile p=10000 gen üretilmiştir. 10000 gen, her biri 100 gen içeren k=100 bloğa

bölünmüştür. Farklı bloklardaki genlerin birbirlerinden bağımsız, aynı bloktaki genlerin ise otoregresif yapı olarak ele alınan ve Eşitlik 3.11.'de gösterilen kovaryans yapısına uygun şekilde ilişkili olduğu varsayılmıştır. Aynı blok içinde $|\rho|=0,9$ alınmıştır (eğitim setindeki 50 blok için pozitif, 50 blok için ise negatif). İlk olarak, her genin ifade düzeyi standart normal dağımdan üretilmiş daha sonra yukarıda tanımlanan kovaryans matrisinin kare kökü ile çarpılan ifade düzeyleri böylece dönüşüme uğramış ve temelde $MVN(0,\Sigma)$ ile çok değişkenli normal dağılıma uymuştur. Daha sonra sadece ikinci sınıftaki ilk 200 gen ifade düzeyine 0,5 sabiti eklenmiştir.

Aşağıda, SNFS yönteminin ilk adımda %1 en iyi genin tabloda belirtilen özellik seçim yöntemi kombinasyonlarından elde edilen tümel sıralamalar kullanılarak seçildiği ve tabloda belirtilen topluluk seçim yönteminin kullanımıyla elde edilen toplulukları bağlılık derecesine göre temsil edebilecek en iyi %5 genin biyobelirteç olarak seçildiği durumda DVM'nin test setinde gösterdiği sınıflama başarımına ilişkin sonuçlar yer almaktadır (Tablo 4.26.).

Tablo 4.26. İki grup bağımlılık yapısına göre $|\rho|=0,90$ için üretilen veri ile SNFS yönteminden elde edilen benzetim sonuçları.

Özellik Seçimi	Topluluk Belirleme	Doğru Sınıflama Oranı	Duyarlılık	Seçicilik	AUC
DVM K-YÖE+CS	Walktrap	0,635	0,657	0,6133	0,6620
DVM K-YÖE+IG		0,682	0,690	0,6733	0,7447
CS+IG		0,650	0,647	0,6533	0,6949
CS+RF		0,650	0,647	0,6533	0,6949
DVM K-YÖE+CS	Infomap	0,658	0,673	0,6433	0,7077
DVM K-YÖE+IG		0,682	0690	0,6733	0,7447
CS+IG		0,650	0,647	0,6533	0,6949
CS+RF		0,648	0,650	0,6467	0,6949
DVM K-YÖE+CS	Louvain	0,583	0,597	0,5700	0,6265
DVM K-YÖE+IG		0,682	0,690	0,6733	0,7447
CS+IG		0,653	0,617	0,6900	0,7194
CS+RF		0,673	0,670	0,6767	0,7176

Benzetim çalışması sonuçlarına göre, SNFS yöntemi ile 10000 özellik yani gen (yapay) arasından 6 ile 8 arasında gen seçilerek elde edilen DVM sınıflama başarımı blok içi yüksek korelasyon yapısına sahip genlerin söz konusu durumda orta seviyededir (Bkz. Tablo 4.26.). Hem hız hem de sınıflama başarımında yarattığı artış açısından SNFS'nin özellik seçim adımında CS ve IG filtrelerinin birlikte kullanılması ve topluluk belirleme yöntemlerinden Louvain'ın seçilmesi yeterli olacaktır. Ancak, topluluk belirleme yöntemlerinden Walktrap ve Infomap tüm özellik seçim kombinasyonları ile elde edilen indirgenmiş genler için daha kararlı sonuçlar üretmektedirler (Bkz. Tablo 4.26.).

Aynı zamanda, genler arası korelasyon yapısı değiştirilerek genler arası korelasyon yapısının etkisi gözlemlenmiştir. Bu nedenle, aynı blokta yer alan genler arası korelasyonun daha düşük olduğu ($|\rho|=0,60$) durum için de veri üretim süreci benzer şekilde yenilenmiş ve benzetim çalışması sonucunda test setinden aşağıdaki sonuçlar elde edilmiştir.

Tablo 4.27. İki grup bağımlılık yapısına göre $|\rho|=0,60$ için üretilen veri ile SNFS yönteminden elde edilen benzetim sonuçları.

Özellik Seçimi	Topluluk Belirleme	Doğru Sınıflama Oranı	Duyarlılık	Seçicilik	AUC
DVM K-YÖE+CS	Walktrap	0,657	0,680	0,6333	0,7319
DVM K-YÖE+IG		0,640	0,627	0,6533	0,7098
CS+IG		0,645	0,677	0,6133	0,6969
CS+RF		0,648	0,650	0,6467	0,7007
DVM K-YÖE+CS	Infomap	0,662	0,670	0,6533	0,7208
DVM K-YÖE+IG		0,640	0,627	0,6533	0,7098
CS+IG		0,658	0,613	0,7033	0,7433
CS+RF		0,697	0,703	0,6900	0,7556
DVM K-YÖE+CS	Louvain	0,657	0,680	0,6333	0,7319
DVM K-YÖE+IG		0,640	0,627	0,6533	0,7098
CS+IG		0,642	0,640	0,6433	0,6854
CS+RF		0,648	0,650	0,6567	0,7007

Benzetim çalışması sonuçlarına göre, SNFS yöntemi ile 10000 özellik yani gen (yapay) arasından 6 ile 8 arasında gen seçilerek elde edilen DVM sınıflama başarımı

blok içi orta düzey korelasyon yapısına sahip genlerin söz konusu durumda blok içi yüksek düzey korelasyon yapısına sahip genlerden elde edilen sonuçlara göre daha iyidir (Bkz. Tablo 4.26., Tablo 4.27.). Genel olarak, aynı blok içinde yer alan genler arası korelasyonun artmasının sınıflama başarımı üzerinde düşük düzeyde olumsuz bir etkiye sahip olduğunu ancak SNFS yöntemi ile 10000 genden 6 ile 8 arasında gen seçilerek hem ciddi oranda boyut indirgenebildiği hem de kabul edilebilir düzeyde iyi bir sınıflama başarımına ulaşıldığı söylenebilir.

Aynı blok içindeki genler arası korelasyon $|\rho|=0,60$ olduğunda 10000 gene ilişkin verinin doğrudan kullanılmasıyla DVM'den (farklı çekirdek fonksiyonlar için parametre en iyileştirmesi yapıldığında) elde edilebilen en yüksek sınıflama başarımı, doğrusal çekirdek fonksiyonunun kullanıldığı durumda yakalanmıştır ve doğru sınıflama oranı, duyarlılık, seçicilik ve eğri altında kalan alan sırasıyla %82,67, %80, %85 ve %90,5 olmuştur; aynı bloktaki genler arası korelasyon $|\rho|=0,90$ olduğunda ise DVM'den elde edilen sınıflama başarımı düşerek doğru sınıflama oranı, duyarlılık, seçicilik ve eğri altında kalan alan sırasıyla %65, %61, %69 ve %70,6 olmuştur. Buna göre özellikle, aynı blok içindeki genler arası korelasyon yüksek düzeyde olduğunda SNFS yönteminin boyut indirgeme açısından tercih edilebilecek bir yöntem olduğu görülmektedir.

5. TARTIŞMA

İlişkisel veri tabanlarından veya büyük veri yığınlarından elde edilen karmaşık sonuçları görselleştirme olanağı nedeniyle, sosyal ağ analizinin tıp ve genetik alanında da kullanımı giderek yaygınlaşmaktadır. Özellikle son yıllarda boyut indirgenmesi ve en ilgili özelliklerin seçimi yoluyla sınıflama başarımını arttıracak olan hastalığa özgü biyobelirteçlerin belirlenmesi amacıyla yapılan çalışmalarda, sosyal ağ analizinin kullanımı ilgi uyandırmakta ve hastalığa özgü biyobelirteçlerin belirlenmesi problemine sosyal ağ analizinin de yer aldığı melez yaklaşımlarla farklı bir bakış açısı getirilmektedir.

Ancak, melez yöntemler kapsamında yer alan birden çok makine öğrenmesi, veri madenciliği ya da sosyal ağ analizi yönteminin çok sayıda farklı kombinasyonunun söz konusu olması nedeniyle problem çözümü için bu yöntemleri uygun biçimde birleştirmenin ve probleme özgü en uygun çözüm arayışının önemi ortaya çıkmıştır (19).

Bu tez çalışmasında, genomik bir ağın analizinde kullanılan SNFS melez biyobelirteç belirleme yaklaşımının (14) boyut indirgeme, kümeleme, topluluk belirleme gibi farklı aşamalarında kullanılan makine öğrenmesi, veri madenciliği ve sosyal ağ analizi yöntemlerinin incelenmesinin yanı sıra SNFS kapsamında kullanılabilecek bazı farklı yöntemler de değerlendirilmiş ve R yazılımında yani tek bir platform kullanılarak tezde kullanılan erişime açık genomik mikrodizi veri setleri üzerinde bu farklı kombinasyonların etkisi incelenmiştir.

Tez kapsamında, ilk olarak platform ve/veya veri farklılıklarının etkisini görmek amacıyla Özyer ve diğ. (14)'nin çalışmasında SNFS yöntemi ile Lösemi veri seti için elde ettikleri biyobelirteç genler kullanılmış ve doğrudan DVM'nin sınıflama başarımına ilişkin sonuçlar elde edilmiştir. DVM kullanılırken tüm çekirdek fonksiyon seçenekleri (doğrusal, radyal, polinomiyal, sigmoid) denenmiş ve alanyazından farklı olarak en iyi sınıflama başarımını veren radyal tabanlı DVM'nin sınıflama başarımı ayrıntılı olarak değerlendirilmiştir. Böylece, aynı 9 gen kullanılarak alanyazınla karşılaştırma yapıldığında R'da farklı sonuçlar elde edildiği görülmüş ve eğitim setinde daha başarılı sonuçlar elde edilmesine rağmen test setinde benzer bir sınıflama başarımı elde edilmiştir. Bu farklılıklar, platform farklılıklarından kaynaklanabileceği

gibi alanyazında da bahsedilen (19) aynı isme sahip ancak farklı yapıdaki veri setlerinin söz konusu olması nedeniyle de ortaya çıkabilmektedir. Bu nedenle özellik seçim yöntemlerinin gerçek genomik veri setleri kullanılarak karşılaştırılmasında objektif karşılaştırma yapma zorluğunun olduğu söylenebilir (19).

Bu uygulamanın ardından SNFS yönteminin R yazılımında hazırlanan fonksiyon ile işletilmesine geçilmiştir. Alanyazında yer alan SNFS yönteminin (14) R yazılımında uygulanmasıyla birlikte her ne kadar yakın sınıflama başarımları elde edilmiş olsa da R yazılımında uygulanan yöntem eğitim setinde daha iyi sonuç vermiştir. Aynı zamanda, alanyazındaki kimi çalışmalara göre daha az gen sayısı ile (30) daha iyi ya da oldukça yakın (42) sınıflama başarımına da ulaşılmıştır. Doğrudan 7129 gen kullanılarak uygulanan DVM sonucunda elde edilen sınıflama başarımına göre DVM sınıflama başarımını yükselten SNFS yönteminin özellikle test setinde doğru sınıflama oranını yükselttiği ve bu nedenle başarılı bir melez özellik seçim yöntemi olduğu söylenebilir.

Tez kapsamında kullanılan yöntem kombinasyonlarının değiştirilmesi yoluyla Lösemi veri seti için farklı sonuçlar elde edilmiştir. DVM-YÖE gibi hesaplama açısından daha uzun süre gerektiren gömülü yöntemlerin yanı sıra hızlı hesaplama süresiyle birlikte yine de sınıflama başarımı üzerinde olumlu etkisi olan CS ve IG gibi filtre kombinasyonları da söz konusu olmuştur. Buna göre, araştırmacılar Lösemi veri seti için IG+RF kombinasyonu dışında bu çalışma kapsamında denenmiş diğer tüm kombinasyonlar ile gen seçerek tüm genlerin kullanılmasından daha iyi bir DVM sınıflama başarımı elde edebilirler. Ancak bu veri seti için IG+RF kombinasyonuyla DVM'nin sınıflama başarımında bir artış elde edilememiştir.

Tez çalışmasının ikinci aşamasında, sentetik veri seti üzerinde SNFS kapsamında yer alan yöntemlerin biyobelirteç belirleme problemi açısından karşılaştırılması amaçlanmıştır. Buna göre, seçilen Madelon veri setinde doğru sınıflama oranını arttırmasına rağmen SNFS, ilgili veri seti için DVM'de çok yüksek bir doğru sınıflama oranı elde edilmesini sağlayamamıştır. Veri yapısının mikrodizi verilerine çok uygun olmaması nedeniyle bu sonuç yöntemin yeterliliğine ilişkin geçerli bir bilgi sağlamamakla birlikte gürültülü ve az sayıda ilgili özellik içeren veri yapılarında SNFS'nin ilgili özellikleri seçmede başarılı olduğunu göstermiştir. Ayrıca farklı yöntem kombinasyonlarının sınıflama başarımı üzerinde ciddi etkileri olduğu

görülmüştür. Bunun ötesinde, alanyazında Madelon veri seti kullanılarak elde edilen sınıflama başarımından daha yüksek bir doğru sınıflama oranı da elde edilmiştir (19). Canedo çalışmasında (19) kullandığı farklı özellik seçim yöntemleri sonrası DVM ile sınıflama yapmış ve DVM için en yüksek doğru sınıflama oranı %67,54 olarak elde edilmiştir. Oysa bu tez çalışmasında, SNFS ile seçilen özellikler kullanılarak DVM'den elde edilen doğru sınıflama oranı %77,5'e kadar yükseltilebilmiştir.

Aynı zamanda bir benzetim çalışması yapılarak tez kapsamında kullanılması planlanan yöntemlerin hangisinin belirli özelliklere sahip mikrodiziler için en uygun çözümü sağladığı araştırılmış ve sonuç olarak hemen hemen tüm özellik seçim ve topluluk belirleme yöntem kombinasyonları ile oluşturulan SNFS sonucunda elde edilen biyobelirteç genler (yapay) ile DVM'den benzer sınıflama başarımı elde edildiği görülmüştür. Ancak, Walktrap ve Infomap topluluk belirleme yöntemlerinin daha kararlı sonuçlar ürettiği düşünülmektedir. Aynı zamanda alanyazında uygulanan gen seçimi sonrası daha yüksek sayıda gen (yapay) ile elde edilen sınıflama başarımına, SNFS ile seçilen çok daha az sayıdaki gen (yapay) ile yaklaşılabildiği düşünülmektedir (42). Guo ve diğ. (42) çalışmalarında tez çalışması ile benzer yapıda ürettikleri 200 örneklik eğitim setinde önerdikleri yöntemlerle 167 ile 282 gen seçerek sırasıyla %89,5 ve %87,5 doğru sınıflama oranına 10-kat çapraz geçerlik sonucunda ulaşmışlarken tez çalışmasında SNFS ile seçilen 6 ile 8 gen arasındaki biyobelirteç genler kullanılarak eğitim setinde DVM ile %81,5-%83 arasında doğru sınıflama oranı elde edilmiştir.

Eğitim setinde elde edilen bu başarı kayda değer olmakla birlikte aynı sınıflama başarımına test setinde ulaşılamamıştır. Alanyazında 1000 örneklik test setinde 167 ile 282 gen seçerek sırasıyla %90,4 ve %89,2 doğru sınıflama oranına ulaşılmıştır (42). Ancak, tez çalışması kapsamında benzer yapıda üretilen 600 örneklik test setinde SNFS ile seçilen 6 ile 8 arasında gen kullanılarak elde edilen DVM sınıflama başarımına ilişkin ölçülere bakıldığında (Bkz. Tablo 4.26.) doğru sınıflama oranı ancak %68'e kadar yükseltilebilmiştir. Ancak, SNFS sonucunda test setinde elde edilen DVM sınıflama başarımları orta düzeyde olmasına rağmen seçilen özellik sayısı alanyazına göre ciddi anlamda daha düşüktür. Bu nedenle, topluluğu temsil edebilecek özellik oranının benzetim çalışmasında seçilmiş olduğu haliyle yani %5'ten daha

yüksek seçilmesi durumunda hem alanyazına (42) göre yine daha az özellik seçilebilir hem de sınıflama başarımında bir iyileşme sağlanabilir.

Bunun yanı sıra genler arası korelasyon yapısı değiştirilerek sınıflama başarımına korelasyonun etkisi gözlemlenmiş ve aynı bloktaki genlere ilişkin korelasyonun artmasıyla SNFS yöntemiyle seçilen genler kullanılarak DVM'den elde edilen sınıflama başarımında bir düşüş gözlenmiştir. Buna göre, genler arası korelasyon yapısı doğrudan sınıflama başarımı üzerinde etkilidir ve boyut indirgemenin sağladığı olumlu etkiyi aynı bloktaki genler arası korelasyondaki artış düşürebilmektedir.

Tez kapsamında hazırlanan fonksiyon yardımıyla SNFS parametrelerinin (özellik seçim yöntem kombinasyonları, seçilecek gen yüzdesi, kümeleme adımının atlanıp atlanmayacağı, topluluk belirleme yöntemleri, topluluğu temsil edebilecek gen yüzdesi, ağa özel metrik, vb.) bir bölümündeki değişimin etkisi incelenmiştir. Ancak, bu parametrelerin tümündeki değişimin DVM ile elde edilecek sınıflama başarımı üzerindeki etkisi de karşılaştırılabilir.

6. SONUÇ ve ÖNERİLER

Melez özellik seçim yöntemlerinin aşamalarında, birden çok makine öğrenmesi, veri madenciliği ve/veya sosyal ağ analizi yöntemlerinin çok sayıda farklı kombinasyonunun söz konusu olması nedeniyle problem çözümü için bu yöntemleri uygun biçimde birleştirmek ve probleme özgü en uygun çözümü bulmak için olası tüm kombinasyonların değerlendirilmesi gerekmektedir.

SNFS yöntemi sosyal ağ analizinin de kullanıldığı melez bir özellik seçim yöntemidir ve yöntemin birçok adımında farklı algoritma ve yöntemlerin söz konusu olma olasılığı nedeniyle aynı veri seti için farklı sınıflama başarımı sağlayabilecek potansiyele sahiptir. Bu nedenle, kullanıcıların aynı platformda SNFS yöntemini uygulayabilmesi ve olası yöntem kombinasyonlarını deneyebilmeleri amacıyla fonksiyonlar geliştirilmiştir. Bu fonksiyonlar geliştirilmeye ve genişletilmeye açıktır.

Tez çalışması kapsamında, SNFS yönteminin ana ve alt adımlarda yer alan çeşitli yöntemler orijinal ve sentetik veri setleri üzerinde denenmiş ve sonuç olarak SNFS'nin içerdiği farklı yöntem kombinasyonlarının sınıflama başarımı üzerinde temel olarak olumlu etkisi olduğu görülmüştür. Ancak kombinasyonlar değiştikçe sınıflama başarımının etkilendiği ve bu nedenle veri yapısına en uygun kombinasyonun seçilebilmesi için araştırmacıların olası tüm kombinasyonları denemesi gerektiği düşünülmektedir.

Tez çalışmasında ele alınan iki sınıflı sınıflama problemi kapsamında değerlendirilen SNFS yöntemi, aynı zamanda çok sınıflı problemlere uygun özellik seçim yöntemlerinin kullanımı yoluyla probleme uyarlanabilir.

7. KAYNAKLAR

1. Dziuda DM. Data mining for genomics and proteomics: analysis of gene and protein expression data. Hoboken, New Jersey: Wiley; 2010.

2. Basaran E, Aras S, Cansaran-Duman D. General Outlook and Applications of Genomics, Proteomics and Metabolomics. Turk Hij Den Biyol Derg. 2010;67(2):85-96.

3. Apitz JC. A Statistical Method for Selection, Classification, and Network Construction in Genetic Systems [Master of Science]: California State University; 2016.

4. Karabulut E, Karaagaoglu, E. Biyoinformatik ve biyoistatistik. Hacettepe Tıp Dergisi. 2010(41):162-70.

5. Cosgun E, Karaağaoğlu E. Veri madenciliği yöntemleriyle mikrodizilim gen ifade analizi. Hacettepe Tıp Dergisi. 2011;42:180-9.

6. Zararsız G. Gen Ekspresyon Verilerinde Kümelemeye Dayalı Yeni Bir Sınıflandırma Yaklaşımı [Yüksek Lisans Tezi]. Kayseri: Erciyes Üniversitesi; 2012.

7. Zengin HY, Karabulut, E, Öğüş, E, Başkent Üniversitesi Tıp Fakültesi'nde 2014-2015 Yılları Arasında Üretilen Bilimsel Yayınların Sosyal Ağ Analizi Yaklaşımı ile İncelenmesi. XVII Ulusal Biyoistatistik Kongresi; 5-9 Kasım 2015; KKTC.

8. McCurdie T, Sanderson P, Aitken LM. Applying social network analysis to the examination of interruptions in healthcare. Applied ergonomics. 2018;67:50-60.

9. Seker SE. Sosyal ağlarda veri madenciliği (data mining on social networks). Ybs Ansiklopedi. 2015;2(2):30-39.

10. Brantingham PL, Ester M, Frank R, Glässer U, Tayebi MA. Co-offending network mining. Counterterrorism and Open Source Intelligence: Springer; 2011. p. 73-102.

11. Naji G, Nagi M, ElSheikh AM, Gao S, Kianmehr K, Özyer T, et al. Effectiveness of social networks for studying biological agents and identifying cancer biomarkers. Counterterrorism and open source intelligence: Springer; 2011. p. 285-313.

12. Dey P, Roy S. Social network analysis. Advanced Methods for Complex Network Analysis: IGI Global; 2016. p. 237-65.

13. Horvath S. Weighted network analysis: applications in genomics and systems biology: Springer Science & Business Media; 2011.

14. Ozyer T, Ucer S, Iyidogan T. Employing social network analysis for disease biomarker detection. International journal of data mining and bioinformatics. 2015;12(3):343-62.

15. Üçer S, Koçak Y, Ozyer T, Alhajj R. Social network Analysis-based classifier (SNAc): A case study on time course gene expression data. Computer methods and programs in biomedicine. 2017;150:73-84.

16. Mason MJ, Fan G, Plath K, Zhou Q, Horvath S. Signed weighted gene co-expression network analysis of transcriptional regulation in murine embryonic stem cells. BMC genomics. 2009;10(1):327.

17. Gates KM, Henry T, Steinley D, Fair DA. A Monte Carlo evaluation of weighted community detection algorithms. Frontiers in neuroinformatics. 2016;10:45.

18. Gysi DM, Fragoso TM, Almaas E, Nowick K. CoDiNA: an R Package for Co-expression Differential Network Analysis in n Dimensions. arXiv preprint arXiv:180200828. 2018.

19. Canedo VB, Marono NS. Novel feature selection methods for high dimensional data [Ph.D. Thesis]. Spain: Universidade da Coruña; 2014.

20. R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

21. Yılmaz E. Nükleikasitlerin Yapisi, Fonksiyonu ve Genom Organizasyonu. [Erişim tarihi: Haziran 2018] www.thd.org.tr/thdData/userfiles/file/molhem_01.pdf

22. Gürel O. Yaşamın Kökeni: Pan Yayıncılık; 1999.

23. Öztürk F. Temel Kavramlar I [Internet]. 2007 [Erişim tarihi: Haziran 2016]. http://80.251.40.59/science.ankara.edu.tr/ozturk/ist432.html.

24. Parmigiani G, Garret E, Izizarry R, Zeger S. Statistics for biology and health. The Analysis of Gene Expression Data: Methods and Software. 2003.

25. Komurcu-Bayrak E, Erginel-Ünaltuna N. Gen Anlatımı Analiz Yöntemlerine Genel Bakış. Deneysel Tıp Araştırma Enstitüsü Dergisi. 2011;1(2):28-35.

26. Zhang A. Advanced analysis of gene expression microarray data: World Scientific Publishing Company; 2006.

27. Bumgarner R. Overview of DNA microarrays: types, applications, and their future. Current protocols in molecular biology. 2013:101(1):22-1.

28. Tan AC, Gilbert D. An empirical comparison of supervised machine learning techniques in bioinformatics. Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics 2003-Volume 19; 2003: Australian Computer Society, Inc.

29. Alpaydin E. Introduction to machine learning, Second Edition: MIT press; 2010.

30. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. science. 1999;286(5439):531-7.

31. Liao J, Chin KV. Logistic regression for disease classification using microarray data: model selection in a large p and small n case. Bioinformatics. 2007;23(15):1945-51.

32. Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American statistical association. 2002;97(457):77-87.

33. Peng HY, Jiang CF, Fang X, Liu JS. Variable selection for Fisher linear discriminant analysis using the modified sequential backward selection algorithm for the microarray data. Applied Mathematics and Computation. 2014;238:132-40.

34. Soukup M, Lee JK. Developing optimal prediction models for cancer classification using gene expression data. Journal of Bioinformatics and Computational Biology. 2004;1(04):681-94.

35. Jelizarow M, Guillemot V, Tenenhaus A, Strimmer K, Boulesteix AL. Over-optimism in bioinformatics: an illustration. Bioinformatics. 2010;26(16):1990-8.

36. Nguyen DV, Rocke DM. Tumor classification by partial least squares using microarray gene expression data. Bioinformatics. 2002;18(1):39-50.

37. Mukherjee S, Tamayo P, Slonim D, Verri A, Golub T, Mesirov J, et al. Support vector machine classification of microarray data. 1999.

38. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics. 2000;16(10):906-14.

39. Geman D, d'Avignon C, Naiman DQ, Winslow RL. Classifying gene expression profiles from pairwise mRNA comparisons. Statistical applications in genetics and molecular biology. 2004;3(1):1-19.

40. Hwang KB, Cho DY, Park SW, Kim SD, Zhang BT. Applying machine learning techniques to analysis of gene expression data: cancer diagnosis. Methods of microarray data analysis: Springer; 2002. p. 167-82.

41. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Machine learning. 2002;46(1-3):389-422.

42. Guo Y, Hastie T, Tibshirani R. Regularized discriminant analysis and its application in microarrays. Biostatistics. 2005;1(1):1-18.

43. Díaz-Uriarte R, De Andres SA. Gene selection and classification of microarray data using random forest. BMC bioinformatics. 2006;7(1):3.

44. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. bioinformatics. 2007;23(19):2507-17.

45. Bolón-Canedo V, Sánchez-Maroño N, Alonso-Betanzos A. A review of feature selection methods on synthetic data. Knowledge and information systems. 2013;34(3):483-519.

46. Amaldi E, Kann V. On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. Theoretical Computer Science. 1998;209(1-2):237-60.

47. Bellman R. Curse of dimensionality. Adaptive control processes: a guided tour Princeton, NJ. 1961

48. Jain A, Zongker D. Feature selection: Evaluation, application, and small sample performance. IEEE transactions on pattern analysis and machine intelligence. 1997;19(2):153-8.

49. Gümüşçü A, Aydilek İB, Taşaltın R. Mikro-dizilim Veri Sınıflandırmasında Öznitelik Seçme Algoritmalarının Karşılaştırılması. Harran Üniversitesi Mühendislik Dergisi. 2016;1(1):1-7.

50. Rokach L, Chizi B, Maimon O. A methodology for improving the performance of non-ranker feature selection filters. International Journal of Pattern Recognition and Artificial Intelligence. 2007;21(05):809-30.

51. Liu H, Setiono R, editors. Chi2: Feature selection and discretization of numeric attributes, 1995: IEEE.

52. Quinlan JR. Induction of decision trees. Machine learning. 1986;1(1):81-106.

53. Ding Y, Wilkins D. Improving the performance of SVM-RFE to select genes in microarray data. BMC bioinformatics; 2006: BioMed Central.

54. Furlanello C, Serafini M, Merler S, Jurman G. Entropy-based gene ranking without selection bias for the predictive classification of microarray data. BMC bioinformatics. 2003;4(1):54.

55. Kohavi R, John GH. Wrappers for feature subset selection. Artificial intelligence. 1997;97(1-2):273-324.

56. Zhao ZA, Liu H. Spectral feature selection for data mining: CRC Press; 2011.

57. Zhang Y, Ding C, Li T. Gene selection algorithm by combining reliefF and mRMR. BMC genomics. 2008;9(2):S27.

58. Peng Y, Wu Z, Jiang J. A novel feature selection approach for biomedical data classification. Journal of Biomedical Informatics. 2010;43(1):15-23.

59. El Akadi A, Amine A, El Ouardighi A, Aboutajdine D. A two-stage gene selection scheme utilizing MRMR filter and GA wrapper. Knowledge and Information Systems. 2011;26(3):487-500.

60. Bennet J, Ganaprakasam C, Kumar N. A Hybrid Approach for Gene Selection and Classification using Support Vector Machine. International Arab Journal of Information Technology (IAJIT). 2015;12.

61. Vainer I, Kraus S, Kaminka GA, Slovin H. Obtaining scalable and accurate classification in large-scale spatio-temporal domains. Knowledge and information systems. 2011;29(3):527-64.

62. Tuv E, Borisov A, Runger G, Torkkola K. Feature selection with ensembles, artificial variables, and redundancy elimination. Journal of Machine Learning Research. 2009;10(Jul):1341-66.

63. Xu M, Setiono R. Gene selection for cancer classification using a hybrid of univariate and multivariate feature selection methods. arXiv preprint arXiv:150602085. 2015.

64. Sun Y, Li J. Iterative RELIEF for feature weighting. Proceedings of the 23rd international conference on Machine learning; 2006: ACM.

65. Sun Y, Todorovic S, Goodison S. A feature selection algorithm capable of handling extremely large data dimensionality. Proceedings of the 2008 SIAM International Conference on Data Mining; 2008: SIAM.

66. Chidlovskii B, Lecerf L. Scalable feature selection for multi-class problems. Joint European Conference on Machine Learning and Knowledge Discovery in Databases; 2008: Springer.

67. Saeys Y, Abeel T, Van de Peer Y. Robust feature selection using ensemble feature selection techniques. Joint European Conference on Machine Learning and Knowledge Discovery in Databases; 2008: Springer.

68. Bolón-Canedo V, Sánchez-Maroño N, Alonso-Betanzos A. An ensemble of filters and classifiers for microarray data classification. Pattern Recognition. 2012;45(1):531-9.

69. Ang JC, Haron H, Hamed HNA. Semi-supervised SVM-based feature selection for cancer classification using microarray gene expression data. International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems; 2015: Springer.

70. Han J, Pei J, Kamber M. Data mining: concepts and techniques. Third Edition ed: Elsevier; 2011.

71. MacQueen J. Some methods for classification and analysis of multivariate observations. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability; 1967: Oakland, CA, USA.

72. Hartigan JA, Wong MA. Algorithm AS 136: A k-means clustering algorithm. Journal of the Royal Statistical Society Series C (Applied Statistics). 1979;28(1):100-8.

73. Lloyd S. Least squares quantization in PCM. IEEE transactions on information theory. 1982;28(2):129-37.

74. Sharan R. Analysis of biological networks: Network modules–clustering and biclustering. lecture notes. 2006.

75. Tan PN. Introduction to data mining: Pearson Education India; 2006.

76. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Statistical applications in genetics and molecular biology. 2005;4(1).

77. Borate BR, Chesler EJ, Langston MA, Saxton AM, Voy BH. Comparison of threshold selection methods for microarray gene co-expression matrices. BMC research notes. 2009;2(1):240.

78. Fortunato S. Community detection in graphs. Physics reports. 2010;486(3-5):75-174.

79. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment. 2008;2008(10):P10008.

80. Rosvall M, Bergstrom CT. Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences. 2008;105(4):1118-23.

81. Lancichinetti A, Fortunato S, Radicchi F. Benchmark graphs for testing community detection algorithms. Physical review E. 2008;78(4):046110.

82. Pons P, Latapy M. Computing communities in large networks using random walks. J Graph Algorithms Appl. 2006;10(2):191-218.

83. Meghanathan N. Advanced methods for complex network analysis: IGI Global; 2016.

84. Freeman LC. Centrality in social networks conceptual clarification. Social networks. 1978;1(3):215-39.

85. Opsahl T, Agneessens F, Skvoretz J. Node centrality in weighted networks: Generalizing degree and shortest paths. Social networks. 2010;32(3):245-51.

86. Golub T. golubEsets: exprSets for golub leukemia data. R package version 1.20.0. 2017.

87. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Sciences. 1999;96(12):6745-50.

88. Merk S. colonCA: exprSet for Alon et al. (1999) colon cancer data. R package version 1.20.0. 2017.

89. Guyon I. Design of experiments for the NIPS 2003 variable selection benchmark 2003 [Available from: http://clopinet.com/isabelle/Projects/NIPS2003/.

90. Ramey JA. sortinghat: sortinghat. R package version 0.1. 2013. https://CRAN.R-project.org/package=sortinghat

91. David Meyer ED, Kurt Hornik, Andreas Weingessel, Friedrich Leisch e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.6-8. 2017. https://CRAN.R-project.org/package=e1071

92. Romanski P, Kotthoff L. FSelector: Selecting Attributes. R package version 0.21. 2016. https://CRAN.R-project.org/package=FSelector.

93. Guyon I. R implementation of the Support Vector Machine Recursive Feature Extraction (SVM-RFE) Algorithm [Internet]. (Erişim Tarihi: Haziran 2016). http://www.uccor.edu.ar/paginas/seminarios/Software/SVM_RFE_R_implementation.pdf

94. Csardi G, Nepusz T. The igraph software package for complex network research, InterJournal, Complex Systems 1695. 2006. http://igraph.org

95. Klockiewicz B, Alvarez A. R implementation of the Force Atlas 2 graph layout designed for Gephi. (Erişim Tarihi: Haziran 2018). https://github.com/adolfoalvarez/Force-Atlas-2

ÖZGEÇMİŞ

Kişisel bilgiler

Ad / Soyad	**Hatice Yağmur Zengin**
Doğum tarihi	18/04/1985
E-posta adresi	yagmurz@baskent.edu.tr

Eğitim Bilgileri

Lisans	Hacettepe Üniversitesi Fen Fakültesi İstatistik Bölümü Mezuniyet Tarihi: 31/01/2011
Yüksek Lisans	Hacettepe Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı İstatistik-Tezli Yüksek Lisans Programı (Mezun) Mezuniyet Tarihi: 28/06/2013

Kişisel beceri ve yeterlilikler

Yabancı Dil	**İngilizce – 2011 KPDS İlkbahar Dönemi (83,75)** **İngilizce – 2017 Kasım YÖKDİL (95,0)** **European Language Portfolio C2 Level Certificate**

İş/Staj Deneyimi

Başkent Üniversitesi Tıp Fakültesi Biyoistatistik Anabilim Dalı, Araştırma Görevlisi (Eylül 2015 – Halen)

Bildiri ve Makaleler

Ödüller:

- Zengin H, Öğüş E, Karabulut E.; (0,1) Aralığında Tanımlı Sürekli Verilerin Modellenmesinde Kullanılan Bir Yöntem: Beta Regresyon Analizi ve Bir Uygulama. XVII. Ulusal Biyoistatistik Kongresi: KKTC; 07/11/2015 (Sözlü Bildiri Üçüncülük Ödülü)
- Zengin H. Y., Öğüş E., Karabulut E.; En İyi Sözlü Bildiri. Afyon Kocatepe Üniversitesi ve Biyoistatistik Derneği, "XVIII. Ulusal Biyoistatistik Kongresi ve 1. Uluslararası Biyoistatistik Kongresi", 29/10/2016 (Sözlü Bildiri Birincilik Ödülü)

Yayınlar:

- Oğuz D., Tuncay E., Zengin H. Y.; PP-036 Diagnostic Value of Platelet Indexes for Massive Pulmonary Embolism. The American Journal of Cardiology, 2016; 117(Suppl1):55-. (elsevier)
- Kızıltan E, Aydın L, Zengin HY. ;Internal Motivation Modulates Voluntary Repetitive Movements: "Ha gayret" Energy. Acta Physiologica, 2017; 221(Suppl714):81-. (SSCI : Social Sciences Citation Index)

Bildiri ve Makaleler

Sözlü Bildiriler:

- Zengin H. Y., Ergün G.; Markov Zinciri Monte Carlo Yönteminin Dinamik Doğrusal Modellere Uygulanması. 8. Uluslararası İstatistik Kongresi: Antalya; 27/09/2013-30/09/2013
- Zengin H. Y., Karabulut E.; Sayımla Elde Edilen Verilerin Modellenmesinde Kullanılan Bazı Regresyon Yöntemlerinin Tahmin Performansları Üzerine Bir Benzetim Uygulaması. XVI. Ulusal Biyoistatistik Kongresi: Antalya; 10/10/2014 - 12/10/2014
- Zengin H. Y., Öğüş E., Karabulut E.; Başkent Üniversitesi Tıp Fakültesi'nde 2014-2015 Yılları Arasında Üretilen Bilimsel Yayınların Sosyal Ağ Analizi Yaklaşımı ile İncelenmesi. XVII. Ulusal Biyoistatistik Kongresi: KKTC; 05/11/2015 - 09/11/2015
- Zengin H. Y., Öğüş E., Karabulut E.; (0,1) Aralığında Tanımlı Sürekli Verilerin Modellenmesinde Kullanılan Bir Yöntem: Beta Regresyon Analizi ve Bir Uygulama. XVII. Ulusal Biyoistatistik Kongresi: KKTC; 05/11/2015 - 09/11/2015
- Oğuz D., Tuncay E., Zengin H.Y.;.Diagnostic Value of Platelet Indexes for Massive Pulmonary Embolism. 12. Uluslararası Kardiyoloji ve Kardiyovasküler Cerrahide Yenilikler Kongresi: Antalya; 10/03/2016 - 13/03/2016
- Zengin H. Y., Öğüş E., Karabulut E.; [0,1] Aralığında Tanımlı Sürekli Sayısal Verilerin Beta Regresyon Analizi ile Modellenmesinde Kullanılan Farklı Dönüşüm Yöntemleri ve Modelleme Stratejilerine İlişkin Bir Uygulama. XVIII. Ulusal Biyoistatistik Kongresi ve I. Uluslararası Biyoistatistik Kongresi: Belek, Antalya; 26/10/2016 - 29/10/2016
- Zengin H. Y., Öğüş E., Karabulut E.; Aykırı Değer veya Heteroskedastisite Varlığında Kantil Regresyon Yaklaşımının Performansı ve Doğrusal Regresyon Yöntemi ile Karşılaştırılması. XVIII. Ulusal Biyoistatistik Kongresi ve I. Uluslararası Biyoistatistik Kongresi: Belek, Antalya; 26/10/2016 - 29/10/2016
- Şençelikel T., Zengin H. Y., Öğüş E.; Veri Setinde Eksik Gözlem Olması Durumunda Goodman-Kruskal Gamma, Gwet AC2 ve Krippendorff Alfa Uyum Ölçütlerinin Karşılaştırılması: Bir Simülasyon Çalışması. XVIII. Ulusal Biyoistatistik Kongresi ve I. Uluslararası Biyoistatistik Kongresi: Belek, Antalya; 26/10/2016 - 29/10/2016
- Şençelikel T, Zengin HY, Öğüş E, Öner KS. ;Kategori Sayısı İkiden Fazla Olan Bağımsız Değişkenlerde Johnson-Neyman Prosedürü. XIX. Ulusal ve II. Uluslararası Biyoistatistik Kongresi: Antalya; 25/10/2017 - 28/10/2017
- Zengin HY, Öğüş E, Karabulut E. ;[0,1] Kapalı Aralığında Tanımlı Sürekli Bir Bağımlı Değişkenin Modellenmesinde Kullanılan Farklı Modelleme Stratejilerinin İncelenmesi. XIX. Ulusal ve II. Uluslararası Biyoistatistik Kongresi: Antalya; 25/10/2017 - 28/10/2017
- Kızıltan E, Aydın L, Zengin HY. ;Tekrarlayan İstemli Hareketlerin İçsel Motivasyonla Modülasyonu: "Ha Gayret" Enerjisi. Türk Fizyolojik Bilimler Derneği, 43. Ulusal Fizyoloji Kongresi: Denizli; 07/09/2017 - 10/09/2017

Posterler:

- Öğüş E., Zengin H. Y.; Biyoistatistik ve Adli Bilimler Üzerine Bir Uygulama. XVII. Ulusal Biyoistatistik Kongresi: KKTC; 05/11/2015 - 09/11/2015
- Öğüş E., Zengin H. Y., Şençelikel T., Akpınar D., Sonsayar D. İ., Yıldız S. R., Aplak S. B.; Klinik Denemelerde Yanlılık Kavramı ve Önleme Yöntemleri. XVIII. Ulusal Biyoistatistik Kongresi ve I. Uluslararası Biyoistatistik Kongresi: Belek, Antalya; 26/10/2016 - 29/10/2016