BAŞKENT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

KANSER SINIFLANDIRMADA mikroRNA VE mRNA ANLATIM BİLGİLERİNİN ENTEGRASYONU

ONUR ALTINDAĞ

YÜKSEK LİSANS TEZİ
2013

KANSER SINIFLANDIRMADA mikroRNA VE mRNA ANLATIM BİLGİLERİNİN ENTEGRASYONU

INTEGRATING microRNA AND mRNA EXPRESSION DATA FOR CANCER CLASSIFICATION

ONUR ALTINDAĞ

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

2013

“Kanser Sınıflandırmada mikroRNA ve mRNA Anlatım Bilgilerinin Entegrasyonu” başlıklı bu çalışma, jürimiz tarafından, _/08/2013 tarihinde, **BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI'nda YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Başkan :

Doç. Dr. Mustafa KOCAKULAK

Üye (Danışman) :

Doç. Dr. Hasan OĞUL

Üye :

Yrd. Doç. Dr. Emre SÜMER

ONAY

..../..../.....

Prof.Dr. Emin AKATA

Fen Bilimleri Enstitüsü Müdürü

TEŞEKKÜR

Bu çalışma TÜBİTAK tarafından 110E160 nolu proje ile desteklenmiştir. TÜBİTAK ve ilgili birimlerine desteklerinden dolayı teşekkür ederiz.

Bu tez çalışmasını tamamlamamda büyük emeği geçen sevgili eşim Tuğba ALTINDAĞ'a, yoğun zamanlarımda desteklerini esirgemeyen anne ve babam Hilal-Hikmet ALTINDAĞ'a, binbir türlü şirinlikleri ile tüm sıkıntılarımı unutturan ve ileride benden çok daha önemli çalışmalar yapmasını umduğum canım kızım, minik bebeğim Arya'ya, yardımını hiç esirgemeyen dostum Didem ÖZIŞIK BAŞKURT ve ihtiyaç duyduğum her zaman yanımda olan diğer tüm arkadaşlarıma çok teşekkür ederim.

Tez danışmanım olmasının ve bu çalışmayı gerçekleştirebilmem için gerekli tüm desteği vermesinin çok ötesinde mesleğimde elde edebildiğim tüm başarılarda büyük payı olduğunu düşündüğüm, beni mesleğimle tanıştıran ve bu mesleğe hayran olmamı sağlayan değerli hocam Doç. Dr. Hasan OĞUL'a sonsuz teşekkürlerimi sunarım.

ÖZ

KANSER SINIFLANDIRMADA mikroRNA VE mRNA ANLATIM BİLGİLERİNİN ENTEGRASYONU

Onur ALTINDAĞ

Başkent Üniversitesi Fen Bilimleri Enstitüsü

Bilgisayar Mühendisliği Anabilim Dalı

Gen ifade verilerinden kanserli doku örneklerinin sınıflandırılması günümüz biyokimyasının en önemli problemlerindendir. Bu problemi zor kılan en önemli durum, tipik bir mikroseri deneyindeki çok yüksek miktarda gen sayısına (mRNA) karşılık çok az sayıda örnek bulunmasıdır. Yapılan son araştırmalarda öznitelik seçimi yöntemlerinin bu sorunu aşmada önemli rolü olduğu raporlanmaktadır. Bunun yanı sıra kanserli doku saptamada mikroRNA ifade biçimlerinin de önemli bir bilgi değeri taşıdığı belirtilmektedir. Bu çalışmada bu iki bulgunun kapsamlı bir şekilde ele alınmasıyla mikroRNA-mRNA entegrasyonu üzerinde öznitelik seçimi yöntemlerinin etkisi değerlendirilmiştir. Çalışmamızın sonucunda bu entegrasyonun etkili bir öznitelik seçim stratejisinin de yardımıyla uygulanan sınıflandırıcıların tahmin doğruluğunu önemli ölçüde arttırdığı ispatlanmıştır.

ANAHTAR SÖZCÜKLER: Kanser sınıflandırma, tümör sınıflandırma, çok kategorili kanser sınıflandırma, kanser alt kategori sınıflandırma, göğüs kanseri, gen ifadesi, mikroRNA, mRNA, veri entegrasyonu, öznitelik seçimi, makine öğrenme.

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

ABSTRACT

INTEGRATING microRNA AND mRNA EXPRESSION DATA FOR CANCER CLASSIFICATION

Onur ALTINDAĞ

Başkent University Institute of Science Engineering

Department of Computer Engineering

Classifying cancer samples from gene expression data is one of the central problems in current systems biomedicine. The problem is challenging due to the small number of samples in comparison to the number of genes (mRNAs) in a typical microarray experiment. Recent reports suggest that feature selection may help to manage the problem. Furthermore, microRNA expression profiles have shown to provide valuable knowledge in detecting cancer signatures. In this study, we present the results of a comprehensive study to assess the effect of feature selection and microRNA-mRNA data integration in cancer type prediction from microarray expression data. We prove that this integration can significantly improve prediction accuracy with a proper feature selection strategy.

KEYWORDS: Cancer classification, tumor classification, multi-class cancer classification, cancer sub-classification, breast cancer, gene expression, microRNA, mRNA, data integration, feature selection, machine learning.

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

İÇİNDEKİLER LİSTESİ

Sayfa

Sayfa

ŞEKİLLER LİSTESİ

Sayfa

ÇİZELGELER LİSTESİ

Sayfa

SİMGELER VE KISALTMALAR LİSTESİ

n	bir sınıflandırma probleminde analiz edilen örnek sayısı
p	bir sınıflandırma probleminde analiz edilen öznitelik sayısı
x_i	veri değeri
t_i	kayıt
C_j	sınıf
p	bağımsız öznitelik
k	KNN sınıflandırıcıda değerlendirilecek en yakın komşu sayısı
D	bir veri setinin herhangi bir zamandaki durumu
d_i	beklenen çıktı
y_i	gerçek çıktı
m	çıktı düğüm sayısı
$\boldsymbol{\beta}$	ağırlık vektörü
$\boldsymbol{\beta_0}$	öngörü değeri
M	marjin
O_{ij}	gözlenen sıklık
E_{ij}	teorik sıklık

DNA	Deoksiribonükleik Asit
PCR	Polymerase Chain Reaction
mRNA	Mesajcı Ribonükleik Asit
DT	Decision Tree
ANN	Artificial Neural Network
SVM	Support Vector Machine
NBM	Naïve Bayes Multinomial
KNN	K-Nearest Neighbors
MSE	Mean Squared Error
RBF	Radial Basis Function
MLP	Multilayer Perceptron
WEKA	Waikato Environment for Knowledge Analysis
CFS	Correlation Based Feature Subset
GCM	Global Cancer Map
GEO	Gene Expression Omnibus
LOOCV	Leave-one-out-cross validation

1 GİRİŞ

Kanser birçok farklı belirtisi ve moleküler düzeyde etkisi olan kompleks bir bozukluktur. Ölümcül bir hastalık olması ve giderek artan görülme ve tekrarlama sıklığı sebebiyle bu hastalık üzerine yoğun bilimsel çalışmalar gerçekleştirilmektedir. Farklı alanlardaki teknolojik gelişmeler ışığında kanser teşhisi ve bu hastalığın anlaşılması ile ilgili bir çok ilerleme kaydedilmiştir. Moleküler biyoloji, hücresel biyoloji ve patolojik teknikler konusundaki ilerlemeler bu gelişimin temelini oluşturmaktadır. Bu gelişmelerle birlikte genetik analiz kanser teşhisinde kullanılan önemli metotlardan biri olmuştur.

Bir doku örneğinin kanserli olup olmadığının anlaşılması ve eğer kanserliyse bu kanserin hangi kategoride olduğunun belirlenmesi işlemine kanser teşhisi veya kanser sınıflandırma adı verilmektedir. Genetik analizde kan, kemik veya çeşitli organlardan alınan doku örnekleri farklı yöntemler ile incelenmektedir. DNA (Deoksiribonükleaik asit), dizileme, sitogenetik, PCR (Polymerase chain reaction) ve mikroseriler (gen çipleri) bu yöntemlerin önde gelenleridir. Özellikle kanser sınıflandırma işleminde mikroseriler sık kullanılan bir yöntemdir. Bu yöntemde mikroskobik boyutlardaki bir örneğin içindeki çok sayıda genin ifade seviyeleri aynı anda ölçülebilmektedir. Bir genin ilgili örnekte ifade edilmesi o gen ile ilgili mRNA (mesajcı ribonükleik asit) üretildiği anlamına gelir ve yüksek ifade değerine sahip genler daha fazla sayıda mRNA üretir [1; 2; 3].

Genetik testlerin zorluğu ve maliyeti sebebiyle, bir mikroseri deneyinde çok sayıda örnek bulunması henüz mümkün değildir. Fakat incelenen her örnek için mümkün olan en yüksek sayıda genetik bilgi çıkarılmaya çalışılır. Genetik analizde de en ciddi problemlerden biri olan **'küçük *n* büyük *p* paradigması'** olarak da anılan, analiz edilecek veri örneğinin çok az, veri özniteliklerinin ise çok fazla olma durumu bilimsel veri analizindeki en uğraştırıcı olaylardan biridir. Bu durumla başa çıkmak için literatürde üç temel yöntem bulunmaktadır. Bunlardan birincisi öznitelik azaltmadır. Bu yöntem temelde, veri özniteliklerinden yapılacak çıkarım için diğerlerine göre etkisinin az veya hiç olmadığı bir bölümünün uygun bir yöntemle diğer özniteliklerden ayrılmasıdır. Sonuç olarak toplam öznitelik sayısı *p* tüm veri özniteliklerinin etkin ama küçük bir alt kümesinin bulunması ile azaltılmaktadır

(Saeys et al., [4]). İkinci yöntem olan veri entegrasyonu veya füzyonu ise uygun bir şekilde birden çok veri kaynağından faydalanma esasıyla sonuç olarak yapılacak tahminin iyileştirilmesidir. Bu yöntem farklı şekillerde gerçekleştirilebilir. İlave ölçümlerin (örnekler) kullanılması veya orijinal öznitelikler üzerinde tamamlayıcı etkisi olabilecek başka özniteliklerin ele alınması bunlar arasında sayılabilir (Oğul and Akkaya, [5]). Üçüncü yöntem ise çıkarım için Bayes varsayımına dayanan istatistiksel modellemedir. Bu kategorideki metotlar çok boyutlu uzayın belirsizliğini modellemek için olasılık dağılımlarını kullanırlar (West, [6]; Klami and Kaski, [7]). Gelişmiş başarılı yöntemler ise çoğunlukla bu üç yöntemin farklı şekillerde birleşiminden oluşmaktadır.

Günümüz biyokimya sistemlerindeki en önemli problemlerden biri verilen bir doku örneğindeki tümör tipinin teşhis edilebilmesidir. Bir doku örneğinin beraberinde çoğunlukla mikroseri deneyi ile elde edilmiş mRNA ifade profilleri de bulunur. Her kanser tipinin ayırt edici bir şekilde ilişkili bazı genlerin düzenleyici özelliklerini değiştiriyor olmasından ötürü bu profillerin tümör sınıflandırma potansiyelleri çok yüksektir. Bu profillerde her doku örneği genomda bilinen tüm genlerin aktivitelerine karşılık gelen sabit sayıda ifade değerleri ile belirtilir. Bu aşamada devam edecek olan işlem artık bir örüntü tanıma problemidir ve gen ifade değerlerinden oluşan bir vektör, bilinen kanser sınıflarından birine karşılık gelecek şekilde belirlenmelidir. Örnek sayılarının tüm genlere oranla çok az olması sebebiyle küçük *n* büyük *p* problemi burada açıkça ortaya çıkar (Ramaswamy et al., [8]; Su et al., [9]; [10]; Peng et al., [11];[12]; Lin et al., [13]; Xu et al., [14];Liu and Xu, [15]).

Makine öğrenme dalında veri sınıflandırma ve öznitelik seçimi konularında önemli gelişmeler yaşanırken biyolojik bilimlerde de gen sistemleri konusunda büyük bir gelişme yaşanmıştır. Bu gelişme ile mikroRNA adı verilen minik moleküllerin de gen düzenleme ağlarında tamamlayıcı veya öncü etkileri olduğu kanıtlanmıştır. Bu moleküllerin binlerce genin transkripsiyon sonrası düzenlenmesinde büyük rol oynadıkları artık kesin olarak bilinmektedir (Bartel, [16]). Yakın zamandaki bazı çalışmalarda ise mikroRNA değerlerindeki değişikliklerin tek başına bile kanser tiplerinin sınıflandırılmasında çok yetenekli olduğunu göstermektedir (Lu et al., [17]; Xua et al., [18]; Chan et al., [19]).

Bu çalışmada mikroRNA düzenleme bilgisi ile mRNA bilgisinin veri füzyonu ve öznitelik seçme stratejilerini entegre ederek kanser sınıflandırmasında karşılaşılan küçük *n* büyük *p* problemini ve dolayısıyla karşılaşılan yetersiz sınıflandırma sonuçları aşılmaya çalışılmıştır. Bu amaçla iyi bilinen beş makine öğrenme algoritması ile her biri için beş önemli öznitelik seçme yöntemi bir arada kullanılarak sınıflandırma sonuçlarının en yetersiz olduğu bilinen çok kategorili kanser sınıflandırma problemine uygulanmıştır. Bu problem çeşidinin kanser teşhisinde kullanılmaktan öte, kanser ve gen bilgileri arasındaki henüz tespit edilmemiş ilişkilerin tespiti için çok özel bir yeri olmasından ötürü başarılı bir sınıflandırma sonucu sağlayacak özniteliklerin ve sınıflandırma metotlarının tespit edilebilmesi yüksek önem taşımaktadır. Bu problem için seçtiğimiz yaygın olarak kullanılan ve bilinen bir değerlendirme veri seti üzerinde yaptığımız deneyler sonucu mRNA ve mikroRNA verilerinin bütünleştirilmesinin uygun bir öznitelik azaltma yönteminin de işleme katılmasıyla kanser sınıflandırıcılarının doğruluk oranlarını önemli ölçüde arttırdığı görülmüştür. Bu kapsamdaki deneylerimiz sonucu, ilgili veri seti için literatürde kayıtlı en yüksek doğruluk oranı olan 95.8% değerinin üzerine çıkılmıştır. Bu çalışmanın ardından kanser sınıflandırma da daha başarılı sonuçların elde edilebildiği ve kanser teşhisinde direkt olarak kullanılan kanser alt kategori sınıflandırma problemlerinde önceki çalışmamızda elde ettiğimiz bulguları kullanmayı öngördük. Bu kapsamda gerçekleştirdiğimiz göğüs kanseri alt kategori sınıflandırma deneylerinde yine mikroRNA ve mRNA bilgisi füzyonunun etkili olabildiği, ayrıca mevcut genetik bilgi dahilinde sınıflandırma ve öznitelik seçimi metotlarının optimizasyonu ile hatasız sınıflandırma sonuçlarına ulaşılabileceği görülmüştür.

2 GEREÇLER VE YÖNTEMLER

2.1 Sınıflandırma

Bu çalışmadaki esas problem bilinmeyen bir doku örneğini verilen kanser kategorilerinden biriyle eşleştirebilmek veya kansersiz doku olduğunu belirleyebilmektir. Literatürde çok kategorili sınıflandırma için birçok makine öğrenme algoritması bulunmaktadır. Biyolojik bilimlerdeki çalışmalarda çok kullanılmaları ve diğer alanlarda da kayıt altına alınmış başarıları göz önünde bulundurularak bu algoritmalardan beş tanesi seçilerek ilgili veri seti üzerinde sınıflandırma performansları değerlendirilmiştir. Bu algoritmalar; C4.5 Decision Tree (DT), Artifical Neural Networks (ANN), Support Vector Machines (SVM), Naïve Bayes multinomial sınıflandırıcı (NBM), and K-Nearest Neighbors (KNN) şeklindedir. Bu algoritmalar ile ilgili özet bilgi ve karşılaştırma sonuçlarına Caruana ve Niculescu-Mizil [20] yayınından ulaşılabilir. Daha detaylı anlatımlar için ise Bishop [21] yayını önemli bir kaynaktır.

2.1.1 Bayes sınıflandırıcı

Bir sınıflandırma probleminde tüm özniteliklerin sınıflandırmadaki katkısının eşit ve birbirinden bağımsız olduğu varsayımıyla koşullu olasılık temeline dayanan **'Naive Bayes'** adı verilen basit bir sınıflandırma şekli ortaya çıkmıştır. Buna göre her bağımsız özniteliğin sonuca katkısının analizi ile bir koşullu olasılık değeri belirlenir. Farklı özniteliklerin, sonuç üzerindeki etkilerinin birleşimi ile sınıflandırma işlemi gerçekleştirilir. Bu metoda **'naive'** yani saf denmesinin sebebi farklı özniteliklerin birbirinden bağımsız olduğu varsayımıdır. Verilen bir veri değeri x_i ile ilişkili t_i kaydının C_j sınıfında bulunma olasılığı $P(C_j|x_i)$ şeklinde gösterilir. Eğitim verisi $P(x_i)$, $P(x_i|C_j)$ ve $P(C_j)$ olasılıklarının hesaplanmasında kullanılabilir. Bu değerlerden ise Bayes teoremi sayesinde önce $P(C_j|x_i)$ ve daha sonra da $P(C_j|t_i)$ sonsal olasılıkları hesaplanır [22].

$$\boldsymbol{P(h_1|\, x_i) = \frac{P(x_i|\, h_1)P(h_1)}{P(x_i|\, h_1)P(h_1) + \, P(x_i|\, h_2)P(h_2)}} \tag{2.1}$$

$$\boldsymbol{P(x_i) = \sum_{j=1}^{m} P(x_i|\, h_j)P(h_j)} \tag{2.2}$$

$$\boldsymbol{P(h_1|\, x_i) = \frac{P(x_i|\, h_1)P(h_1)}{P(x_i)}} \tag{2.3}$$

Bir eğitim veri seti için **'Naive Bayes'** algoritması ilk olarak her sınıf için eğitim verisinde bulunma sıklıklarına göre önsel olasılık $P(C_j)$ değerlerini hesaplar. Her öznitelik için x_i değerinin bulunma sıklığı sayılarak $P(x_i)$ elde edilir. Benzer şekilde $P(x_i|C_j)$ değerleri de her x_i değerinin eğitim verisinde bulunan her C_j sınıfında bulunma sıklıkları sayılarak hesaplanır. Bu durumda sadece ilgili öznitelik değerleri için gerekli hesaplamalar yapılmıştır. Eğitim verisindeki bir kayıt için birçok öznitelik tanımlı olabilir ve bu özniteliklerin herbirinin de birçok değeri olabilir. Yukarıda bahsedilen hesaplamalar her öznitelik ve her öznitelik değeri için gerçekleştirilmelidir. Bu hesaplamaların sonuçları eğitim sonrasında yeni bir kayıt sınıflandırılacağı zaman da kullanılmaktadır. Bu nedenle **'Naive Bayes'** sınıflandırma hem betimleyici hem de kestirimci bir algoritma örneğidir [22].

Bir kaydın sınıflandırılmasında eğitim veri setinden hesaplanmış koşullu olasılıklar ve önsel olasılıklar kullanılmaktadır. Bu sayede bu kayda ait farklı öznitelik değerlerinin etkisinin birleşimi hesaplanabilir. Örneğin t_i kaydına ait p kadar bağımsız öznitelik değeri varsa $\{x_{i1}, x_{i2}, \ldots, x_{ip}\}$, betimleyici aşamada her C_j sınıfı için ve x_{ik} özniteliği için hesaplanmış $P(x_{ik}|C_j)$ değerlerinden $P(t_i|C_j)$ hesaplanır (Eşitlik 2.4) [22].

$$P(t_i|C_j) = \prod_{k=1}^{p} P(x_{ik}|\, C_j) \tag{2.4}$$

Algoritmanın bu aşamasında $P(t_i)$ değerini hesaplamak için gerekli olan, her sınıf için $P(C_j)$ önsel olasılıkları ve $P(t_i|C_j)$ koşullu olasılık değeri artık elimizde mevcuttur. $P(t_i)$ değerini hesaplamak için t_i kaydının her sınıfta bulunma sayılarının toplamı kullanılmalıdır. t_i kaydının bir sınıfta olma olasılığı, her öznitelik değeri için bulduğumuz koşullu olasılıkların çarpımı ile bulunabilir. Bu noktada ise her sınıf için sonsal olasılık değeri $P(C_j|t_i)$ bulunur. En yüksek olasılık değerine sahip sınıf bu kayıt için sonuç sınıfı olarak seçilmiş olur [22].

‘**Naive Bayes’** algoritmasının birçok avantajı bulunur. Öncelikle kullanımı çok kolaydır. İkinci olarak ise diğer sınıflandırma algoritmalarının büyük çoğunluğunun aksine eğitim verisinin sadece bir kez taranması yeterlidir. Ayrıca kayıp (boş) değerler de olasılık hesaplarında yok sayılarak kolaylıkla ele alınabilmektedir. Basit ilişkilerin bulunduğu durumlarda çoğunlukla çok iyi sonuç veren bir tekniktir.

Basit kullanımına karşın **‘Naive Bayes’** bazı durumlarda tatmin edici sonuçlar vermeyebilir. Öncelikle öznitelikler bağımsız olmayabilir. Bu durumda öznitelik alt kümeleri kullanılabilir. İkinci olarak ise bu teknik, sürekli değerleri ele alamaz. Bu durumda da sürekli değerleri aralıklara bölmek gerekir. Bu tip sorunlarının yanı sıra ilişkili çözümleri olsa da bu çözümlerin uygulanması kolay olmamakla beraber bu çözümleri uygulama şekli de sonuçları ciddi ölçüde etkiler.

Bu çalışmada kullanılan Bayes sınıflandırıcı **‘Naive Bayes Multinomial’** aslında **‘Naive Bayes’** sınıflandırıcının çalışma prensibini temel alan birçok gerçekleştirimden biridir ve diğerlerinden farklı olarak özniteliklerin sınıflar içindeki dağılımının hesabında sadece çok terimli dağılım kullanır. Bu farklılık, en belirgin olarak, öznitelik değerlerinin bir kayıtta o özniteliğin sayısal olarak ne kadar bulunduğu bilgisini içerdiğinde daha iyi sonuç vermesi şeklinde kendini gösterir. Bu çalışmada da her öznitelik değeri ilgili mRNA veya mikroRNA tipinin örnek içinde bulunma sıklığı olarak değerlendirilebilir. Zaten diğer Bayes sınıflandırıcılarını da kullanarak yaptığımız ön değerlendirme çalışmalarında, bu sınıflandırıcının diğer Bayes sınıflandırıcılardan çok farklı sonuçlar verdiği görülmüştür [22].

Bayes algoritmasının eğitim ve sınıflandırma aşamalarına ait **‘python’** dilinde yazılmış basit bir gerçekleştirim kodu aşağıda verilmiştir:

Bayes eğitim işlemi için örnek kod

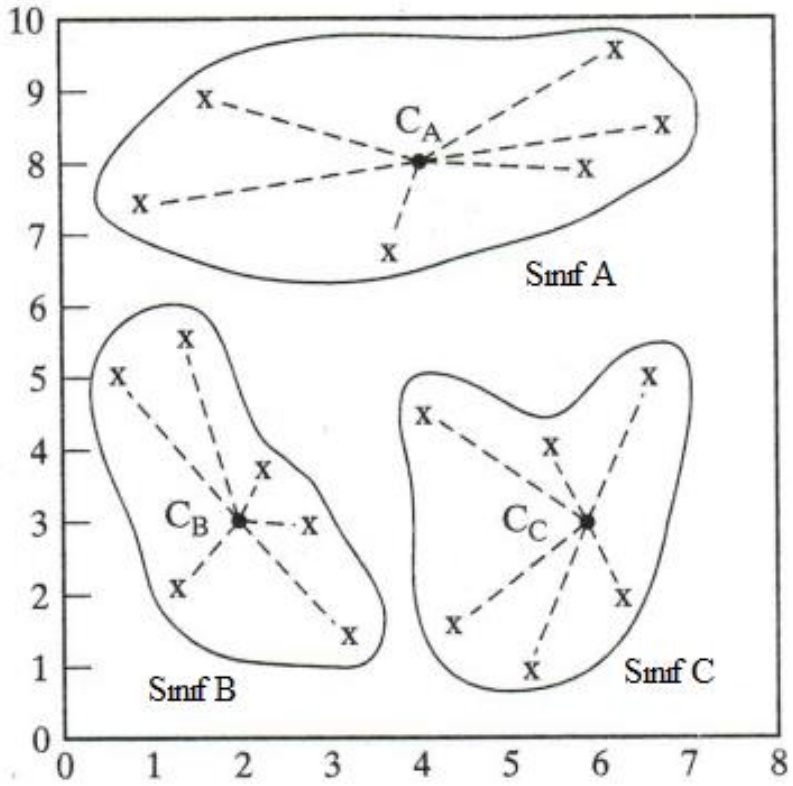
```
def TrainClassifier(self):
    for fv in self.featureVectors:
        self.labelCounts[fv[len(fv)-1]] += 1
        for counter in range(0, len(fv)-1):
            self.featureCounts[(fv[len(fv)-1],
                self.featureNameList[counter],fv[counter])] += 1
            for label in self.labelCounts:
                for feature in self.featureNameList[:len(self.featureNameList)-1]:
                    self.labelCounts[label] += len(self.features[feature])
```

Bayes sınıflandırma işlemi için örnek kod

```
def Classify(self, featureVector):
    probabilityPerLabel = {}
    for label in self.labelCounts:
        logProb = 0
        for featureValue in featureVector:
            logProb += math.log(self.featureCounts[
                (label,self.featureNameList[
                featureVector.index(featureValue)],featureValue)]/
                self.labelCounts[label])
            probabilityPerLabel[label] = (self.labelCounts[label]/sum(
                self.labelCounts.values())) * math.exp(logProb)
    print probabilityPerLabel
    return max(probabilityPerLabel, key = lambda classLabel:
                probabilityPerLabel[classLabel])
```

2.1.2 Uzaklık ölçümü temelli sınıflandırıcılar

Bir problemdeki aynı sınıfa ait örneklerin birbirleri ile başka sınıftaki örneklerle olduklarından daha benzer veya matematiksel ifadeyle daha yakın olduğu temeline dayanan birçok sınıflandırıcı mevcuttur. Bu sınıflandırıcılar arasındaki temel fark sınıf içi ve dışı benzerlik oranını yani örnekler arası uzaklığı ölçme metodudur. Bu ölçüm metotları hem uzaklık ölçümü için seçilen matematiksel yola hem de ölçümün test edilen örneğin sınıfının belirlenmesinde nereye göre uzaklığının ölçüleceğine göre farklılık göstermektedir. En basit yapıdaki bir uzaklık ölçümüne dayalı sınıflandırıcıyı ele alacak olursak, eğitim verisi içindeki farklı sınıflara ait örneklerin sınıflara göre oluşturdukları geometrik bölgelerin orta noktaları bulunmalıdır. Eğitim aşaması sadece bu işlem ile tamamlanmaktadır. Sınıflandırma aşaması ise test için verilen örnek ile tüm sınıfların orta noktaları arası mesafe ölçümü yapıldıktan sonra en yakın mesafedeki sınıfın seçilmesinden ibarettir. Bu durumda sınıflandırma için sadece toplam sınıf sayısı kadar karşılaştırma yapmak yeterlidir [22].



Şekil 2.1 Uzaklık ölçümü temelli sınıflandırıcı

Böyle bir sınıflandırıcı geliştirmek için gerekli algoritma aşağıda verilmiştir [22].

Girdi:

$c_1, \ldots, c_m$ //Her sınıf için orta noktalar
t //Sınıflandırılacak kayıt

Çıktı:

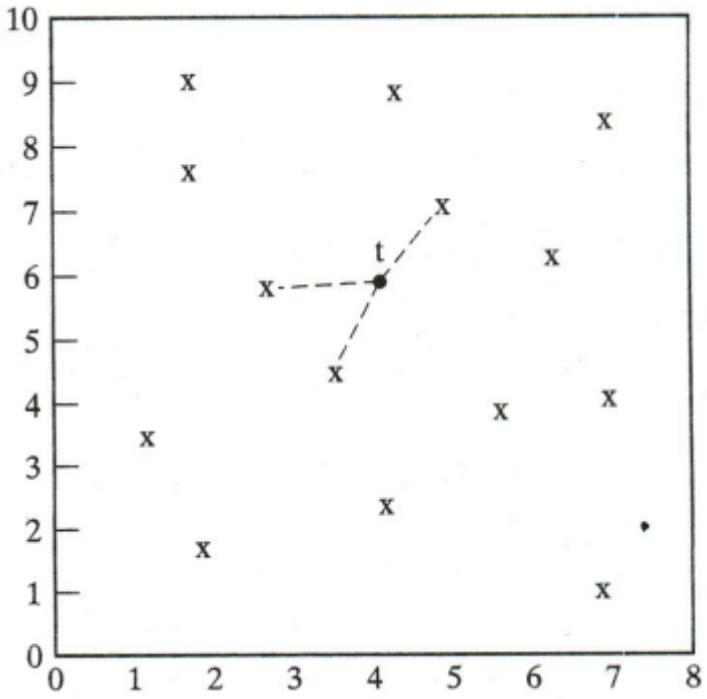
c //t kaydının atandığı sınıf

Uzaklık ölçümü temelli sınıflandırma algoritması:

dist=∞;
for $i \coloneqq 1$ to m **do**
if $dist(c_i, t) < dist$, **then**
$c = i$;
$dist = dist(c_i, t)$;

2.1.2.1 KNN sınıflandırıcı

Uzaklık ölçümü temelli sınıflandırma biçimlerinden en çok kullanılanlarından biri KNN sınıflandırıcıdır. KNN tekniğinde eğitim verisinin sadece veri değerlerinden oluşmadığı bunun yanında ilgili problemde test edilecek herhangi bir verinin sınıflandırma sonucunu da önceden içerdiği varsayılır. Bu varsayım sonucu eğitim verisi sınıflandırma modelinin kendisi olmuş olur. Yani sınıflar için orta nokta veya başka herhangi bir tanımlayıcıya gerek yoktur. Yeni bir örnek sınıflandırılacağı zaman eğitim verisindeki tüm örneklere olan uzaklıkları tek tek hesaplanır. Bu hesaplama sonrası değerlendirme, sadece en yakın olduğu k sayıdaki örneğe göre devam eder. Bu örneklere algoritmanın adını da oluşturan k en yakın komşu denilmektedir. Sınıflandırma sonucu bu en yakın k komşu içerisinde en çok bulunan sınıf olarak belirlenir [22].



Şekil 2.2 KNN sınıflandırıcı

KNN Sınıflandırıcı gerçekleştirimi için gerekli algoritma aşağıdaki gibidir [22]:

Girdi:

```
T            //Eğitim verisi
K            //Komşu sayısı
t            //Sınıflandırılacak kayıt
```

Çıktı:

```
c            //t kaydının atandığı sınıf
```

KNN algoritması:

```
N=∅;
             //t kaydı için N komşu kümesinin bulunması
for each d ∈ T do
   if |N| ≤ K, then
      N = N ∪ {d};
   else if ∃ u ∈ N öyle ki sim(t,u) ≤ sim(t,d), then
      begin
         N = N − {u};
         N = N ∪ {d};
      end
             //Sınıflandırma sonucunun bulunması
c = en çok u ∈ N sayısına sahip sınıf;
```

2.1.3 Karar ağacı sınıflandırıcı

Karar ağaçları yaklaşımı ile sınıflandırmanın temeli arama uzayını dikdörtgensel bölgelere ayırmaktır. Değerlendirilecek kayıt içinde bulunduğu bölgeye göre sınıflandırılacaktır. Farklı bazı yaklaşımlar olsa da bir karar ağacı aşağıdaki gibi tanımlanabilir [22]:

> D={t_1, ….. ,t_n} şeklinde verilen bir veri setinde, kayıt t_i=<t_{i1}, …… , t_{ih}> olsun ve kayıtlar {A_1, ……., A_h} özniteliklerinden oluşsun. İlgili kayıtlara ait sınıflar da C = {C_1, …. , C_m } şeklinde olsun. Bu durumda D ile ilişkili bir karar ağacı şu özelliklere sahip olmalıdır:
>
> - Kök düğüm ve her ara düğüm bir A_i özniteliği ile tanımlanmalı,
> - Her dal kendi başlangıç noktasındaki özniteliğe ait bir değer veya değer aralığını temsil etmeli,
> - Her yaprak düğümü bir C_j sınıfı şeklinde gösterilmelidir.

Bu durumda kök düğümden başlayarak bir yaprak düğüme kadar giden her farklı yol "VE" işlemlerinden oluşan bir kuralı tanımlamış olur. Bu kurallar herhangi bir veriye uygulandığında kurallardan sadece ve mutlaka bir tanesi başarılı bir şekilde sonlanır. Bu kuralı oluşturan yoldaki yaprak düğüm değeri sınıflandırma sonucudur [22].

Kök ve ara düğümlerde kullanılacak özniteliklerin bulunması ve bu özniteliklere ait hangi değer aralıklarının kullanılacağı gibi karar ağaçlarında belirlenmesi gereken önemli noktalar bulunmaktadır. Karar ağaçlarında genel prensip eğitim verisini sırayla en iyi şekilde farklı sınıflara bölen öznitelikleri ve bu özniteliklerin veri setine uygulanacakları düğümlerdeki değer aralıklarını bulmaktır. Seçilen bu özniteliklere ayırıcı öznitelikler denilebilir. Bu kapsamda geliştirilecek temel özelliklere sahip bir karar ağacı sınıflandırıcısı için aşağıdaki algoritma verilebilir [22]:

```
Girdi:
   D            //Eğitim verisi

Çıktı:
   T            //Karar ağacı

Karar ağacı oluşturma algoritması:
```

```
   T = Ø;
   En iyi parçalama kriterini belirle;
   T = En iyi ayırıcı özniteliği ile kök düğüm yarat;
   T = Her aday parça için kök düğüme kenar ekle;
for each kenar do
   D = Veri seti D'ye aday parçanın uygulanması ile oluşan
       yeni veri seti D;
if bu kenar için durma noktasına erişildi, then
   T'= uygun sınıf değeri ile yaprak düğümü oluştur;
else
      T'= Karar ağacı oluşturma algoritması(D);
      T = Kenara T' alt ağacını ekle;
```

Herhangi bir karar ağacı algoritmasının performansını belirleyecek ana unsurlar aşağıdaki gibi sıralanabilir [22]:

- Ayırıcı öznitelikler olarak hangi özniteliklerin kullanılacağı çok önemli bir unsurdur. Bazı öznitelikler diğerlerinden daha ayırıcı yapıdadır. Bu özniteliklerin bulunması için eğitim verilerinin incelenmesinin yanı sıra veri ile ilgili alanda uzmanlık da önemli olabilir.
- Ayırıcı özniteliklerin uygulanma sırası da önemli bir unsurdur. Doğru sıralama ile gerekli olan karşılaştırma işlemi sayısı minimize edilebilir.
- Öznitelik değer aralıklarının belirlenmesi de önemli bir başka unsurdur. Bu aralıkların belirlenmesi bazı öznitelikler için çok kolay olabilirken, özellikle sürekli değerlere sahip özniteliklerde doğru aralıkların bulunması ciddi bir probleme dönüşebilir.
- Ağacın yapısı ise performans için önemli bir faktördür. Genel olarak dengelenmiş ve az sayıda seviyeden oluşan ağaçlar tercih edilse de bu durumda daha kompleks karşılaştırmalar ve dallanmalara sebep olabilmektedir. Bazı algoritmalar bundan ötürü sadece ikili ağaçları kullanmaktadır.
- Normalde ağacın oluşturulması işlemi bütün eğitim verisi doğru olarak sınıflandırıldığında bitecek olsa da bu noktada ağacın istenecek seviyenin çok üzerinde büyüyeceği bazı durumlarda bir durma kriteri belirtilmesi gerekebilir. Bu durum daha yüksek performans için başarı oranından ödün vermek anlamına gelir. Bunun yanı sıra önceden durdurma işlemi **'overfitting'** sorununu engellemek için de gerekli olabilir.

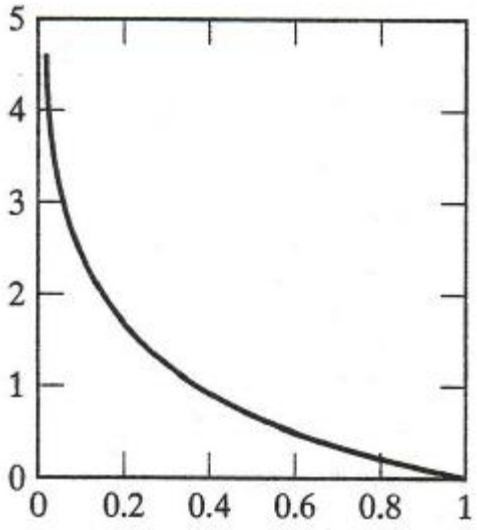
- Eğitim verisinin boyutu da oluşacak ağaç yapısında çok etkilidir. Az sayıda eğitim verisi daha genel yapıdaki farklı veriler için yeterince spesifik bir ağaç oluşturamayabilecekken çok sayıda eğitim verisi ise **'overfitting'** problemine neden olabilir.

Eğitilmiş bir ağaçta da performans artışı için bazı modifikasyonlar yapılabilir. Bunların başında budama işlemi sayılabilir. Budama ile ağaçta oluşmuş gereksiz dallanmalar veya alt ağaçlar giderilebilir.

2.1.3.1 ID3 ile karar ağacı oluşturma

Karar ağaçlarının oluşturulmasında kullanılan yöntemlerden biri olan ID3 tekniği bilgi teorisine dayanır ve gerekli karşılaştırma işlemi sayısını minimize etmeye çalışır. Bunun için ID3 ayırıcı özniteliklerin seçiminde en yüksek bilgi kazancı değerine sahip özniteliği seçme stratejisini kullanır. Bir özniteliğin bilgi değeri bu özniteliğin veri setinde görülme sıklığı ile bağıntılıdır [23].

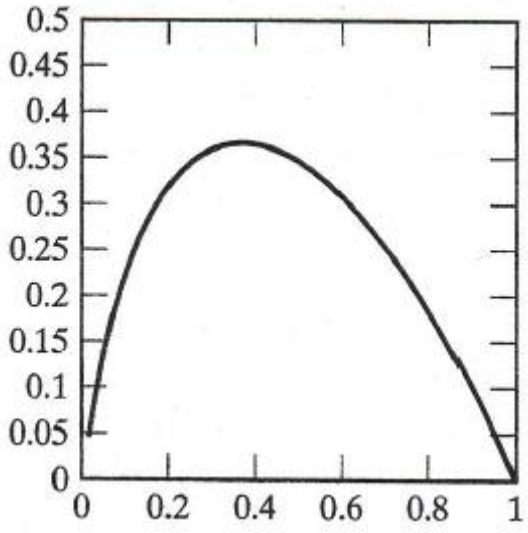
Bilgiyi ölçmek için kullanılan kavrama entropi (bilgi yitimi) adı verilir. Entropi bir veri setindeki belirsizlik oranını ölçmeye yarar. Mantık olarak bir veri setindeki tüm örnekler tek bir sınıfa ait ise belirsizlik yoktur. Bu durumda entropi değeri sıfır değerine sahiptir. Eğer bir olayın gerçekleşme olasılığı p = 1 değerine sahip ise bu olayın gerçekleşmesinde bir belirsizlik olmadığı söylenir. Bu olasılık sıfır değerine yaklaştıkça ise belirsizlik artar. Olasılığa bağlı olarak belirsizlik değerinin değeri log(1/p) formülü ile gösterilmektedir (Şekil 2.3a) [22].



(a)

1. log(1/p)

Karar ağaçlarının oluşturulmasında izlenilen böl ve fethet yaklaşımı sonucu toplamları bire eşit olan olasılık değerleri oluşur. Böyle bir bölümleme işlemi ile ilgili bilgi değerlerini ölçebilmek için problemdeki tüm olaylara ait bilgilerin birleştirilebilmesi gerekir. Bunun için bölümlemedeki ortalama bilgi değerinin hesaplanması gerekir. Bir olaya ait bilgi değeri p log (1/p) formülü ile gösterilebilir (Şekil 2.3b) [22].

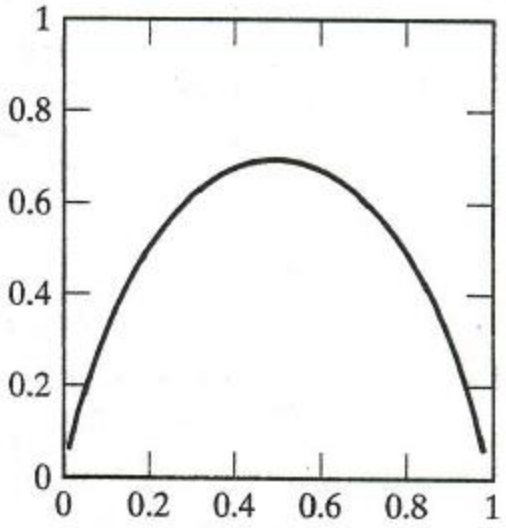


(b)

2. p log(1/p)

İki olaylı bir problemi ele alırsak ortalama bilgi değeri H(p, 1-p) her iki olaya ait bilgi değerlerinin toplamı şeklinde gösterilir. Bu ortalama değerine ait olası değerleri gösteren fonksiyon p log(1/p) + (1-p)log (1/(1-p)) şeklindedir ve Şekil 2.3c'de verilen çizimden de görüleceği üzere en büyük değerini iki olasılık değeri birbirine eşit olduğunda almaktadır. Bu durum bir özniteliğin alabileceği değer aralıkların ilgili veri setindeki olasılıklarının eşit olması durumunda bu veri setini eşit parçalara böleceğini ve bilgi değerinin maksimum olacağını gösterir [23].

Şekil 2.3 devam ediyor



(c)

3. H(p,1-p)

Şekil 2.3 Entropi

Entropinin genel formülü eşitlik 2.5'te verilmiştir [22].

$$(p_1, p_2, \dots, p_s) \text{ olasılıkları için } \sum_{i=1}^{s} p_i = 1 \; ise \; entropi:$$

$$H(p_1, p_2, \dots, p_s) = \sum_{i=1}^{s} (p_i \log(1/p_i)) \tag{2.5}$$

Bir veri setinin herhangi bir zamandaki durumunu D ile gösterecek olursak H(D) entropi değeri bu durumda veri setinin ne kadar sıralı olduğunu gösterir. Eğer bu durumdaki veri setini s kadar yeni duruma parçalarsak, parçalanmış bu durum S = { D_1, D_2, …, D_s } şeklinde gösterilebilir. Bu durumda her durumun ayrı ayrı entropileri hesaplanmalıdır. ID3'teki her adımda o esnadaki parçalanmış durumu en iyi sıralayan durum parçası seçilir. Bir durum eğer içindeki tüm örnekler aynı sınıftaysa tam olarak sıralıdır. ID3 en yüksek bilgi kazancına sahip parçalanmayı seçer. Burada bilgi kazancı, doğru sınıflandırma yapmak için parçalama öncesi gereken bilgi miktarı ile parçalama sonrası gereken bilgi miktarının farkıdır (Eşitlik 2.6) [22].

$$Kazanç(D, S) = H(D) - \sum_{i=1}^{s} P(D_i H(D_i) \tag{2.6}$$

2.1.3.2 C4.5 ile karar ağacı oluşturma

C4.5 karar ağacı oluşturma algoritması ID3 algoritmasını birçok yönden geliştiren daha kompleks bir algoritmadır. Bu geliştirmeler aşağıdaki gibi sıralanabilir [22]:

- Eksik veriler basitçe yok sayılır. Kısacası kazanç oranı bu öznitelikler için sadece değere sahip olan kayıtlara bakılarak hesaplanır. Bu öznitelikler için değeri olmayan kayıtlar sınıflandırılırken diğer kayıtlardaki mevcut değerlerden bir çıkarım yapılır.
- Sürekli veriler için temel işlem eğitim verisinde bulunan değerlere ve bilgi kazancına bağlı olarak değer aralıklarını belirlemektir.
- Budama işlemi için tanımlı iki yöntemi kullanır. Bunlar alt ağaç değişimi ile alt ağaç yükseltme yöntemleridir.
- C4.5 karar ağaçlarından kuralların çıkarımında kuralları basitleştirici bazı teknikler kullanır.
- ID3'te bilgi kazancının kullanılmasından ötürü **'overfitting'** problemi ile karşılaşılabilir. Çok fazla değere bölünmüş özniteliklerin bilgi kazancında gereksiz yere değerli görülmesinden kaynaklanan bu durum, C4.5'te özniteliklerin entropilerini de dikkate alan kazanç oranı kullanılarak giderilmeye çalışılır. Kazanç oranının C4.5'te veri seti parçalarına uygulama şekli eşitlik 2.7'deki gibi gösterilir.

$$KazançOranı(D,S) = \frac{Kazanç(D,S)}{H\left(\frac{|D_1|}{|D|},\dots,\frac{|D_s|}{|D|}\right)} \tag{2.7}$$

2.1.4 Yapay sinir ağları tabanlı sınıflandırıcılar

Yapay sinir ağları sınıflandırıcı olarak kullanıldıklarında verilen bir örneğin mevcut sınıflarda olma olasılıklarını çıktı olarak verecek şekilde bir model oluşturulur. Bu model uygulandığında sonuç olarak en yüksek değeri veren çıkış, sonuç sınıfı olarak kabul edilir. Eğitim aşamasında ise her veri yapay sinir ağı içerisine verildiğinde çıkan sonuç ile beklenen sonucun karşılaştırılmasına dayanarak ağ içerisindeki ağırlık değerleri değişir. Yapay sinir ağı ile sınıflandırma problemi aşağıdaki adımlardan oluşur [22]:

1. Sınıflandırıcıda girdi için kullanılacak öznitelikler ve çıktı sayılarının belirlenmesi.
2. Gizli katman sayılarının ve gizli katmanlardaki gizli düğüm sayılarının ilgili problem alanına bağlı olarak belirlenmesi.
3. Ağda kullanılacak ağırlıkların ve fonksiyonların belirlenmesi.
4. Eğitim veri setinin belirlenmesi. Gereğinden fazla eğitim verisi **'overfitting'**, az sayıda eğitim verisi ise **'underfitting'** problemine yol açabilir.
5. Öğrenme tekniğinin belirlenmesi. Bu teknik ağırlıkların nasıl düzenleneceğini belirler. Çoğunlukla geri yayılım (backpropagation) formunda bir yaklaşım izlenir.
6. Durma koşulunun belirlenmesi. Tüm eğitim verisi ağ üzerinde değerlendirildiğinde veya istenilen bir hata oranı ya da belirlenen bir zaman kriterine ulaşıldığında öğrenme işlemi sonlandırılabilir.
7. Durma koşulu sağlanana kadar her eğitim verisinin ağda değerlendirilmesi ile bulunan sonucun istenilen sonuç ile karşılaştırılması. Eğer sonuç istenen şekildeyse benzer bir veri geldiğinde de aynı sonucun verilme olasılığını arttıracak şekilde ağırlıkların düzenlenmesi. Sonuç istenenden farklıysa bu sonucun tekrar verilme olasılığını azaltacak ağırlık değişiklikleri yapılması.
8. Test edilecek verilerin eğitilmiş ağda değerlendirilmesi sonucu sınıflandırılmaları.

2.1.4.1 Yayılma (propagation) tekniği

Yapay sinir ağlarında öğrenme işlemi için genel olarak izlenilen yönteme yayılma adı verilir. Bu işleme örnek olarak tek gizli katmana sahip bir yapay sinir ağının kullanıldığı algoritma aşağıda verilmiştir. Bu örnekte gizli katman düğümleri için hiperbolik tanjant aktivasyon fonksiyonu kullanılmış ve çıkış düğümlerinde ise sigmoid fonksiyon kullanılmıştır. c sabiti ile tanımlanan öğrenme oranı (learning rate) kullanıcı tarafından sağlanmalıdır. k değeri ise bir düğüme giren kenar sayısıdır [23].

Girdi:

```
N              //Yapay sinir ağı
X=⟨x1,…,xh⟩    //Sadece giriş öznitelik değerlerinden oluşan
               girdi
```

Çıktı:

```
   Y=⟨y_1, …, y_m⟩   //Yapay sinir ağı çıkış değerlerinden oluşan
                     çıktı
Yayılma algoritması:
                     //Bir kaydın yapay sinir ağı içindeki sürecini
                     gösteren algoritma
     for each gizli katman do
          for each düğüm i do
          S_i = (∑_{j=1}^{k} (w_ji x_ji));
          for each i den çıkan kenar do
                Çıktı (1−e^{−S_i}) / (1+e^{−cS_i});
     for each çıkış katmanındaki i düğümü do
          S_i = (∑_{j=1}^{k} (w_ji x_ji));
          Çıktı y_i = 1 / (1+e^{−cS_i});
```

2.1.4.2 Gözetimli öğrenme

Birçok öğrenme algoritmasında olduğu gibi yapay sinir ağları ile sınıflandırma için öğrenme işlemi de verilen bir örnek için istenen sınıf sonucunun sınıflandırıcıyı eğitmek üzere sağlanması ve sınıflandırıcı performansının bu bilgiyi kullanarak eğitim esnasında gelişmesi esasına dayanır. Bu şekilde öğrenme yöntemine gözetimli öğrenme (supervised learning) adı verilir. Yapay sinir ağları için gözetimli öğrenme işlemine örnek olarak aşağıdaki algoritma verilebilir [22].

```
Girdi:
   N          //Başlangıçtaki yapay sinir ağı
   X          //Eğitim veri setinden bir kayıt
   D          //İstenen çıkış kaydı
Çıktı:
   N          //İyileştirilmiş yapay sinir ağı
Gözetimli öğrenme algoritması:
              //Yapay sinir ağı ile öğrenme için basit bir
              örnek oluşturan algoritma
   X'i N içinde ilerleterek Y çıktısını oluştur;
   D ile Y'yi karşılaştırarak hatayı hesapla;
   N'deki kenarların ağırlıklarını düzenleyerek hatayı azalt;
```

Bu algoritmada hata oranını hesaplamak ve ağırlıkları ayarlamak için bazı yöntemlere ihtiyaç duyulduğu görülmektedir. Hata oranı hesabı için çok sayıda yöntem bulunmakla beraber temel olarak ortalama karesel hata yani MSE (Mean squared error) kullanılır. Bir i düğümü için beklenen çıktı d_i ve gerçek çıktı y_i olarak gösterilirse herhangi bir katmanda bulunan bu düğüm için hata, eşitlik 2.8 ile

gösterilir. Bu durumda MSE eşitlik 2.9 ile hesaplanır. Bir yapay sinir ağındaki m sayıda çıkış düğümü için toplam MSE ise eşitlik 2.10 de gösterilmiştir [22].

$$|y_i - d_i| \tag{2.8}$$

$$\frac{(y_i - d_i)^2}{2} \tag{2.9}$$

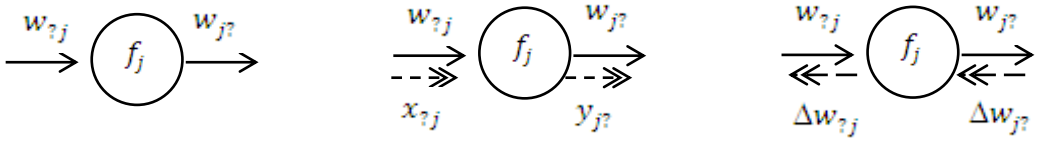
$$\sum_{i=1}^{m} \frac{(y_i - d_i)^2}{m} \tag{2.10}$$

Öğrenme tekniğinin amacı bir kayıt için giriş değerleri sonucu elde edilen çıkış değerine bağlı olarak ağırlıkları değiştirmektir. Bu işleme öğrenme kuralı adı verilir ve temel olarak iki yöntem kullanılır. **'Hebb'** kuralı ve **'Delta'** kuralı olarak bilinen bu yöntemlerin her ikisi de çıkış değerinin hatalı olması durumunda ağırlıkları değiştirme esasına dayanır. Verilen bir j düğümü için ağırlıklar <w_{1j}, … , w_{kj}>, giriş değerleri <x_{1j}, … , x_{kj}> ve çıkış değeri y_j ise **'Hebb'** öğrenme kuralı eşitlik 2.11 ile gösterilir. Öğrenme oranı olarak adlandırılan c sabiti için genel olarak 1 / (eğitim veri sayısı) değeri kullanılır. **'Delta'** kuralı ise **'Hebb'** kuralından farklı olarak sadece y_j çıkış değerini değil d_j beklenen değeri de göz önüne alarak ağırlık değişimini gerçekleştirir (Eşitlik 2.12) [22].

$$\Delta w_{ij} = c x_{ij} y_j \tag{2.11}$$

$$\Delta w_{ij} = c x_{ij} (d_j - y_j) \tag{2.12}$$

Geri yayılma (back propagation) **'Delta'** kuralı kullanılarak ve çıkış düğümünden kaynak düğümlere geriye doğru dolaşarak her düğümdeki hatayı azaltan bir yöntemdir. Şekil 2.4 (a), (b), (c) sırasıyla bir yapay sinir ağındaki tek bir düğümün genel yapısını, bu düğüm üzerinde yayılma tekniğinin kullanımını ve son olarak geri yayılma tekniğinin ağırlık değerlerini geriye doğru değiştirmesini göstermektedir [23].



(a) (b) (c)

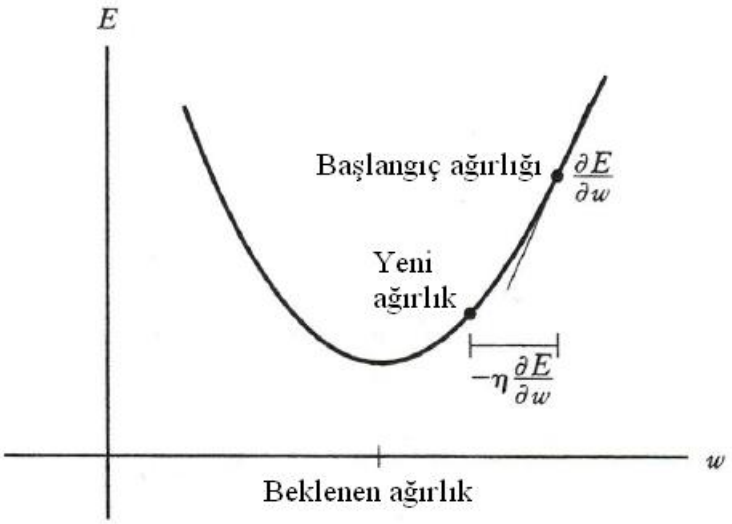
Şekil 2.4 Yapay sinir ağları kullanımı

Geri yayılma için temel algoritma ise aşağıdaki algoritmada gösterilmiştir [22].

Girdi:
N //Başlangıçtaki yapay sinir ağı
X=$\langle x_1, \dots, x_h \rangle$ //Eğitim veri setinden bir kayıt
D=$\langle d_1, \dots, d_m \rangle$ //İstenen çıkış kaydı
Çıktı:
N //İyileştirilmiş yapay sinir ağı
Geri yayılma algoritması:
//Geri yayılma için basit bir örnektir.
Yayılma(N,X);
E=$1/2 \sum_{i=1}^{m} (d_i - y_i)^2$;
Gradient(N,E);

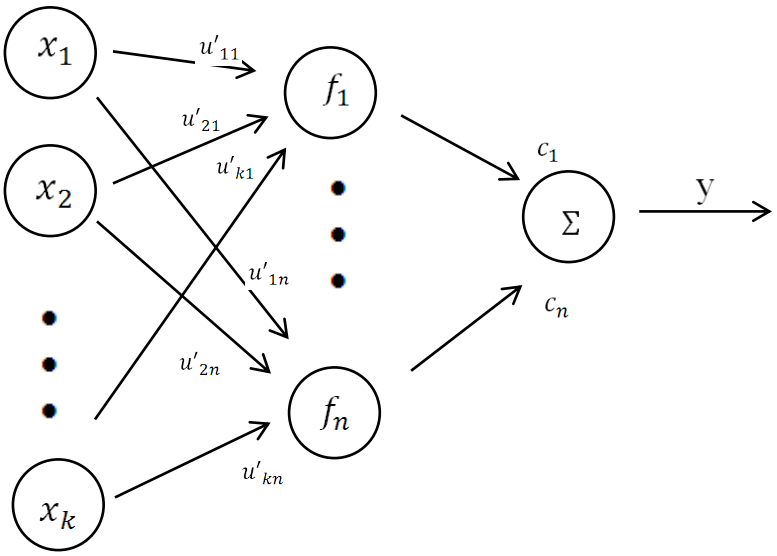
Bu algoritmada kullanılan **'gradient descent'** tekniği ile ağırlıklar düzenlenmektedir. Bu tekniğin temel mantığı MSE değerini minimize edecek ağırlıklar kümesini bulmaktır. Bu tekniğin çalışma prensibi aşağıdaki algoritmada verilmiştir. Hata oranına karşılık ağırlık değeri grafiği ve **'gradient descent'** tekniğinin bu grafikte eğimin sıfır olma durumuna kadar uygulanışı Şekil 2.8'de gösterilmiştir [24].

Girdi:
N //Başlangıçtaki yapay sinir ağı
E //Yayılma algoritmasından hesaplanan hata
Çıktı:
N //İyileştirilmiş yapay sinir ağı
Gradient algoritması:
//Arttırımlı gradient descent için basit bir örnek
for each çıkış katmanındaki düğüm *i* **do**
for each *i* düğümüne girişi olan *j* düğümü **do**
$\Delta w_{ji} = \eta(d_i - y_i) y_j (1 - y_i) y_i$;
$w_{ji} = w_{ji} + \Delta w_{ji}$;
katman = önceki katman;
for each bu katmandaki her *j* düğümü **do**
for each *j* düğümüne girişi olan her *k* düğümü **do**
$\Delta w_{kj} = \eta y_k \frac{1-(y_j)^2}{2} \sum_m (d_m - y_m) w_{jm} y_m (1 - y_m)$;
$w_{kj} = w_{kj} + \Delta w_{kj}$;



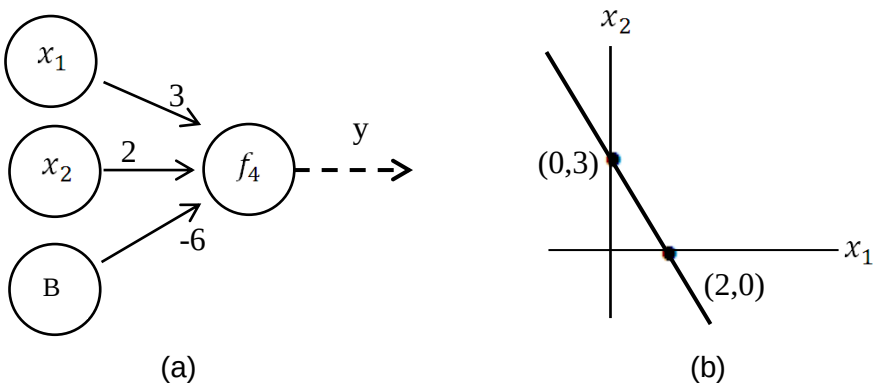
Şekil 2.5 Gradient descent

RBF (radial basis function) yani radyal bazlı fonksiyon bir merkezden uzaklaştıkça değeri artan veya azalan fonksiyon türüdür. RBF **'Gauss'** şeklindedir ve RBF ağı üç katmandan oluşan tipik bir yapay sinir ağıdır. Giriş katmanı klasik olarak sadece veri girişinden sorumludur. Gizli katmanda **'Gauss'** aktivasyon fonksiyonu kullanılırken çıkış katmanında doğrusal bir aktivasyon fonksiyonu kullanılır. RBF kullanıldığında gizli katman düğümleri giriş değerlerinin sadece bazı alt kümelerine duyarlıdır. Tipik bir RBF ağı Şekil 2.6'da verilmiştir [24].



Şekil 2.6 Radyal bazlı fonksiyon

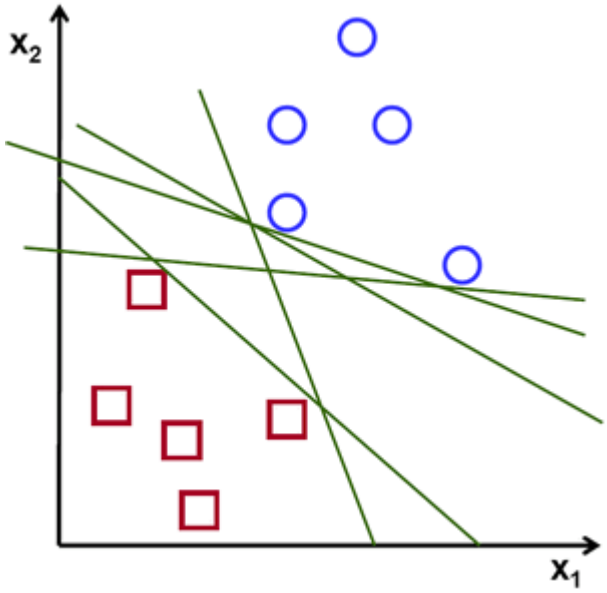
En basit yapıdaki yapay sinir ağına **‘perceptron’** adı verilir. **‘Perceptron’** tek bir sinir hücresi görünümünde, çok giriş ve tek çıkışa sahip bir algılayıcıdır. Bu basit algılayıcı temel olarak iki sınıflı bir sınıflandırma yapmada kullanılabilir. Şekil 2.7 (a) ve (b) böyle bir sınıflandırmaya örnek olarak verilebilir. Çok sayıda **‘perceptron’** birarada kullanılarak MLP (Multilayer perceptron) adı verilen çok sınıflı sınıflandırma yapma kabiliyetine sahip yapay sinir ağı bu çalışmada tercih edilmiştir [24].



Şekil 2.7 Perceptron sınıflandırma örneği

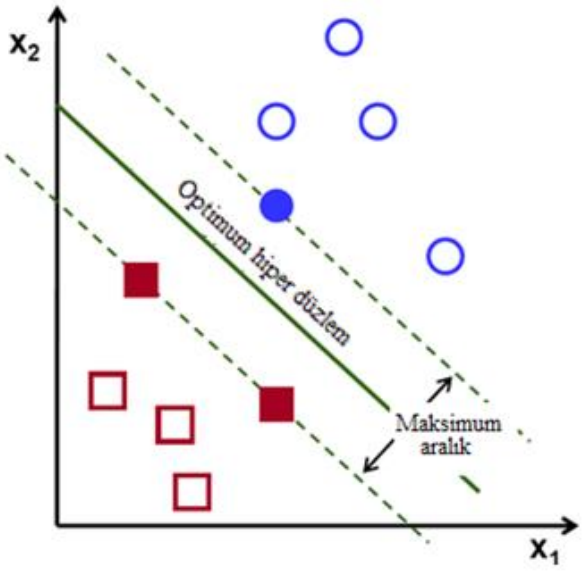
2.1.5 SVM sınıflandırıcı

SVM problem uzayını çeşitli parçalara ayıran hiper düzlemlerle tanımlanan ayrıcalıklı bir sınıflandırıcıdır. Diğer bir deyişle verilen bir eğitim veri seti ile gözetimli öğrenme sonucu ilgili verileri ayıran optimum hiper düzlemleri bulmaya yarayan bir algoritmadır. Optimum hiper düzlemin ne olduğu SVM’in ayrıcalığını oluşturan ve sınıflandırma performansını belirleyen temel unsurdur. İki sınıflı ve iki boyutlu bir problemi ele alacak olursak problem uzayını doğru sınıflandırmayı yapacak şekilde ayıran çok sayıda hiper düzlem olabilir. Bu durum Şekil 2.8’de gösterilmiştir [25].



Şekil 2.8 Ayırıcı hiper düzlemler

Normalde çok sayıda sınıf ve çok fazla sayıda öznitelikten oluşmuş çok boyutlu problemlere de rahatlıkla uygulanabilen SVM'in daha kolay anlaşılabilmesi için kullanılan bu iki boyutlu örnekte çok boyutlu uzaydaki vektörler noktalarla, hiper düzlemler ise doğrularla gösterilebilmektedir. Şekil 2.8'de rahatlıkla göründüğü üzere problem uzayını, tüm örnekleri başarılı sınıflandıracak şekilde ayıran birçok doğru bulunabilir. Bu doğrulardan hangisinin diğerlerinden daha iyi yani optimum olduğunu belirlemek için bir kritere ihtiyaç duyulmaktadır. Eğer bir doğru, noktalara çok yakın ise basitçe bu doğrunun çok iyi olmadığı söylenebilir. Çünkü bu doğru kullanılarak yapılacak sınıflandırma gürültüye daha hassas olacak ve yeterince genelleyici olamayacaktır. Bu yaklaşıma göre en iyi ayırıcı doğru her noktadan mümkün olduğunca uzaktan geçen olmalıdır. Çok boyutlu uzaya dönecek olursak bu durumda SVM'in çalışma mantığı eğitim verilerinden kendine en yakın örnek için en uzak mesafaye sahip ayırıcı hiper düzlemi bulmaktır. Bu mesafe SVM teorisinde marjin adıyla anılmaktadır. Şekil 2.9'da aynı örnek için maksimum marjine sahip optimum hiper düzlemin bulunma yöntemi basitçe gösterilmiştir [25].



Şekil 2.9 SVM için optimum hiper düzlem

Eşitlik 2.13 ile bir hiper düzlem ifade edilebilir. Bu eşitlikte β ağırlık vektörü ve β_0 öngörü değeridir [25].

$$f(x) = \beta_0 + \beta^T x \tag{2.13}$$

Optimum hiper düzlem farklı β ve β_0 değerleri kullanılarak sonsuz farklı şekilde ifade edilebilir. Ortak bir düzen sağlayabilmek adına SVM için optimum hiper düzlemin ifadelerinden eşitlik 2.14 ile verilen kullanılır [25].

$$|\beta_0 + \beta^T x| = 1 \tag{2.14}$$

x hiper düzleme en yakında bulunan eğitim verilerini sembolize etmektedir ve bunlara destek vektörleri adı verilir. x ile (β, β_0) hiper düzlemi arasındaki uzaklık için geometrik tanım eşitlik 2.15 ile gösterilmiştir [25].

$$uzaklık = \frac{|\beta_0 + \beta^T x|}{\|\beta\|} \tag{2.15}$$

Eşitlik 2.14 göz önünde bulundurulduğu takdirde destek vektörleri için bu uzaklık formülü eşitlik 2.16'da gösterildiği gibi olur [25].

$$uzaklık_{destek\ vektörleri} = \frac{|\beta_0 + \beta^T x|}{\|\beta\|} = \frac{1}{\|\beta\|} \tag{2.16}$$

Bu durumda marjin M eşitlik 2.17'de verildiği üzere en küçük uzaklık değerinin iki katına eşit olacaktır [25].

$$M = \frac{2}{\|\beta\|} \tag{2.17}$$

Bu bilgiler ışığında M'yi maksimize etme problemi aslında bazı kısıtlara bağlı olarak $L(\beta)$ mesafesini minimize etmek şeklinde gösterilebilir. Bu kısıtlar tüm x_i eğitim verilerini doğru sınıflandırmak üzere belirlenir ve genel bir ifadeyle eşitlik 2.18'de verilmiştir [25].

$$\min_{\beta,\beta_0} L(\beta) = \frac{1}{2}\|\beta\|^2 \ \ öyle\ ki\ y_i(\beta^T x_i + \beta_0) \geq 1\ \forall i \tag{2.18}$$

Bu eşitlikte y_i eğitim verisindeki her bir sınıfı gösterir. Bu aslında bir **'Lagrange'** optimizasyon problemidir ve **'Lagrange'** çarpanları kullanılarak çözülür [25].

2.2 Öznitelik Seçimi

Öznitelik seçimi tüm öznitelikler içinden azaltılmış ve muhtemelen daha iyi sınıflandırma performansına sahip bir öznitelik alt kümesinin bulunması işidir. mRNA bilgisi kullanılarak doğruluk oranı yüksek ve güvenilir kanser sınıflandırma sonuçları elde etmek söz konusu olduğunda öznitelik seçiminin kritik bir gereklilik olduğu kanıtlanmıştır (Guyon et al., [26]; Cai et al., [27]). Bu çalışma için benzer problemlerdeki başarılarını da göz önünde bulundurarak pek çok öznitelik seçme algoritması değerlendirilmiş ve sonuçta aralarından beş tanesi seçilmiştir. Bunlar; SVM öznitelik seçimi, bilgi kazancı kullanılarak öznitelik seçimi, kazanç oranı kullanılarak öznitelik seçimi, Correlation Based Feature Subset (CFS) öznitelik seçimi, Ki-Kare öznitelik seçimi şeklindedir (Saeys et al., [4]; Guyon et al., [26]; Hall, [28]).

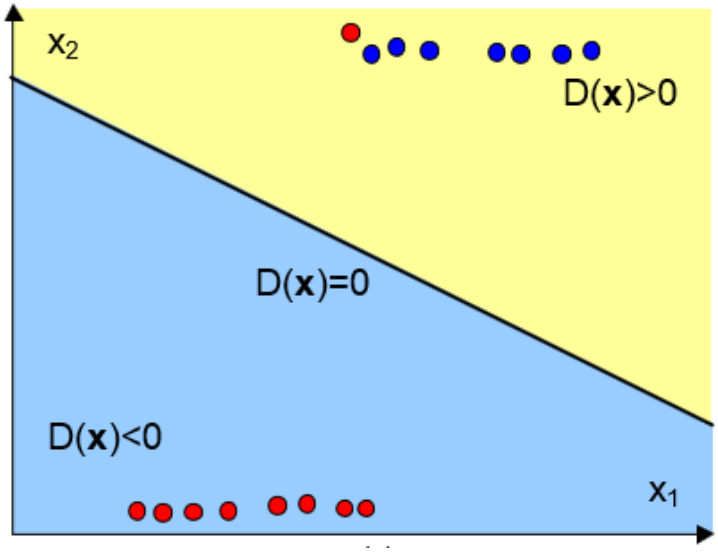
Bu öznitelik seçimi metotlarının tümü WEKA (Waikato Environment for Knowledge Analysis) makine öğrenme araç seti üzerinde ve özel bir parametre optimizasyon yöntemi olmaksızın çoğunlukla ön tanımlı parametreleriyle kullanılmıştır. Bu seçim metotları sonucu çoğunlukla tüm özniteliklerin ilgili metot tarafından belirlenen bilgi değerlerine göre sıralamaları oluşmaktadır. Bu durumda ilgili değerlendirme sonuçları kapsamında kaç tane özniteliğin kullanılacağına da karar verilmesi gerekmektedir. Bunun için tarafımızdan gerçekleştirilen kullanıcı destekli seçimlere ve aç gözlü yöntemlere dayalı yarı otomatik deneyler ve WEKA'da gerçekleştirmek istenilen işlemleri otomatize etmeyi kolaylaştıran deney ara yüzünde **(WEKA Experimenter)** tanımlı çeşitli metotlar kullanılmıştır.

2.2.1 SVM ile öznitelik seçimi

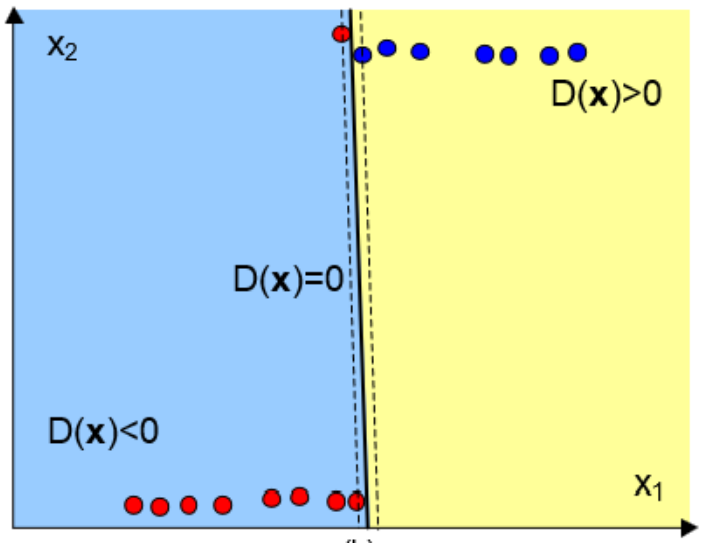
Şekil 2.10 ve Şekil 2.11'de iki boyutlu ve iki sınıflı bir sınıflandırma problemi örneği verilmiştir. Örnekleri temsil eden noktaların yerleşimine göre x_2 özniteliği tek başına kullanıldığında problem uzayı sadece tek bir hatalı sınıflandırma olacak şekilde ve küçük bir varyans ile doğrusal ayrılabilmekteyken, x_1 özniteliği kullanıldığında çok yüksek bir varyans ile de olsa tüm örnekler doğru sınıflandırılabilmektedir. İstatistiksel tabanlı birçok sınıflandırıcı sınıf ortalamalarını baz alarak problem uzayını ayırır ve bu tarz bir durumda öznitelik seçimi için kullanıldıklarında Şekil 2.10'da da görüleceği üzere küçük varyanslı öznitelik olan x_2'yi x_1 özniteliğine tercih ederler. Halbuki SVM sınıf ortalamarından bağımsız olarak en iyi sınıflandırma sonucunu sağlayan en büyük marjinli hiper düzlemi seçtiğinden öznitelik seçimi için kullanıldığında bu yanılgıya düşmez ve Şekil 2.11'de görüldüğü üzere x_1 özniteliği seçilir [29].

Bu seçimlerin temelinde problem uzayını ayıran düzlemin durumu etkilidir. Bu düzleme kısaca karar hattı denebilir ve $D(x) = 0$ şeklinde gösterilir. Buna göre $D(x) > 0$ durumunu sağlayan bir örnek bir sınıfta iken $D(x) < 0$ durumunu sağlayan bir örnek ise karşı sınıftadır. Karar hattına göre hangi özniteliğin tercih edilmesi gerektiği ise basit bir geometrik işlemle bulunabilir. Buna göre karar hattının eğimi 45 dereceden büyük ise x_1 yoksa x_2 özniteliği seçilmelidir [29]. SVM'in bu yaklaşımı özellikle çok öznitelikli yani çok boyutlu sınıflandırma problemlerinde daha genelleyici bir sonuç elde etmeyi ve gürültüden daha az etkilenmeyi sağlamasına

karşın iki özniteliğin birbirine göre değerini baz alarak öznitelik seçimi yapmada kullanıldığında ise tersi bir durum söz konusu olabilir. Bu durum verilen örnekte de açıkça görülebilir. SVM ile x_1 özniteliğinin seçilmiş olması her ne kadar daha doğru bir yaklaşım gibi görünse de mavi sınıf örneklerinin çok yakınında bulunan ve kendi sınıf örneklerine de çok uzak olan kırmızı sınıf örneği bu seçimde etkili olmuştur. Bu örneğin gürültü olması durumunda x_2 özniteliği seçilmeliyken x_1 özniteliği hatalı olarak seçilmiş olur. Bu da SVM öznitelik seçiminde SVM ile sınıflandırmaya göre gürültünün çok daha kritik bir önemi olduğu anlamına gelmekte ve bu öznitelik seçiminin buna uygun veri setlerinde uygulanması gereğini ortaya koymaktadır.



Şekil 2.10 İstatistiksel tabanlı sınıflandırma örneği



Şekil 2.11 SVM tabanlı sınıflandırma örneği

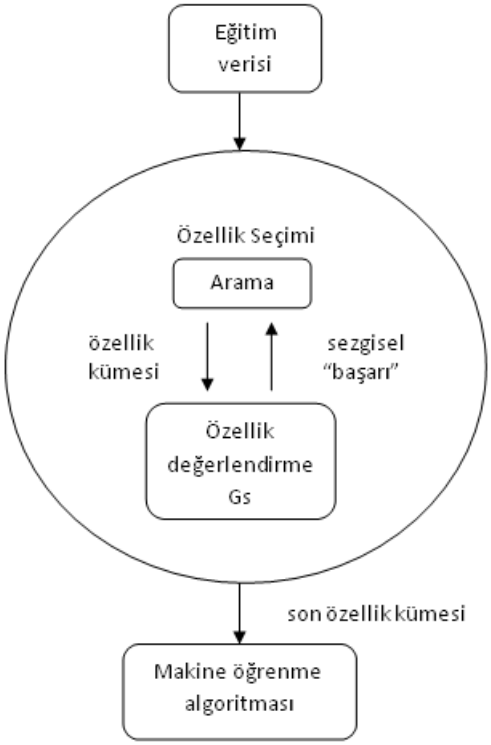
2.2.2 CFS öznitelik seçimi

Diğer öznitelik seçimi algoritmalarının bir çoğunda olduğu gibi CFS metodu da öznitelik alt kümelerinin bilgi değerlerini ölçen bir fonksiyonun yanında bir arama algoritması da kullanır. CFS'nin öznitelik altkümelerinin değerini ölçerken kullandığı sezgisel yaklaşım her özniteliğin sınıf etiketlerini tahmin etmedeki bireysel yeteneklerinin yanı sıra aralarındaki iç korelasyonu da dikkate alır. Bu yaklaşımın temel olarak aldığı hipoteze göre; iyi öznitelik altkümeleri ilgili sınıf ile yüksek, birbirleri ile ise düşük korelasyona sahip özniteliklerden oluşurlar [29].

$$G_s = \frac{k\overline{r_{cı}}}{\sqrt{k + k(k-1)\overline{r_{ıı'}}}} \tag{2.8}$$

Eşitlik 2.8, çıkış noktası toplanmış varlıkların bulunduğu bir testin güvenilirliğinin tekil varlıkların güvenilirliğine göre ölçülmesinde kullanıldığı test teorisidir (Ghiselli, [30]). Örneğin bir insanın eğitimindeki başarısının ölçülmesinde birçok farklı becerisinin ölçüldüğü birden fazla farklı testin bileşkesi, sınırlı sayıda becerinin aynı anda ölçüldüğü tek bir testten daha doğru bir ölçüt olabilmektedir. Bu eşitlikte *k* veri alt setindeki öznitelik sayısı, r_{ci} ortalama öznitelik korelasyonu ve $r_{ii'}$ ortalama öznitelik iç korelasyonudur.

Eşitlik 2.8 aslında tüm değişkenlerin standartlaştırıldığı **'Pearson korelasyonu'**dur. Eşitlikteki payın bir grup özniteliğin sınıf üzerindeki tahmin yeteneğini, paydanın ise bu özniteliklerin arasındaki fazlalığı temsil ettiğini düşünebiliriz. Sezgisel iyilik ölçümü ile alakasız öznitelikler sınıf tahmininde kötü oldukları için elenirken, fazlalık öznitelikler ise bir veya daha fazla öznitelikle yüksek korelasyona sahip oldukları için eleneceklerdir. Şekil 2.12 CFS öznitelik seçimindeki elemanları göstermektedir [29].



Şekil 2.12 CFS öznitelik seçimi

2.2.3 Öznitelik korelasyonları

Makine öğrenmedeki sınıflandırma işlemleri genel olarak nominal sınıf değerlerini birbirinden ayırt edebilmek için öğrenme işlemi ihtiva ederken sıralı veya sürekli öznitelikler üzerinde çalışmaktadırlar. Eşitlik 2.8 için gerekli korelasyonların hesabı için ortak bir kural elde edebilmek için sürekli öznitelikler veri gruplama ile nominal değerlere dönüştürülür [29].

Eşitlik 2.8'deki öznitelik-sınıf korelasyonları ile öznitelik iç korelasyonlarının hesabında birtakım bilgi tabanlı ölçütler denenmiştir. Bunların arasında belirsizlik katsayısı ve simetrik belirsizlik katsayısı (Press et al., [31]), kazanç oranı (Quinlan, [32]) ve minimum tanımlama prensibine dayalı birtakım yöntemler bulunmaktadır. En iyi sonuçlar ise öznitelik-sınıf korelasyonu için kazanç oranı, öznitelik iç korelasyonları içinse simetrik belirsizlik katsayısı kullanıldığında elde edilmiştir [29].

Eğer X ve Y ayrık iki rastgele değişken ise eşitlik 2.9 ve 2.10 X'in incelenmesinden önceki ve sonraki Y'nin entropisini verir. Eşitlik 2.11 X'in incelenmesinden sonra Y

ile ilgili elde edilen bilgi değerini verir. Aynı zamanda X ile ilgili elde edilen bilgi değeri de elde edilebilir [29].

$$H(Y) = \sum_{y} p(y) \log_2\big(p(y)\big) \tag{2.9}$$

$$H(Y|X) = \sum_{x} p(x) \sum_{y} p(y|x) \log_2\big(p(y|x)\big) \tag{2.10}$$

$$\begin{aligned} kazanç &= H(Y) - H(Y|X) \\ &= H(X) - H(X|Y) \\ &= H(Y) + H(X) - H(X,Y) \end{aligned} \tag{2.11}$$

Kazanç daha fazla değere sahip özniteliklerden yana olma yanılgısındadır, bu da aslında bilgi değeri fazla olmasa da çok daha fazla değere sahip özniteliklerin az sayıda değere sahip özniteliklerden daha fazla kazanca sahip olması anlamına gelir. Kazanç oranı eşitlik 2.12 ise bu yanılgının önüne geçmeye çalışan simetrik olmayan bir ölçüttür. Eğer Y tahmin edilecek değişkense kazanç oranı, kazancı X'in entropisine bölerek normalize eder. Simetrik belirsizlik katsayısı ise kazancın normalize edilmesini, kazancı X ve Y'nin entropilerinin toplamına bölerek gerçekleştirir. Kazanç oranı da simetrik belirsizlik katsayısı da 0 ile 1 arası değere sahiptir. 0 değeri her iki ölçütte de X ve Y'nin hiçbir ilişkisinin olmadığı anlamına gelmektedir. Kazanç oranında 1 değeri Y'nin bilgisinin X'in tam olarak tahmin edilmesini sağladığı anlamındadır, simetrik belirsizlik katsayısında ise bir değişkenin bilgisinin diğer değişkeni tam olarak tahmin edebildiğini göstermektedir.Her iki ölçütte az sayıda değere sahip özniteliklerin üzerinde durmaktadır [29].

$$kazanç\ oranı = \frac{kazanç}{H(X)} \tag{2.12}$$

Yapılan deneyler göstermektedir ki CFS agresif sayılabilecek bir filtredir çünkü genel olarak verilen özniteliklerin yarısından çoğunu eler ve hatta çoğunlukla en iyi birkaç özniteliği bırakır. Bu bazı veri setlerinde gelişme sağlarken bazı veri setlerinde ise daha fazla özniteliğin daha iyi sonuç vereceği açıkça görülebilir. Eşitlik

2.8'deki iç korelasyonların etkisini azaltmak performansı arttırabilmektedir. 0.25 oranında bir ölçeklendirme ise genel olarak birçok veri seti ve öğrenme algoritmasında iyi sonuç vermektedir ve birçok araştırmada bu şekilde kullanılmıştır [29].

Bilgi kazancı gibi tek değişkenli filtrelerin en büyük dezavantajı, öznitelikler arası etkileşimleri hiçe saymalarıdır ve bu eksiklik CFS ve benzeri çok değişkenli filtreler tarafından giderilmiştir. CFS bir öznitelik alt kümesinin değerini ölçerken her özniteliğin tek başına tahmin yeteneğini değerlendirmenin yanında özniteliklerin birbirlerine göre gereksizliklerini de hesaba katar. Öznitelik alt kümeleri ile sınıflar arası korelasyonun hesabında ve ayrıca özniteliklerin birbirleriyle olan iç korelasyonlarının hesaplanmasında korelasyon katsayıları kullanılır. Bir öznitelik grubunun geçerliliği sınıflar ile özniteliklerin korelasyonu ile artarken özniteliklerin iç korelasyonu ile azalmaktadır [29].

CFS en iyi öznitelik alt kümesini bulmaya çalışır ve bunun için çeşitli arama stratejilerinden yararlanabilir. Bunlara örnek olarak; ileri doğru seçim, geriye doğru eleme, çift yönlü arama, önce-en iyi arama ve genetik arama verilebilir.

2.2.4 Genetik algoritma

Genetik algoritma genel olarak kullanılan olasılıksal bir arama metodudur ve özellikle öznitelik seçme gibi durumlarda sık rastlanılan geniş uzayda arama alanında çok başarılıdır. Bunun da ötesinde diğer birçok arama algoritmasının aksine lokal değil genel arama gerçekleştirir. Bir genetik algoritma temel olarak üç operatörden oluşur: çoğaltma, çaprazlama, mutasyon. Çoğaltma iyi dizilimleri seçer, çaprazlama iyi dizilimleri birleştirerek daha iyi döl dizilimler elde etmeye çalışır ve mutasyon bir dizilimi değişikliğe uğratarak daha iyi bir dizilim elde etmeye çalışır. Bu operatörler kullanılarak elde edilen her yeni nesil popülasyon algoritmanın sonlanma koşulu için test edilir. Eğer sonlama koşulu sağlanamamışsa mevcut popülasyon tekrar aynı işlemlerden geçirilerek yeniden bir popülasyon üretilir ve sonlanma kriterleri sağlanana kadar bu döngü devam eder [33].

2.2.5 Bilgi kazancı

Bir eğitim veri seti (Ör: S) için entropi bu veri setinin katışıklığının değerlendirilmesinde önemli bir ölçüttür. X bilgisi sağlandığı taktirde Y'nin entropisinin azalma oranı dikkate alınarak Y bilgisi ile ilgili daha fazla bilgi sağlayacak bir ölçüt tanımlanabilmektedir. Bu ölçüte bilgi kazancı adı verilmektedir [34].

$$Bilgi\ Kazancı = H(Y) - H(Y|X) = H(X) - H(X|Y) \tag{2.13}$$

Bilgi kazancı simetrik bir ölçüttür (Eşitlik 2.13). X gözlemlendikten sonra Y ile ilgili kazanılan bilgi Y gözlemlendikten sonra X ile ilgili kazanılan bilgiye eşittir. Bilgi kazancının önemli bir zayıf noktası çok sayıda değere sahip özniteliklere daha fazla bilgi değerleri olmasa bile daha fazla eğilimli olmasıdır [34].

2.2.6 Kazanç oranı

Kazanç oranı simetrik olmayan bir ölçüttür ve bilgi kazancının yukarıda bahsedilmiş yanılgı eğilimini aşmak üzere öne sürülmüştür (Eşitlik 2.14) [34].

$$Kazanç\ Oranı = \frac{Bilgi\ Kazancı}{H(X)} \tag{2.14}$$

Örneğin Y değeri tahmin edilmek istendiğinde bilgi kazancı X'in entropisine bölünerek normalize edilir. Aynı işlem ters yönde de gerçekleştirilebilir. Bu normalizasyon ile kazanç oranı değerleri mutlaka 0-1 aralığına düşmüş olur. Kazanç oranı 1'e eşit olduğunda X bilgisinin Y bilgisini tam olarak tahmin etmeye yettiği, 0 değeri ise Y ve X arasında hiçbir ilişki olmadığı anlamına gelmektedir. Bilgi kazancının aksine kazanç oranı az sayıda değere sahip özniteliklere önem verir [34].

2.2.7 Simetrik belirsizlik katsayısı

Simetrik belirsizlik katsayısı, bilgi kazancının olumsuz yönünü telafi edebilmek üzere onu X ve Y nin entropi toplamlarına böler (Eşitlik 2.15) [34].

$$Simetrik\ Belirsizlik\ Katsayısı = 2\frac{Bilgi\ Kazancı}{H(Y) + H(X)} \quad (2.15)$$

Eşitlik 2.15 içinde geçen düzeltme faktörü 2 sebebiyle simetrik belirsizlik katsayısı da 0-1 arası normalize olmuş değerler alır. 1 değeri bir özniteliğin bilinmesiyle diğer bir özniteliğin tamamıyla tahmin edilebildiğini gösterirken, 0 değeri ise bu iki özniteliğin birbirine bağımlılığının olmadığını göstermektedir. Simetrik belirsizlik katsayısı da kazanç oranı gibi az değere sahip özniteliklerden yana eğilimlidir [34].

2.2.8 Ki-Kare öznitelik seçimi

Chi Square ile öznitelik seçimi sık kullanılan metotlar arasındadır. Bu metot ile bir özniteliğin bilgi değeri onun sınıfa göre chi-square istatistiksel değerinin hesaplanması ile ölçülür. Başlangıç hipotezine göre iki özniteliğin birbirleriyle hiçbir ilişkisi olmadığı varsayılır ve Ki-Kare formülü ile test edilir [34].

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

O_{ij} gözlenen sıklık ve E_{ij} de beklenen yani teorik sıklık değeridir ve başlangıç hipotezi ile ileri sürülmüştür. X^2 değeri ne kadar büyük olursa başlangıç hipotezine karşı o denli büyük bir iddia söz konusudur [34].

2.2.9 One-R

One-R Holte tarafından önerilmiş basit bir algoritmadır. Eğitim veri setindeki her veri için bir kural oluşturur ve en az hata veren kuralı seçer. Tüm nümerik değerli özniteliklerin sürekli değerler olduğunu varsayar ve düz mantık bir metot ile değerleri birçok ayrık aralığa böler [34].

2.3 Veri Setleri

2.3.1 Çok kategorili kanser sınıfları veri setleri

Çalışmamızda aynı deneklere ait hem mRNA hem de mikroRNA bilgilerinin beraber bulunduğu ve hem normal hem de hastalıklı doku örneklerinden oluşan bilindik bir veri seti kullanılmıştır. Bu veri setindeki hastalıklı doku örneklerine ait kanser tipleri

şu şekildedir: kolon, pankreas, böbrek, mesane, prostat, rahim, yumurtalık, akciğer, akciğer zarı, cilt ve göğüs. İlgili kaynaklardan elde edilen mikroRNA tekil, mRNA tekil ve mikroRNA-mRNA birlikte olacak şekilde üç veri seti kullanılmıştır.

2.3.1.1 mRNA ifade biçimleri veri seti

Bu veri seti Ramaswamy et al. [8] yayınında sunulan GCM (Global Cancer Map) mRNA veri setinin büyük bir kısmını içeren alt kümesidir. Bu sette 16,063 genle ifade edilmiş ve 11 farklı kanser sınıfı ile aynı bölgelere ait normal doku örneklerinin de içinde olduğu toplam 89 doku örneği bulunmaktadır. Deneylerimizde ilgili normal doku örnekleri tek bir normal sınıf altında gruplanmıştır.

2.3.1.2 mikroRNA ifade biçimleri veri seti

Lu et al. [17] yayınında Ramaswamy et al. [8] yayınındakiaynı 217 memeli örnekleri üzerinde sistematik bir ifade biçimi analizi yapılmak üzere boncuk tabanlı akım sitometrik çalışması ile mikroRNA ifade biçimleri elde edilmiştir. Bizim tarafımızdan ise bu veri setinin mRNA veri setimizdeki aynı 89 örneğe karşılık gelecek ve toplam 217 mikroRNA'dan oluşacak şekilde bir alt kümesi kullanılmıştır.

2.3.1.3 mikroRNA ve mRNA ifade biçimleri veri seti

Bu veri seti kullandığımız iki veri setinin (mikroRNA ve mRNA veri setlerimiz) bileşkesi olup toplamda 16,280 öznitelik içermektedir.

2.3.2 Göğüs kanseri alt kategorileri veri setleri

2.3.2.1 mRNA ifade biçimleri veri seti

Bu veri seti Enerly et al. [35] yayınında sunulan, **'GSE19536'** kodlu ve **'Molecular Characterization of Breast Cancer Subtypes Derived from Joint Analysis of High Throughput miRNA and mRNA Data**' başlıklı GEO (Gene Expression Omnibus) mRNA veri setinin büyük bir kısmını içeren alt kümesidir. Bu sette 40493 genle ifade edilmiş ve 4 farklı göğüs kanseri alt sınıfı ile göğüs bölgesine ait normal doku örneklerinin de içinde olduğu toplam 94 doku örneği bulunmaktadır.

2.3.2.2 mikroRNA ifade biçimleri veri seti

Bu veri seti Enerly et al. [35] yayınında sunulan, **'GSE19536'** kodlu ve **'Molecular Characterization of Breast Cancer Subtypes Derived from Joint Analysis of High Throughput miRNA and mRNA Data**' başlıklı GEO mikroRNA veri setinin büyük bir kısmını içeren alt kümesidir. Bu sette 490 mikroRNA ile ifade edilmiş ve 4 farklı göğüs kanseri alt sınıfı ile göğüs bölgesine ait normal doku örneklerinin de içinde olduğu toplam 94 doku örneği bulunmaktadır. Bu örnekler mRNA veri setinde verilen 94 örnekle aynı deneklerden elde edilmiştir.

2.3.2.3 mikroRNA ve mRNA ifade biçimleri veri seti

Bu veri seti kullandığımız iki veri setinin (mikroRNA ve mRNA veri setlerimiz) bileşkesi olup aynı 94 örnek için toplamda 40982 öznitelik içermektedir.

3 ÇOK KATEGORİLİ KANSER SINIFLANDIRMASINDA mikroRNA VE mRNA BİLGİSİNİN BİRLİKTE KULLANIMI

3.1 Çalışma Kapsamı ve Geçmiş Çalışmalar

Bu çalışma kapsamında ilgili beş sınıflandırıcı ile oluşturulan üç veri seti üzerinde, hem orijinal veriler hem de seçilen beş öznitelik azaltma metodunun çıktıları olan azaltılmış veriler kullanılarak toplam doksan deney geçekleştirilmiştir. Bu deneylerde ilgili sınıflandırıcıların seçilen veri setleri üzerindeki performanslarının test edilebilmesi için LOOCV (Leave-one-out cross validation) tekniği kullanılmıştır. Bu doğrulama tekniğinin küçük *n* büyük *p* problemlerinde en gerçekçi test sonuçlarını sağladığı çok iyi bilinmektedir. Bu tip deneyler eğitim verisi üzerinde yeterli doğrulama yapılmaz ise **'overfitting'** problemi ile karşılaşmaya aşırı meyillidir. Deneylerin doğruluk oranlarının belirtilmesinde doğru sınıflandırılan örnek yüzdesi kullanılmıştır.

Peng et al. [11] yayınında da aynı veri setleri ve benzer deneysel kurulum kullanılarak bu çalışmadakine yakın bir kanser sınıflandırılması çalışması yapılmıştır. O çalışmada da LOOCV kullanılmış ve geçmiş çalışmalar ile gerçekçi ve tekrarlanabilir karşılaştırmalar sunulmuştur. O çalışma sonucundayazarlar Lu et al. [17] yayınında savunulanın tam aksine mikroRNA bilgisinin tek başına kanser sınıflandırma yeteneğinin mRNA bilgisine göre çok yetersiz olduğunu savunmuşlardır. Lu et al. [17] mRNA verilerinin düzgün sınıflar oluşturacak şekilde kümelenmelerinin mikroRNA verilerinin aynı sınıflar için olan kümelenmelerine göre çok düzensiz olması nedeniyle mRNA bilgisinin mikroRNA bilgisine göre yetersiz bir sınıflandırıcı olacağını savunurken, Peng et al. [11] ise doğru bir öznitelik seçme veya azaltma uygulaması ile mRNA bilgisinin mikroRNA bilgisinden daha iyi sınıflandırma performansına sahip olabildiğini başarılı bir şekilde göstermiştir.

Geçmiş çalışmaları detaylı bir biçimde incelediğimizde ve disiplinler arası bilgi birikimimiz ışığında bizim düşüncemiz ise ne Lu et al. [17] yayınında ne de Peng et al. [11] yayınında savunulan bilgilerin yanlış olmadığı fakat yetersiz olduklarıdır. Çünkü hem mikroRNA hem de mRNA tümör sınıflandırmada önemli değere sahiptir. Ayrıca birbirleriyle ilişkili oldukları da bilimsel olarak kanıtlanmıştır. Bu bilgiye dayanarak, bu iki veri kaynağının etkin bir şekilde füzyonu, gelişmiş makine

öğrenme algoritmaları ve optimize edilmiş öznitelik seçimi metotlarının birlikte kullanımıyla çok kategorili kanser sınıflandırma problemlerinde daha iyi tahmin sonuçları elde edebileceğimizi öne sürdük.

Çoğu makine öğrenme algoritması işleyişini değiştiren parametre ayarlarına sahiptir ve bu parametrelerden büyük ölçüde etkilenir. Bu kapsamda kullanılan algoritmalar için mümkün olduğunca iyi parametre ayarlarını bulmak üzere bazı kullanıcı destekli ve aç gözlü optimizasyon yöntemlerinden faydalandık. Bu parametre seçimi işlemini KNN, ANN ve SVM algoritmaları için gerçekleştirirken C4.5 ve NBM için ise WEKA ortamındaki ön tanımlı ayarları kullandık.

MikroRNA-mRNA veri entegrasyonunun başarısını ölçmek için skor tabanlı bir sistem geliştirdik. Bu sistemde her sınıflandırıcı ve öznitelik seçimi metodu çifti için (ayrıca hiçbir öznitelik seçimi olmaksızın her sınıflandırıcı için) her üç veri seti üzerinde elde edilen en iyi deney sonuçlarını kayıt altına aldık. Her deney sonrası kazanan veri setine ilişkin skoru arttırarak sonuçta maksimum değeri 30 olabilecek genel bir skor elde ettik. Bu kapsamlı test tasarımının tamamı gerçekleştikten sonra beklenen, bu iki biyolojik veri kaynağının birlikte kullanımının genel anlamda daha iyi kanser sınıflandırması yapabileceği sonucuydu.

3.2 Bulgular ve Tartışma

Bu çalışmada elde ettiğimiz deney sonuçları Çizelge 3.1'de gösterilmiştir. İlk adımda mikroRNA, mRNA ve bileşke veri setinin herhangi bir öznitelik seçimi olmadan ilgili beş algoritmada kullanılmaları sonucunda elde edilen sınıflandırma performansları değerlendirilmiştir. Bu adımda elde edilen sonuçlar Lu et al. [17] yayınında da bahsedildiği şekilde mikroRNA verisinin tek başına mRNA verisinden daha ayırt edici nitelikte olduğunu göstermiştir. mRNA verisinde bir öznitelik seçimi uygulanmadan alınan bu sonuç (Peng et al. [11] yayınında mikroRNA verisine göre çok büyük boyutlardaki mRNA verisinin mutlaka öznitelik azaltma veya seçme gibi işlemlerden geçmesi gerektiğini kanıtlamıştır) zaten bizim tarafımızdan da bekleniyor olmasına karşın yine de füzyon veri seti ile bazı algoritmalar (C4.5 karar ağaçları ile SVM) daha iyi sonuç vermeyi başarmıştır.

Çizelge 3.1. Gerçekleştirilen çok kategorili kanser sınıflandırma deneylerinde elde edilen LOOCV sonuçları.

Öznitelik Seçimi Metodu	Veri seti (ve azaltılan öznitelik sayısı)	LOOCV sonucu (%)					Kazanan Veri Seti Skorları		
		KNN	ANN	DT	NBM	SVM	mRNA	miRNA	miRNA &mRNA
Yok	mRNA (16063 öznitelik)	60,7	23,6	38,2	55,1	**75,3**	**0/5**	**3/5**	**2/5**
	miRNA (217 öznitelik)	68,5	**83,1**	51,7	75,3	77,5			
	miRNA & mRNA (16280)	60,7	23,6	52,8	57,3	**77,5**			
SVM-tabanlı	mRNA (100)	88,8	**95,8**	41,6	85,4	92,1	**0/5**	**0/5**	**5/5**
	miRNA (100)	73,0	**86,5**	46,1	75,3	82,0			
	miRNA & mRNA (100)	92,1	**96,6**	70,8	91,0	93,3			
Bilgi Kazancı	mRNA (365)	80,9	**89,9**	53,9	75,3	84,3	**1/5**	**0/5**	**4/5**
	miRNA (76)	73,0	**83,1**	40,4	70,8	82,0			
	miRNA & mRNA (441)	85,4	88,8	67,4	87,6	**88,8**			
Kazanç Oranı	mRNA (<365)	80,9	**89,9**	55,1	75,3	84,3	**0/5**	**0/5**	**5/5**
	miRNA (<76)	76,4	**85,4**	40,4	70,8	84,3			
	miRNA & mRNA (<441)	87,6	**92,1**	67,4	87,6	88,8			
Ki-Kare	mRNA (<365)	80,9	**89,9**	56,2	77,5	84,3	**0/5**	**0/5**	**5/5**
	miRNA (76)	73,0	**83,1**	40,4	70,8	82,0			
	miRNA & mRNA (<=441)	85,4	**89,9**	69,7	87,6	88,8			
CFS	mRNA (90)	88,8	**93,3**	55,1	84,3	91,0	**0/5**	**0/5**	**5/5**
	miRNA (18)	68,5	74,2	55,1	46,1	71,9			
	miRNA & mRNA (91)	88,8	93,3	67,4	89,9	**93,3**			
GENEL SKOR							**1/30**	**3/30**	**26/30**

Bu çizelgede gösterilen renkler yukarıda anlatılan skor belirleme işleminde her deney grubunda kazanan veri setini belirlemek için kullanılmıştır. Veri setlerine karşılık gelen renkler kazanan veri seti skorları başlığı altındaki en sağdaki 3 sütunda verilmiştir. Sonraki adımlarda ise aynı deneyler beş öznitelik seçme metodu, ilgili veri setlerine uygulanarak ve beş sınıflandırıcı ile kullanılarak gerçekleştirilmiştir. Bu deneylerin sonuçları Peng et al. [11] yayınını destekleyecek şekilde mRNA verisinin öznitelik seçme işlemi ile birlikte kullanıldığında mikroRNA verisine göre önemli derecede üstünlük sağladığını göstermektedir. Buna karşın füzyon veri seti yapılan 30 deneyin 26'sında en iyi sonucu vererek (kalan dört deneyin 3 tanesi zaten hiçbir öznitelik seçimi olmadan yapılanlardır) bu iki veri kaynağının bir arada kullanımının yalnız kullanımlarından daha iyi performans sağladığı gösterilmiştir. Ayrıca öznitelik seçimi metotlarının bileşik veri seti üzerindeki sonuçları da incelendiğinde, her zaman her iki veri setinden de özniteliklerin seçildiği görülmüştür.

Deneylerde elde ettiğimiz en iyi sınıflandırma sonucu 96.6% LOOCV değeri ile ANN sınıflandırıcısı-SVM Öznitelik seçimi-füzyon veri seti birlikte kullanımına aittir. Bu sonuç, aynı veri seti üzerinde elde edilmiş literatürdeki en yüksek sonuç olan elde ve Peng et al. [11] yayınında belirtilen 95.8% LOOCV değerini geride bırakmıştır. Çizelge 3.2'de GCM veri setleri üzerinde diğer yayınlarda da belirtilmiş çok kategorili kanser sınıflandırma LOOCV sonuçları gösterilmiştir. Yaptığımız bu değerlendirmelerin yanı sıra kullandığımız sınıflandırıcılarında bu deneylerdeki performanslarını karşılaştırmış olduk. Sonuçta en iyi sınıflandırıcıların başta ANN ve onu az farkla takip eden SVM olduğu görüldü. Fakat SVM sınıflandırıcısının ANN'e göre çok daha fazla optimize edilebilme imkanı bulunduğu görülmüş ve bu yüzden ileriki çalışmalarımızda daha yoğun optimizasyonlar sonucu daha iyi performans kaydedebileceği düşünülmüştür.

Çizelge 3.2 Diğer yayınlardaki sonuçlar ile karşılaştırma (GCM veri setleri üzerinde çok kategorili kanser sınıflandırma LOOCV sonuçları)

Yayınlar	**Doğruluk (%)**
Ramaswamy et al. [8]	78.0
Su et al.[9]	81.3
Peng et al. [12]	85.2
Lin et al. [13]	84.3
Liu and Xu [15]	91.8
Peng et al. [11]	95.8
Bu çalışma	**96.6**

4 GÖĞÜS KANSERİ ALT KATEGORİ SINIFLANDIRMASINDA mikroRNA VE mRNA BİLGİSİNİN BİRLİKTE KULLANIMI

4.1 Çalışma Kapsamı ve Geçmiş Çalışmalar

Bir önceki çalışmamızda mikroRNA ve mRNAnın birlikte kullanımının ve etkili öznitelik azaltma metotlarının uygulamasının çok kategorili kanser sınıflandırmasında umut verici gelişmeler sağladığının belirlenmesinden sonra, aynı yöntemler ile bu kategorilerden birinin alt kategorilerine sınıflandırılması üzerinde çalışılmıştır. Buradaki fark tüm hasta örneklerin göğüs kanseri olması, fakat alt-kanser türlerine göre farklılık göstermesidir. Bu nedenle, kanser sınıflandırma yerine alt-kanser sınıflandırma diyebileceğimiz bu deneyde 5 farklı göğüs kanseri alt-türü sınıf olarak belirlenmiştir: Lum-A, Lum-B, normal-like, Basal ve ERBB2.

Önceki çalışmalarımızda elde ettiğimiz bulgular doğrultusunda (çok kategorili sınıflandırma ile alt kategori sınıflandırmasının benzer sonuçlar verebileceği düşüncesi kapsamında) beklentilerimiz;

1. mRNA bilgisinin başarılı bir öznitelik azaltma işlemi ile kullanıldığında mikroRNA bilgisinden daha iyi sınıflandırma sonuçları vermesi,
2. mikroRNA bilgisinin ve mRNA bilgisinin birlikte kullanımının (birlikte bir öznitelik azaltma işlemine tabi tutulduklarında) mRNA ile elde edilen iyileştirilmiş (öznitelik azaltma ile) sınıflandırma sonuçlarını daha da iyileştirebilmesi,
3. mikroRNA bilgisinin ve mRNA bilgisinin birlikte tabi tutuldukları öznitelik azaltma işlemlerinde ilgili metot tarafından seçilen öznitelik kümesinin her iki kaynaktan da öznitelikleri içermesi,
4. En iyi ve tutarlı sınıflandırma sonuçlarını (farklı öznitelik seçme metotları ile) ANN ve SVM sınıflandırıcılarının vermesi. ANN'in SVM'e göre çok az farkla daha iyi olabileceği gibi SVM'in bir önceki çalışmamızda gerçekleştirdiğimiz optimizasyonları genişletebileceğimiz daha fazla parametresinin olması dolayısıyla ANN'in üzerinde sonuçlar verebileceği,
5. Bu veriler üzerinde en iyi sonuç veren öznitelik azaltma metotlarının sırasıyla SVM öznitelik seçimi, CFS öznitelik seçimi ve Ki-Kare öznitelik seçimi olması,
6. SVM öznitelik azaltma yöntemi ile ANN (veya SVM) sınıflandırıcısının birlikte kullanımı ve uyguladığımız parametre optimizasyonları sonucu literatürde

ilgili veri seti için belirtilmiş LOOCV sonuçlarının üzerinde bir başarı elde etmek,

şeklinde sıralanabilir.

4.2 Bulgular ve Tartışma

Yaptığımız deneyler sonucunda önceki bulgularımız doğrultusundaki beklentilerimiz çok büyük ölçüde gerçekleşmiştir. Çizelge 4.1'de verilen sonuçlara göre yine veri bütünleştirmenin sınıflandırma işleminde belirgin bir faydası görünmektedir. Bu çizelge bölüm 3.1 ve 3.2'de anlatılan skor tabanlı sistem kullanılarak oluşturulmuştur. Buna karşın en iyi sonuç sadece mRNAların kullanımıyla, hem öznitelik azaltma hem de sınıflandırma işleminin SVM ile yapılması durumunda elde edilebilmiştir.

Çizelge 4.1 Gerçekleştirilen göğüs kanseri alt-türü sınıflandırma deneylerinde elde edilen LOOCV sonuçları

Öznitelik Seçimi Metodu	Veri seti (ve azaltılan öznitelik sayısı)	LOOCV sonucu (%)				Kazanan Veri Seti Skorları		
		KNN	ANN	DT	SVM	miRNA	mRNA	miRNA&mRNA
Yok	mRNA (40493)	57,4	70,2	66	72,3	**1/4**	**1/4**	**2/4**
	miRNA (490)	50	76,6	51,1	71,3			
	miRNA&mRNA (40982)	59,6	68,9	64,9	74,5			
SVM-tabanlı	mRNA (100)	86,2	98,9	68,1	100	**0/4**	**2/4**	**2/4**
	miRNA (100)	63,8	88,3	67	88,3			
	miRNA & mRNA (100)	84	98,9	69,1	97,9			
Ki-Kare	mRNA (100)	69,1	75,5	69,1	75,5	**2/4**	**1/4**	**1/4**
	miRNA (81)	64,9	76,6	67	76,6			
	miRNA & mRNA (100)	69,1	73,4	66	75,5			
CFS	mRNA (154)	81,9	85,1	73,4	86,2	**0/4**	**3/4**	**1/4**
	miRNA (30)	63,8	72,3	55,3	74,5			
	miRNA & mRNA (162)	79,8	81,9	68,1	89,4			
GENEL SKOR						**3/16**	**7/16**	**6/16**

Bu çalışma kapsamında beklentilerimiz ile ilgili belirttiğimiz maddelere karşılık elde ettiğimiz sonuçlar aşağıda verilmiştir:

1. mRNA bilgisi seçtiğimiz öznitelik azaltma metotları ile kullanıldığında mikroRNA bilgisinden (mikroRNA için öznitelik seçimi olsun ya da olmasın) daha iyi sınıflandırma sonuçları vermiştir. Önceki çalışmada da bu çok net anlaşılmış olsa da bu çalışma da öznitelik azaltma işlemlerinin bu alandaki önemi ve mRNA bilgisinin sınıflandırma değerini ortaya koyma da önemli rol oynamıştır.
2. mikroRNA bilgisinin ve mRNA bilgisinin birlikte kullanımı önceki çalışmadan farklı olarak iyileştirilmiş mRNA (öznitelik azaltma ile) sınıflandırma sonuçlarından daha iyi sonuç verememiştir. Skorlara bakıldığında her ikisi de yakın sonuçlar verse de (7'ye 6) sadece mRNA kullanımı ile en iyi sonuca ulaşılabilmiştir. Bu durum aslında beklediğimize ters sayılmamalıdır. Çünkü farklı deneylerde iki verinin birlikte kullanımı sonuçları biraz daha iyileştirmiştir. Fakat bu sınıflandırma probleminde mRNA bilgisi çok yüksek sınıflandırma başarısı için yeterli olmuş mikroRNA bilgisinin belirttiğimiz metotlarla bunu daha da iyileştirmesi mümkün olmamıştır. Hem mRNA nın tek başına kullanıldığı hem de füzyon veri setinin kullanıldığı iki durumda 98.9% gibi çok yüksek bir sonuç elde edilmiştir. Buna karşın sadece mRNA kullanılan bir durumda 100% lük mükemmel sonuca ulaşılmıştır. Sadece mRNA kullanılan durumların füzyon veri setinin kullanıldığı durumlara göre gerek sayıca gerekse performans bakımından önemli bir fark yaratmamasından ötürü füzyon işleminin değerine aykırı bir durum olmamakla birlikte bazı sınıflandırmalarda çok önemli olmayabileceği söylenebilir.
3. mikroRNA bilgisinin ve mRNA bilgisinin birlikte tabi tutuldukları öznitelik azaltma işlemlerinde ilgili metot tarafından seçilen öznitelik kümesinin her iki kaynaktan da öznitelikleri içerdiği görülmüştür. Zaten bu durum bile tek başına bu iki verinin birbiriyle ilişkisi ve birlikte kullanımlarının önemini tekrar göstermiştir. Her iki bilgi kaynağından da veriler içeren füzyon veri setleri 98.9% ve 97.9% gibi çok önemli sonuçlar vererek mikroRNA bilgisinin mRNA bilgisinin tek başına çok belirleyici olduğu durumlarda bile destekleyici rol oynayabileceğini göstermiştir.

4. En iyi ve tutarlı sınıflandırma sonuçları (farklı öznitelik seçme metotları ile) yine ANN ve SVM sınıflandırıcıları ile elde edilmiştir. Fakat daha önce de belirttiğimiz SVM'in optimizasyonlarını genişletebileceğimiz daha fazla parametresinin olmasından dolayı ve bu çalışmamızda bu optimizasyon işlemlerini geliştirdiğimizden ötürü ANN'e göre çoğunlukla daha iyi sonuçlar elde edilmiştir.
5. Bu veriler üzerinde en iyi sonuç veren öznitelik azaltma metotları yine sırasıyla SVM öznitelik seçimi, CFS öznitelik seçimi ve Ki-Kare öznitelik seçimi olarak belirlenmiştir.
6. SVM öznitelik azaltma yöntemi ile hem ANN hem de SVM sınıflandırıcısının birlikte kullanımı ve uyguladığımız parametre optimizasyonları sonucu gerçektende çok önemli LOOCV sonuçları elde edilmekle beraber, 100% lük mükemmel LOOCV sınıflandırma sonucu da elde edilerek uyguladığımız metodolojinin doğruluğu kanıtlanmıştır.

5 SONUÇ

Bu çalışma ile mikroRNA ve mRNA bilgisinin bütünleştirerek kullanımının gerek çok kategorili kanser sınıflandırmada gerekse kanser alt kategori sınıflandırmasında ne ölçüde yararlı olacağı incelenmiştir. Bölüm 3 ve bölüm 4 bulgular kısımlarında sonuçlarına daha detaylı yer verilmiş olan, yaptığımız geniş kapsamlı deneylerden elde ettiğimiz sonuçlara göre doğru öznitelik azaltma stratejisi ve mikroRNA ile mRNA bilgisinin öznitelik seviyesinde füzyonu ile bütünleştirilmesinin sınıflandırma sonuçlarını önemli ölçüde geliştirebildiği anlaşılmıştır.

Gerçekleştirdiğimiz deney sonuçları incelendiğinde veri füzyonunun deney sonuçlarının çok büyük bir kısmında daha iyi sınıflandırma yapılabilmesini sağladığı görülmüştür. Tüm sonuçlar incelendiğinde ise sınıflandırıcı olarak ANN ya da SVM kullanımının, öznitelik seçimi olarak ise SVM tabanlı öznitelik azaltma yönteminin tercih edilmesinin en başarılı sonuçlara ulaşmayı sağladığı görülebilir. Sınıflandırıcı olarak ANN ile SVM arasında ise SVM' in çok daha hızlı ve başarılı optimizasyonunun mümkün olduğu görülmüş ve özellikle çalışmanın ikinci safhası olan kanser alt kategori sınıflandırma deneylerinde bu konunun daha çok üzerinde durulması ile kusursuz sınıflandırma sonucuna ulaşılabilmiştir.

SVM ile öznitelik seçimi ise her koşulda diğer öznitelik seçimi yöntemlerinden daha iyi sonuçlar elde edilmesini sağlamıştır. Bunun sebepleri incelendiğinde diğer öznitelik seçimi yöntemlerinin ya problemden bağımsız olarak çok sayıda değere sahip öznitelikleri gereğinden çok veya gereğinden az önemseme yanılgısına düştükleri ya da probleme yani eğitim verisine fazla bağlı kalarak fazla katı seçimler yaptıkları gözlenmiştir. Buna karşın SVM'in snıflandırma işleminde **'overfitting'** problemi ile baş etmesini sağlayan yapısı ile aynı sebepten ötürü öznitelik seçimi için kullanıldığında da hem problem için en uygun seçimleri yaparken hem de eğitim verisinde bulunmayan durumları en uygun şekilde içerecek seçimleri yapma eğilimi gösterdiği anlaşılmıştır.

Bölüm 3'te Çizelge 3.1 ile ve bölüm 4'te Çizelge 4.1 ile verilen gerçekleştirdiğimiz tüm deneylere ait test sonuçlarına ek olarak, öznitelik seçiminin bu kapsamdaki önemini göstermek ve ilgili veri füzyonunun sınıflandırma sonuçlarına olan etkisinin

olabildiğince farklı koşulda incelenebilmesi amacıyla hem çok kategorili kanser sınıflandırma hem de kanser alt kategori sınıflandırma için birbirinden farklı eğilimler gösteren sınıflandırıcılar ile hiçbir öznitelik seçimi kullanmadığımız ve en iyi öznitelik azaltma sonuçlarını aldığımız SVM tabanlı öznitelik seçimi yöntemini kullandığımız deneylere ait detaylı test sonuçları eklerde verilmiştir.

Elde ettiğimiz sonuçlar sayesinde çok kategorili sınıflandırma ve alt kategori sınıflandırması arasındaki benzerlikler ve farkların büyük ölçüde anlaşılabileceği verilere ulaşılmış ve hem mikroRNA hem de mRNA bilgilerinin bu konulardaki önemleri detaylıca incelenebilmiştir. Her iki çalışmada da çok yüksek başarıdasonuçlara ulaşılmış olmasına karşın bu sınıflandırma problemlerine daha da iyi bir çözüm bulunabilirliği de göz ardı edilmeyip bu konuda da incelemeler yapılmıştır.

Kanser sınıflandırma konusunda böyle bir çözümün yeni bir model tabanlı (mevcut öznitelik seçme ve sınıflandırma metotlarının, elde ettiğimiz sonuçlar doğrultusunda özelleştirilmesi ve birlikte kullanımları ile) sınıflandırıcı tasarımı ile mümkün olabileceği elde ettiğimiz sonuçlar doğrultusunda mümkün gözükmektedir.

Bu iki çalışmada da füzyon işleminde hem mikroRNA hem de mRNA özniteliklerinin eşit değerde ele alındığı bir yöntem uygulanmış ve öznitelik seçme metotları da buna göre bir seçim gerçekleştirmiştir. Tasarlanabilecek yeni modelde ise mikroRNA ve mRNA'nın birlikte kullanımının farklı şekillerde gerçekleştirilebilmesi için (ağırlık tabanlı bir sistem geliştirilmesi durumunda farklı mikroRNA ve mRNA öznitelikleri için değişik ağırlıkların belirlenebilmesi gibi) gerekli olacak bilgilere yaptığımız bu iki detaylı çalışma sayesinde elde edilen verilerle ulaşılabilir

KAYNAKLAR LİSTESİ

[1] MANDAL, A., Cancer Research, http://www.news-medical.net/health/Cancer-Research.aspx, 2013.

[2] MANDAL, A., Cancer Diagnosis, http://www.news-medical.net/health/Cancer-Diagnosis.aspx, 2013.

[3] http://en.wikipedia.org/wiki/Genetic_analysis

[4] SAEYS, Y., INZA, I. and LARRAÑAGA, P., A review of feature selection techniques in bioinformatics, Bioinformatics, vol.23, s.2507-2517, 2007.

[5] OGUL, H. and AKKAYA, M.S., Data integration in functional analysis of microRNAs, Current Bioinformatics, vol.6, s.462-472, 2011.

[6] WEST M., Bayesian factor regression models in the large p, small n paradigm, Bayesian Statistics, vol.7, s.723-732, 2003.

[7] KLAMI, A. and KASKI, S., Probabilistic approach to detecting dependencies between data sets, Neurocomputing, vol.72, s.39-46, 2008.

[8] RAMASWAMY, S., TAMAYO, P., RIFKIN, R., MUKHERJEE, S., YEANG, C.H., ANGELO, M., LADD, C., REICH, M., LATULIPPE, E., MESIROV, J.P., POGGIO, T., GERALD, W., LODA, M., LANDER, E.S. and GOLUB, T.R., Multiclass cancer diagnosis using tumor gene expression signatures, Proc. Natl. Acad. Sci., vol.98, s.15149-15154, 2001.

[9] SU, Y., MURALİ, T.M., PAVLOVİC, V., SCHAFFER, M. and KASİF, S., RankGene: Identification of diagnostic genes based on expression data, Bioinformatics, vol.19, s.1578-1579, 2003.

[10] SU, A.I., WELSH, J.B., SAPINOSO, L.M., KERN, S.G., DIMITROV, P., LAPP, H., SCHULTZ, P.G., POWELL, S.M., MOSKALUK, C.A., FRIERSON, H.F. and HAMPTON, G.M., Molecular classification of human carcinomas by use of gene expression signatures, Cancer Res., vol.61, s.7388-7393, 2001.

[11] PENG, S., ZENG, X., Li. X., PENG, X. and CHEN, L., Multi-class cancer classification through gene expression profiles: microRNA versus mRNA,J. Genet. Genomics, vol.36, s.409-416, 2009.

[12] PENG, S., XU, Q., LING, X.B., PENG, X., DU, W., and CHEN, L., Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines, FEBS Lett., vol.555, s.358-362, 2003.

[13] LIN, T.C., LIU, R.S., CHEN, C.Y., CHAO, Y.T. and CHEN, S.Y., Pattern classification in DNA microarray data of multiple tumor types, Pattern Recognit., vol.39, s.2426-2438, 2006.

[14] XU, R., ANAGNOSTOPOULOS, G.C., and WUNSCH, D.C., Multiclass cancer classification using semisupervised ellipsoid ARTMAP and particle swarm optimization with gene expression data, IEEE/ACM Trans. Comput. Biol. Bioinform., vol.4, s.65-77, 2007.

[15] LIU, K.H., and XU, C.G., A genetic programming-based approach to the classification of multiclass microarray datasets, Bioinformatics, vol.25, s.331-337, 2009.

[16] BARTEL, D.P., MicroRNAs: genomics, biogenesis, mechanism, and function, Cell, vol.116, s.281-297, 2004.

[17] LU, J., GETZ, G., MISKA, E.A., ALVAREZ-SAAVEDRA, E., LAMB, J., PECK, D., SWEET-CORDERO, A.,EBERT, B.L., Mak, R.H., FERRANDO, A.A., DOWNING, J.R.,JACKS, T., HORVITZ, H.R. and GOLUB, T.R., MicroRNA expression profiles classify human cancers, Nature, vol.435, s.83-838, 2005.

[18] XUA, R., XUB, J. and WUNSCH, D.C., MicroRNA expression profile based cancer classification using Default ARTMAP, Neural Networks, vol.22, s.774-780, 2009.

[19] CHAN, E., PATEL, R., NALLUR, S., RATNER, E., BACCHIOCCHI, A., HOYT, K., SZPAKOWSKI, S., GODSHALK, S., ARIYAN, S., SZNOL, M., HALABAN, R., KRAUTHAMMER, M., TUCK, D., SLACK, F.J. and WEIDHAAS, J.B., MicroRNA signatures differentiate melanoma subtypes, Cell Cycle, vol.10, s.1845-1852, 2011.

[20] CARUANA, R. and NICULESCU-MIZIL, A., An Empirical Comparison of Supervised Learning Algorithms, Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, 2006.

[21] BISHOP C.M., Pattern Recognition and Machine Learning. Springer-Verlag New York, NJ, USA, 2006.

[22] DUNHAM, M.H., Data mining introductory and advanced topics, Prentice Hall, 2003.

[23] ALPAYDIN, E., Introduction to Machine Learning, The MIT Press, 2009.

[24] DUDA, O.D., HART, P.E. and STORK, D.G., Pattern Classification, Wiley, 2000.

[25] CORINNA, C. and VLADIMIR, V.N., Support-Vector Networks, Machine Learning, vol.20, no.3, s.273-297, 1995

[26] GUYON, I., WESTON, J., BARNHILL, S., and VAPNIK, V., Gene selection for cancer classification using support vector machines, Machine Learning, vol.46, s.389-422, 2002.

[27] CAI, Z., GOEBEL, R., SALAVATIPOUR, M.R., and LIN, G., Selecting dissimilar genes for multi-class classification, an application in cancer subtyping. BMC Bioinformatics, vol.8, s.206, 2007.

[28] HALL, M.A., Correlation-based feature subset selection for machine learning, PhD Thesis, Hamilton, New Zealand, 1998.

[29] HALL, M.A. and SMITH, L.A., Practical feature subset selection for machine learning, Proceedings of the 21st Australasian Computer Science Conference ACSC'98, Perth, 4-6 February 1998, Berlin: Springer, s.181-191, 1998.

[30] GHISELLI, E. E., Theory of Psychological Measurement, McGraw-Hill, 1964.

[31] PRESS, W. H., FLANNERY, B. P., TEUKOLSKY, S. A. and VETTERLING, W. T., Numerical Recipes in C, Cambridge University Press, 1988.

[32] QUINLAN, J. R., Induction of decision trees, Machine Learning, vol.1, s.81–106, 1986.

[33] MITCHELL, T.M., Machine Learning, McGraw-Hill.

[34] NOVAKOVIĆ, J., STRBAC, P. and BULATOVIĆ, D., Toward optimal feature selection using ranking methods and classification algorithms, Yugoslav Journal of Operations Research 21, vol.1, s.119-135, 2011.

[35] ENERLY, E., STEINFELD, I., KLEIVI, K., LEIVONEN, S. K., AURE, M.R., RUSSNES, H.G., RØNNEBERG, J. A., JOHNSEN, H., NAVON, R., RØDLAND, E., MÄKELÄ, R., NAUME, B., PERÄLÄ, M., KALLIONIEMI, O., KRISTENSEN, V. N. and YAKHINI, Z., miRNA-mRNA integrated analysis reveals roles for miRNAs in primary breast tumors, PLoS One, vol.6, no.2, 2011.

EKLER LİSTESİ

Sayfa

Sayfa

Sayfa

Sayfa

Ek 1 Çok kategorili kanser sınıflandırması sonuçları

Ek 1.1 mikroRNA bilgisi ile çok kategorili kanser sınıflandırması sonuçları

Ek 1.1.1 Öznitelik seçimi olmadan mikroRNA bilgisi ve ANN sınıflandırıcı ile çok kategorili kanser sınıflandırması sonuçları

```
Scheme:       weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M
0.2 -N 500 -V 0 -S 0 -E 20 -H a
Relation:     miRNA_mRNA_tek_normal
Instances:    89
Attributes:   218
Test mode:    89-fold cross-validation

=== Classifier model (full training set) ===

Class NORMAL
    Input
    Node 0
Class T_COLON
    Input
    Node 1
Class T_PAN
    Input
    Node 2
Class T_KID
    Input
    Node 3
Class T_BLDR
    Input
    Node 4
Class T_PROST
    Input
    Node 5
Class T_OVARY
    Input
    Node 6
Class T_UT
    Input
    Node 7
Class T_LUNG
    Input
    Node 8
Class T_MESO
    Input
    Node 9
Class T_MELA
    Input
    Node 10
Class T_BRST
    Input
    Node 11
```

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          74               83.1461 %
Kappa statistic                          0.809
Mean absolute error                      0.0452
Root mean squared error                  0.1577
Relative absolute error                 30.2007 %
Root relative squared error             57.4556 %
Coverage of cases (0.95 level)          93.2584 %
Mean rel. region size (0.95 level)      16.9476 %
Total Number of Instances               89

=== Detailed Accuracy By Class ===

TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
0,952    0,044    0,870      0,952    0,909      0,881    0,989     0,969     NORMAL
0,857    0,000    1,000      0,857    0,923      0,920    0,972     0,901     T_COLON
0,750    0,000    1,000      0,750    0,857      0,856    0,997     0,970     T_PAN
1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     T_KID
0,333    0,012    0,667      0,333    0,444      0,446    0,867     0,462     T_BLDR
0,667    0,012    0,800      0,667    0,727      0,713    0,934     0,766     T_PROST
1,000    0,012    0,833      1,000    0,909      0,907    1,000     1,000     T_OVARY
0,800    0,025    0,800      0,800    0,800      0,775    0,954     0,805     T_UT
0,600    0,024    0,600      0,600    0,600      0,576    0,876     0,693     T_LUNG
1,000    0,025    0,800      1,000    0,889      0,883    0,995     0,953     T_MESO
0,667    0,000    1,000      0,667    0,800      0,812    0,992     0,867     T_MELA
1,000    0,036    0,667      1,000    0,800      0,802    0,988     0,836     T_BRST
W. Avg.   0,831    0,022    0,839      0,831    0,824      0,811    0,968     0,871

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l   <-- classified as
 20  0  0  0  0  0  0  1  0  0  0  0 |  a = NORMAL
  0  6  0  0  1  0  0  0  0  0  0  0 |  b = T_COLON
  1  0  6  0  0  0  0  0  1  0  0  0 |  c = T_PAN
  0  0  0  4  0  0  0  0  0  0  0  0 |  d = T_KID
  0  0  0  0  2  1  1  0  1  0  0  1 |  e = T_BLDR
  1  0  0  0  0  4  0  0  0  0  0  1 |  f = T_PROST
  0  0  0  0  0  0  5  0  0  0  0  0 |  g = T_OVARY
  1  0  0  0  0  0  0  8  0  0  0  1 |  h = T_UT
  0  0  0  0  0  0  0  1  3  1  0  0 |  i = T_LUNG
  0  0  0  0  0  0  0  0  0  8  0  0 |  j = T_MESO
  0  0  0  0  0  0  0  0  0  1  2  0 |  k = T_MELA
  0  0  0  0  0  0  0  0  0  0  0  6 |  l = T_BRST
```

Ek 1.1.2 SVM öznitelik seçimi uygulanarak mikroRNA bilgisi ve ANN sınıflandırıcı ile çok kategorili kanser sınıflandırması sonuçları

```
Scheme:       weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M
0.2 -N 500 -V 0 -S 0 -E 20 -H a
Relation:     miRNA_mRNA_tek_normal_SVM_Select
Instances:    89
Attributes:   101
Test mode:    89-fold cross-validation

=== Classifier model (full training set) ===

Class NORMAL
    Input
    Node 0
Class T_COLON
    Input
    Node 1
Class T_PAN
    Input
    Node 2
Class T_KID
    Input
    Node 3
Class T_BLDR
    Input
    Node 4
Class T_PROST
    Input
    Node 5
Class T_OVARY
    Input
    Node 6
Class T_UT
    Input
    Node 7
Class T_LUNG
    Input
    Node 8
Class T_MESO
    Input
    Node 9
Class T_MELA
    Input
    Node 10
Class T_BRST
    Input
    Node 11
```

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          77               86.5169 %
Kappa statistic                          0.8471
Mean absolute error                      0.037
Root mean squared error                  0.1339
Relative absolute error                 24.7138 %
Root relative squared error             48.7672 %
Coverage of cases (0.95 level)          96.6292 %
Mean rel. region size (0.95 level)      17.0412 %
Total Number of Instances               89

=== Detailed Accuracy By Class ===

TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
0,952    0,044    0,870      0,952    0,909      0,881    0,990     0,965     NORMAL
0,857    0,000    1,000      0,857    0,923      0,920    0,997     0,968     T_COLON
0,875    0,000    1,000      0,875    0,933      0,930    0,992     0,952     T_PAN
1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     T_KID
0,667    0,000    1,000      0,667    0,800      0,807    0,928     0,765     T_BLDR
0,667    0,012    0,800      0,667    0,727      0,713    0,940     0,815     T_PROST
1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     T_OVARY
0,800    0,025    0,800      0,800    0,800      0,775    0,981     0,920     T_UT
0,600    0,024    0,600      0,600    0,600      0,576    0,986     0,844     T_LUNG
1,000    0,025    0,800      1,000    0,889      0,883    1,000     1,000     T_MESO
0,667    0,000    1,000      0,667    0,800      0,812    1,000     1,000     T_MELA
1,000    0,024    0,750      1,000    0,857      0,856    0,998     0,976     T_BRST
W. Avg.   0,865    0,019    0,876      0,865    0,863      0,851    0,985     0,937

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l   <-- classified as
 20  0  0  0  0  0  0  1  0  0  0  0 |  a = NORMAL
  1  6  0  0  0  0  0  0  0  0  0  0 |  b = T_COLON
  0  0  7  0  0  0  0  0  1  0  0  0 |  c = T_PAN
  0  0  0  4  0  0  0  0  0  0  0  0 |  d = T_KID
  0  0  0  0  4  1  0  0  0  0  0  1 |  e = T_BLDR
  1  0  0  0  0  4  0  0  1  0  0  0 |  f = T_PROST
  0  0  0  0  0  0  5  0  0  0  0  0 |  g = T_OVARY
  1  0  0  0  0  0  0  8  0  0  0  1 |  h = T_UT
  0  0  0  0  0  0  0  1  3  1  0  0 |  i = T_LUNG
  0  0  0  0  0  0  0  0  0  8  0  0 |  j = T_MESO
  0  0  0  0  0  0  0  0  0  1  2  0 |  k = T_MELA
  0  0  0  0  0  0  0  0  0  0  0  6 |  l = T_BRST
```

Ek 1.1.3 Öznitelik seçimi olmadan mikroRNA bilgisi ve SVM sınıflandırıcı ile çok kategorili kanser sınıflandırması sonuçları

```
Scheme:       weka.classifiers.functions.SMO -C 1.0 -L 0.001 -P 1.0E-12 -
N 2 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -C
250007 -E 1.0"
Relation:     miRNA_mRNA_tek_normal
Instances:    89
Attributes:   218
Test mode:    89-fold cross-validation

=== Classifier model (full training set) ===
SMO

Kernel used:
  Linear Kernel: K(x,y) = <x,y>

=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances          69               77.5281 %
Kappa statistic                          0.7445
Mean absolute error                      0.1402
Root mean squared error                  0.2577
Relative absolute error                 93.5912 %
Root relative squared error             93.8712 %
Coverage of cases (0.95 level)         100      %
Mean rel. region size (0.95 level)      76.8727 %
Total Number of Instances               89

=== Detailed Accuracy By Class ===

TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
0,905    0,074    0,792      0,905    0,844      0,795    0,955     0,777     NORMAL
0,714    0,012    0,833      0,714    0,769      0,754    0,964     0,694     T_COLON
0,625    0,012    0,833      0,625    0,714      0,699    0,892     0,652     T_PAN
0,750    0,000    1,000      0,750    0,857      0,861    0,999     0,950     T_KID
0,500    0,000    1,000      0,500    0,667      0,695    0,757     0,610     T_BLDR
0,667    0,036    0,571      0,667    0,615      0,587    0,842     0,440     T_PROST
1,000    0,000    1,000      1,000    1,000      1,000    0,998     0,943     T_OVARY
0,600    0,063    0,545      0,600    0,571      0,515    0,897     0,489     T_UT
0,800    0,024    0,667      0,800    0,727      0,713    0,970     0,592     T_LUNG
1,000    0,025    0,800      1,000    0,889      0,883    0,988     0,800     T_MESO
0,667    0,000    1,000      0,667    0,800      0,812    0,992     0,810     T_MELA
0,833    0,012    0,833      0,833    0,833      0,821    0,990     0,794     T_BRST
W. Avg.   0,775    0,033    0,795      0,775    0,773      0,751    0,934     0,704

=== Confusion Matrix ===
  a  b  c  d  e  f  g  h  i  j  k  l   <-- classified as
 19  0  0  0  0  1  0  1  0  0  0  0 |  a = NORMAL
  1  5  1  0  0  0  0  0  0  0  0  0 |  b = T_COLON
  1  1  5  0  0  1  0  0  0  0  0  0 |  c = T_PAN
  0  0  0  3  0  0  0  0  0  1  0  0 |  d = T_KID
  0  0  0  0  3  1  0  2  0  0  0  0 |  e = T_BLDR
  1  0  0  0  0  4  0  0  1  0  0  0 |  f = T_PROST
  0  0  0  0  0  0  5  0  0  0  0  0 |  g = T_OVARY
  2  0  0  0  0  0  0  6  1  0  0  1 |  h = T_UT
  0  0  0  0  0  0  0  1  4  0  0  0 |  i = T_LUNG
  0  0  0  0  0  0  0  0  0  8  0  0 |  j = T_MESO
  0  0  0  0  0  0  0  0  0  1  2  0 |  k = T_MELA
  0  0  0  0  0  0  0  1  0  0  0  5 |  l = T_BRST
```

Ek 1.1.4 SVM öznitelik seçimi uygulanarak mikroRNA bilgisi ve SVM sınıflandırıcı ile çok kategorili kanser sınıflandırması sonuçları

```
Scheme:       weka.classifiers.functions.SMO -C 1.0 -L 0.001 -P 1.0E-12 -
N 2 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -C
250007 -E 1.0"
Relation:     miRNA_mRNA_tek_normal_SVM_Select
Instances:    89
Attributes:   101
Test mode:    89-fold cross-validation

=== Classifier model (full training set) ===
SMO
Kernel used:
  Linear Kernel: K(x,y) = <x,y>
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances          73               82.0225 %
Kappa statistic                          0.7966
Mean absolute error                      0.1399
Root mean squared error                  0.2572
Relative absolute error                 93.4208 %
Root relative squared error             93.6845 %
Coverage of cases (0.95 level)         100      %
Mean rel. region size (0.95 level)      77.4345 %
Total Number of Instances               89

=== Detailed Accuracy By Class ===

TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
0,905    0,044    0,864      0,905    0,884      0,847    0,972     0,848     NORMAL
0,714    0,037    0,625      0,714    0,667      0,638    0,970     0,628     T_COLON
0,750    0,012    0,857      0,750    0,800      0,784    0,913     0,721     T_PAN
1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     T_KID
0,667    0,000    1,000      0,667    0,800      0,807    0,802     0,697     T_BLDR
0,667    0,024    0,667      0,667    0,667      0,643    0,855     0,595     T_PROST
0,800    0,012    0,800      0,800    0,800      0,788    0,986     0,867     T_OVARY
0,700    0,038    0,700      0,700    0,700      0,662    0,947     0,652     T_UT
0,800    0,012    0,800      0,800    0,800      0,788    0,988     0,740     T_LUNG
1,000    0,012    0,889      1,000    0,941      0,937    0,994     0,889     T_MESO
0,667    0,000    1,000      0,667    0,800      0,812    0,988     0,778     T_MELA
1,000    0,012    0,857      1,000    0,923      0,920    0,994     0,857     T_BRST
W. Avg.   0,820    0,024    0,827      0,820    0,819      0,800    0,951     0,774

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l   <-- classified as
 19  0  0  0  0  1  0  1  0  0  0  0 |  a = NORMAL
  1  5  1  0  0  0  0  0  0  0  0  0 |  b = T_COLON
  0  1  6  0  0  1  0  0  0  0  0  0 |  c = T_PAN
  0  0  0  4  0  0  0  0  0  0  0  0 |  d = T_KID
  0  1  0  0  4  0  1  0  0  0  0  0 |  e = T_BLDR
  1  0  0  0  0  4  0  0  1  0  0  0 |  f = T_PROST
  0  0  0  0  0  0  4  1  0  0  0  0 |  g = T_OVARY
  1  1  0  0  0  0  0  7  0  0  0  1 |  h = T_UT
  0  0  0  0  0  0  0  1  4  0  0  0 |  i = T_LUNG
  0  0  0  0  0  0  0  0  0  8  0  0 |  j = T_MESO
  0  0  0  0  0  0  0  0  0  1  2  0 |  k = T_MELA
  0  0  0  0  0  0  0  0  0  0  0  6 |  l = T_BRST
```

Ek 1.1.5 Öznitelik seçimi olmadan mikroRNA bilgisi ve karar ağacı sınıflandırıcı ile çok kategorili kanser sınıflandırması sonuçları

```
Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     miRNA_mRNA_tek_normal
Instances:    89
Attributes:   218
Test mode:    89-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

EAM208 <= 5.24708
|   EAM225 <= 5.0387
|   |   EAM288 <= 7.41526: T_MESO (8.0)
|   |   EAM288 > 7.41526: T_KID (4.0)
|   EAM225 > 5.0387: T_MELA (3.0)
EAM208 > 5.24708
|   EAM273 <= 6.21537
|   |   EAM200 <= 5.41606: T_LUNG (5.0)
|   |   EAM200 > 5.41606: T_PROST (4.0)
|   EAM273 > 6.21537
|   |   EAM298 <= 5.72849
|   |   |   EAM276 <= 5.6665
|   |   |   |   EAM342 <= 7.78799
|   |   |   |   |   EAM103 <= 5.34993: T_BLDR (5.0)
|   |   |   |   |   EAM103 > 5.34993: T_BRST (5.0)
|   |   |   |   EAM342 > 7.78799
|   |   |   |   |   EAM317 <= 5.3395: T_OVARY (5.0)
|   |   |   |   |   EAM317 > 5.3395: T_PROST (2.0)
|   |   |   EAM276 > 5.6665: T_UT (7.0/1.0)
|   |   EAM298 > 5.72849
|   |   |   EAM270 <= 8.91708
|   |   |   |   EAM238 <= 6.67626: T_PAN (6.0)
|   |   |   |   EAM238 > 6.67626: T_COLON (7.0)
|   |   |   EAM270 > 8.91708
|   |   |   |   EAM317 <= 7.119
|   |   |   |   |   EAM229 <= 7.7224: NORMAL (22.0/1.0)
|   |   |   |   |   EAM229 > 7.7224: T_UT (3.0)
|   |   |   |   EAM317 > 7.119: T_PAN (3.0/1.0)

Number of Leaves  :    15

Size of the tree :     29
```

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          46               51.6854 %
Kappa statistic                          0.4518
Mean absolute error                      0.0812
Root mean squared error                  0.276
Relative absolute error                 54.2027 %
Root relative squared error            100.5476 %
Coverage of cases (0.95 level)          56.1798 %
Mean rel. region size (0.95 level)       9.4569 %
Total Number of Instances               89

=== Detailed Accuracy By Class ===

TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
0,714    0,118    0,652      0,714    0,682      0,579    0,830     0,620     NORMAL
0,571    0,024    0,667      0,571    0,615      0,587    0,774     0,415     T_COLON
0,500    0,037    0,571      0,500    0,533      0,492    0,728     0,331     T_PAN
0,000    0,012    0,000      0,000    0,000      -0,023   0,488     0,045     T_KID
0,167    0,048    0,200      0,167    0,182      0,129    0,561     0,112     T_BLDR
0,667    0,048    0,500      0,667    0,571      0,542    0,809     0,356     T_PROST
0,400    0,048    0,333      0,400    0,364      0,324    0,681     0,234     T_OVARY
0,300    0,101    0,273      0,300    0,286      0,191    0,628     0,221     T_UT
0,600    0,024    0,600      0,600    0,600      0,576    0,788     0,382     T_LUNG
0,625    0,037    0,625      0,625    0,625      0,588    0,794     0,424     T_MESO
0,667    0,000    1,000      0,667    0,800      0,812    0,833     0,678     T_MELA
0,500    0,048    0,429      0,500    0,462      0,421    0,714     0,248     T_BRST
W. Avg.  0,517    0,062      0,507    0,517      0,509    0,452     0,737     0,379

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l   <-- classified as
 15  0  0  0  2  0  0  4  0  0  0  0 |  a = NORMAL
  2  4  1  0  0  0  0  0  0  0  0  0 |  b = T_COLON
  2  1  4  0  0  0  0  1  0  0  0  0 |  c = T_PAN
  0  0  1  0  0  0  0  0  1  2  0  0 |  d = T_KID
  0  0  0  0  1  1  1  1  0  0  0  2 |  e = T_BLDR
  0  0  0  0  0  4  1  1  0  0  0  0 |  f = T_PROST
  0  0  0  0  0  2  2  0  0  0  0  1 |  g = T_OVARY
  3  1  0  0  0  0  2  3  0  0  0  1 |  h = T_UT
  1  0  1  0  0  0  0  0  3  0  0  0 |  i = T_LUNG
  0  0  0  1  1  0  0  0  1  5  0  0 |  j = T_MESO
  0  0  0  0  0  0  0  0  0  1  2  0 |  k = T_MELA
  0  0  0  0  1  1  0  1  0  0  0  3 |  l = T_BRST
```

Ek 1.1.6 SVM öznitelik seçimi uygulanarak mikroRNA bilgisi ve karar ağacı sınıflandırıcı ile çok kategorili kanser sınıflandırması sonuçları

```
Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     miRNA_mRNA_tek_normal_SVM_Select
Instances:    89
Attributes:   101
Test mode:    89-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

EAM208 <= 5.24708
|   EAM225 <= 5.0387
|   |   EAM288 <= 7.41526: T_MESO (8.0)
|   |   EAM288 > 7.41526: T_KID (4.0)
|   EAM225 > 5.0387: T_MELA (3.0)
EAM208 > 5.24708
|   EAM273 <= 6.21537
|   |   EAM200 <= 5.41606: T_LUNG (5.0)
|   |   EAM200 > 5.41606: T_PROST (4.0)
|   EAM273 > 6.21537
|   |   EAM298 <= 5.72849
|   |   |   EAM276 <= 5.6665
|   |   |   |   EAM342 <= 7.78799
|   |   |   |   |   EAM103 <= 5.34993: T_BLDR (5.0)
|   |   |   |   |   EAM103 > 5.34993: T_BRST (5.0)
|   |   |   |   EAM342 > 7.78799
|   |   |   |   |   EAM225 <= 5.0387: T_PROST (2.0)
|   |   |   |   |   EAM225 > 5.0387: T_OVARY (5.0)
|   |   |   EAM276 > 5.6665: T_UT (7.0/1.0)
|   |   EAM298 > 5.72849
|   |   |   EAM331 <= 9.21451
|   |   |   |   EAM238 <= 6.67626: T_PAN (5.0)
|   |   |   |   EAM238 > 6.67626: T_COLON (7.0)
|   |   |   EAM331 > 9.21451
|   |   |   |   EAM238 <= 6.08924: T_PAN (3.0/1.0)
|   |   |   |   EAM238 > 6.08924
|   |   |   |   |   EAM229 <= 7.7224: NORMAL (23.0/2.0)
|   |   |   |   |   EAM229 > 7.7224: T_UT (3.0)

Number of Leaves  :    15

Size of the tree :     29
```

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          41               46.0674 %
Kappa statistic                          0.3896
Mean absolute error                      0.093
Root mean squared error                  0.2917
Relative absolute error                 62.0557 %
Root relative squared error            106.2355 %
Coverage of cases (0.95 level)          49.4382 %
Mean rel. region size (0.95 level)      11.4232 %
Total Number of Instances               89

=== Detailed Accuracy By Class ===

TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
0,619    0,132    0,591      0,619    0,605      0,479    0,702     0,456     NORMAL
0,429    0,061    0,375      0,429    0,400      0,346    0,692     0,302     T_COLON
0,125    0,062    0,167      0,125    0,143      0,072    0,627     0,179     T_PAN
0,000    0,024    0,000      0,000    0,000      -0,033   0,482     0,045     T_KID
0,167    0,024    0,333      0,167    0,222      0,198    0,547     0,140     T_BLDR
0,667    0,072    0,400      0,667    0,500      0,472    0,797     0,289     T_PROST
0,200    0,036    0,250      0,200    0,222      0,183    0,585     0,145     T_OVARY
0,600    0,076    0,500      0,600    0,545      0,485    0,773     0,480     T_UT
0,600    0,048    0,429      0,600    0,500      0,473    0,776     0,280     T_LUNG
0,500    0,037    0,571      0,500    0,533      0,492    0,731     0,331     T_MESO
0,667    0,012    0,667      0,667    0,667      0,655    0,828     0,456     T_MELA
0,500    0,024    0,600      0,500    0,545      0,518    0,720     0,334     T_BRST
W. Avg.  0,461    0,068      0,442    0,461      0,445    0,383     0,694     0,324

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l   <-- classified as
 13  2  2  0  0  0  0  3  0  0  1  0 |  a = NORMAL
  2  3  2  0  0  0  0  0  0  0  0  0 |  b = T_COLON
  4  2  1  0  0  0  0  1  0  0  0  0 |  c = T_PAN
  0  0  0  0  0  0  0  0  2  2  0  0 |  d = T_KID
  1  0  0  0  1  1  1  0  1  0  0  1 |  e = T_BLDR
  0  0  0  0  0  4  1  1  0  0  0  0 |  f = T_PROST
  0  0  0  0  0  3  1  0  0  0  0  1 |  g = T_OVARY
  2  0  0  0  0  1  1  6  0  0  0  0 |  h = T_UT
  0  1  1  0  0  0  0  0  3  0  0  0 |  i = T_LUNG
  0  0  0  2  1  0  0  0  1  4  0  0 |  j = T_MESO
  0  0  0  0  0  0  0  0  0  1  2  0 |  k = T_MELA
  0  0  0  0  1  1  0  1  0  0  0  3 |  l = T_BRST
```

Ek 1.1.7 Öznitelik seçimi olmadan mikroRNA bilgisi ve KNN sınıflandırıcı ile çok kategorili kanser sınıflandırması sonuçları

```
Scheme:       weka.classifiers.lazy.IBk -K 1 -W 0 -A
"weka.core.neighboursearch.LinearNNSearch -A
\"weka.core.EuclideanDistance -R first-last\""
Relation:     miRNA_mRNA_tek_normal
Instances:    89
Attributes:   218
Test mode:    89-fold cross-validation

=== Classifier model (full training set) ===
IB1 instance-based classifier
using 1 nearest neighbour(s) for classification

=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances          61               68.5393 %
Kappa statistic                          0.6436
Mean absolute error                      0.0645
Root mean squared error                  0.2174
Relative absolute error                 43.0429 %
Root relative squared error             79.1683 %
Coverage of cases (0.95 level)          86.5169 %
Mean rel. region size (0.95 level)      58.3333 %
Total Number of Instances               89

=== Detailed Accuracy By Class ===

TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
0,857    0,074    0,783      0,857    0,818      0,760    0,892     0,705     NORMAL
0,714    0,024    0,714      0,714    0,714      0,690    0,845     0,533     T_COLON
0,625    0,037    0,625      0,625    0,625      0,588    0,794     0,424     T_PAN
1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     T_KID
0,500    0,000    1,000      0,500    0,667      0,695    0,750     0,534     T_BLDR
0,667    0,024    0,667      0,667    0,667      0,643    0,821     0,467     T_PROST
0,600    0,012    0,750      0,600    0,667      0,654    0,794     0,472     T_OVARY
0,400    0,063    0,444      0,400    0,421      0,353    0,668     0,245     T_UT
0,800    0,060    0,444      0,800    0,571      0,566    0,870     0,367     T_LUNG
0,750    0,037    0,667      0,750    0,706      0,676    0,856     0,522     T_MESO
0,333    0,012    0,500      0,333    0,400      0,392    0,661     0,189     T_MELA
0,667    0,012    0,800      0,667    0,727      0,713    0,827     0,556     T_BRST
W. Avg.   0,685    0,040    0,702      0,685    0,683      0,652    0,823     0,524

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l   <-- classified as
 18  0  0  0  0  1  0  1  1  0  0  0 |  a = NORMAL
  1  5  1  0  0  0  0  0  0  0  0  0 |  b = T_COLON
  1  1  5  0  0  0  0  0  1  0  0  0 |  c = T_PAN
  0  0  0  4  0  0  0  0  0  0  0  0 |  d = T_KID
  0  0  1  0  3  0  0  1  0  1  0  0 |  e = T_BLDR
  1  0  0  0  0  4  0  0  1  0  0  0 |  f = T_PROST
  0  0  1  0  0  0  3  1  0  0  0  0 |  g = T_OVARY
  2  1  0  0  0  0  1  4  1  0  0  1 |  h = T_UT
  0  0  0  0  0  0  0  1  4  0  0  0 |  i = T_LUNG
  0  0  0  0  0  0  0  0  1  6  1  0 |  j = T_MESO
  0  0  0  0  0  0  0  0  0  2  1  0 |  k = T_MELA
  0  0  0  0  0  1  0  1  0  0  0  4 |  l = T_BRST
```

Ek 1.1.8 SVM öznitelik seçimi uygulanarak mikroRNA bilgisi ve KNN sınıflandırıcı ile çok kategorili kanser sınıflandırması sonuçları

```
Scheme:       weka.classifiers.lazy.IBk -K 1 -W 0 -A
"weka.core.neighboursearch.LinearNNSearch -A
\"weka.core.EuclideanDistance -R first-last\""
Relation:     miRNA_mRNA_tek_normal_SVM_Select
Instances:    89
Attributes:   101
Test mode:    89-fold cross-validation

=== Classifier model (full training set) ===
IB1 instance-based classifier
using 1 nearest neighbour(s) for classification
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances          65               73.0337 %
Kappa statistic                          0.6935
Mean absolute error                      0.0579
Root mean squared error                  0.2016
Relative absolute error                 38.6423 %
Root relative squared error             73.4377 %
Coverage of cases (0.95 level)          87.6404 %
Mean rel. region size (0.95 level)      58.3333 %
Total Number of Instances               89

=== Detailed Accuracy By Class ===

TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
0,857    0,088    0,750      0,857    0,800      0,736    0,884     0,677     NORMAL
0,714    0,000    1,000      0,714    0,833      0,835    0,857     0,737     T_COLON
0,625    0,012    0,833      0,625    0,714      0,699    0,806     0,555     T_PAN
1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     T_KID
0,667    0,012    0,800      0,667    0,727      0,713    0,827     0,556     T_BLDR
0,667    0,024    0,667      0,667    0,667      0,643    0,821     0,467     T_PROST
0,400    0,012    0,667      0,400    0,500      0,495    0,694     0,300     T_OVARY
0,500    0,076    0,455      0,500    0,476      0,407    0,712     0,283     T_UT
0,800    0,024    0,667      0,800    0,727      0,713    0,888     0,545     T_LUNG
1,000    0,025    0,800      1,000    0,889      0,883    0,988     0,800     T_MESO
0,667    0,000    1,000      0,667    0,800      0,812    0,833     0,678     T_MELA
0,667    0,036    0,571      0,667    0,615      0,587    0,815     0,403     T_BRST
W. Avg.  0,730    0,040      0,744    0,730      0,728    0,698     0,845     0,583

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l   <-- classified as
 18  0  0  0  0  1  0  1  1  0  0  0 |  a = NORMAL
  1  5  1  0  0  0  0  0  0  0  0  0 |  b = T_COLON
  2  0  5  0  0  0  1  0  0  0  0  0 |  c = T_PAN
  0  0  0  4  0  0  0  0  0  0  0  0 |  d = T_KID
  0  0  0  0  4  0  0  1  0  1  0  0 |  e = T_BLDR
  1  0  0  0  0  4  0  0  0  0  0  1 |  f = T_PROST
  0  0  0  0  1  0  2  2  0  0  0  0 |  g = T_OVARY
  2  0  0  0  0  0  0  5  1  0  0  2 |  h = T_UT
  0  0  0  0  0  0  0  1  4  0  0  0 |  i = T_LUNG
  0  0  0  0  0  0  0  0  0  8  0  0 |  j = T_MESO
  0  0  0  0  0  0  0  0  0  1  2  0 |  k = T_MELA
  0  0  0  0  0  1  0  1  0  0  0  4 |  l = T_BRST
```

Ek 1.1.9 Öznitelik seçimi olmadan mikroRNA bilgisi ve NBM sınıflandırıcı ile çok kategorili kanser sınıflandırması sonuçları

```
=== Run information ===

Scheme:       weka.classifiers.bayes.NaiveBayesMultinomial
Relation:     miRNA_mRNA_tek_normal
Instances:    89
Attributes:   218
Test mode:    89-fold cross-validation

=== Classifier model (full training set) ===

The independent probability of a class
--------------------------------------
NORMAL      0.21782178217821782
T_COLON     0.07920792079207921
T_PAN       0.0891089108910891
T_KID       0.04950495049504951
T_BLDR      0.06930693069306931
T_PROST     0.06930693069306931
T_OVARY     0.0594059405940594
T_UT        0.10891089108910891
T_LUNG      0.0594059405940594
T_MESO      0.0891089108910891
T_MELA      0.039603960396039604
T_BRST      0.06930693069306931


=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          67               75.2809 %
Kappa statistic                          0.7201
Mean absolute error                      0.0535
Root mean squared error                  0.1742
Relative absolute error                 35.7366 %
Root relative squared error             63.4593 %
Coverage of cases (0.95 level)          92.1348 %
Mean rel. region size (0.95 level)      19.4757 %
Total Number of Instances               89
```

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0,810	0,088	0,739	0,810	0,773	0,700	0,946	0,843	NORMAL
0,857	0,024	0,750	0,857	0,800	0,784	0,981	0,884	T_COLON
0,750	0,025	0,750	0,750	0,750	0,725	0,975	0,850	T_PAN
0,750	0,000	1,000	0,750	0,857	0,861	1,000	1,000	T_KID
0,500	0,024	0,600	0,500	0,545	0,518	0,908	0,638	T_BLDR
0,667	0,012	0,800	0,667	0,727	0,713	0,924	0,810	T_PROST
0,600	0,000	1,000	0,600	0,750	0,766	0,986	0,860	T_OVARY
0,500	0,013	0,833	0,500	0,625	0,614	0,881	0,677	T_UT
0,800	0,024	0,667	0,800	0,727	0,713	0,979	0,793	T_LUNG
1,000	0,025	0,800	1,000	0,889	0,883	1,000	1,000	T_MESO
0,667	0,000	1,000	0,667	0,800	0,812	1,000	1,000	T_MELA
1,000	0,048	0,600	1,000	0,750	0,756	0,990	0,856	T_BRST
W. Avg.	0,753	0,036	0,773	0,753	0,748	0,723	0,956	0,838

=== Confusion Matrix ===

a	b	c	d	e	f	g	h	i	j	k	l	<-- classified as
17	2	2	0	0	0	0	0	0	0	0	0	a = NORMAL
1	6	0	0	0	0	0	0	0	0	0	0	b = T_COLON
1	0	6	0	0	0	0	0	1	0	0	0	c = T_PAN
0	0	0	3	0	0	0	0	0	1	0	0	d = T_KID
0	0	0	0	3	1	0	0	0	0	0	2	e = T_BLDR
1	0	0	0	0	4	0	0	0	0	0	1	f = T_PROST
0	0	0	0	1	0	3	1	0	0	0	0	g = T_OVARY
3	0	0	0	0	0	0	5	1	0	0	1	h = T_UT
0	0	0	0	1	0	0	0	4	0	0	0	i = T_LUNG
0	0	0	0	0	0	0	0	0	8	0	0	j = T_MESO
0	0	0	0	0	0	0	0	0	1	2	0	k = T_MELA
0	0	0	0	0	0	0	0	0	0	0	6	l = T_BRST

Ek 1.1.10 SVM öznitelik seçimi uygulanarak mikroRNA bilgisi ve NBM sınıflandırıcı ile çok kategorili kanser sınıflandırması sonuçları

```
=== Run information ===

Scheme:       weka.classifiers.bayes.NaiveBayesMultinomial
Relation:     miRNA_mRNA_tek_normal_SVM_Select
Instances:    89
Attributes:   101
Test mode:    89-fold cross-validation

=== Classifier model (full training set) ===

The independent probability of a class
--------------------------------------
NORMAL     0.21782178217821782
T_COLON    0.07920792079207921
T_PAN      0.0891089108910891
T_KID      0.04950495049504951
T_BLDR     0.06930693069306931
T_PROST    0.06930693069306931
T_OVARY    0.0594059405940594
T_UT       0.10891089108910891
T_LUNG     0.0594059405940594
T_MESO     0.0891089108910891
T_MELA     0.039603960396039604
T_BRST     0.06930693069306931


=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          67               75.2809 %
Kappa statistic                          0.7189
Mean absolute error                      0.0599
Root mean squared error                  0.1668
Relative absolute error                 39.9799 %
Root relative squared error             60.7475 %
Coverage of cases (0.95 level)          95.5056 %
Mean rel. region size (0.95 level)      26.4981 %
Total Number of Instances               89
```

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0,762	0,132	0,640	0,762	0,696	0,595	0,924	0,703	NORMAL
0,714	0,024	0,714	0,714	0,714	0,690	0,979	0,872	T_COLON
0,750	0,037	0,667	0,750	0,706	0,676	0,981	0,889	T_PAN
0,750	0,000	1,000	0,750	0,857	0,861	1,000	1,000	T_KID
0,667	0,012	0,800	0,667	0,727	0,713	0,910	0,732	T_BLDR
0,667	0,012	0,800	0,667	0,727	0,713	0,918	0,827	T_PROST
0,800	0,012	0,800	0,800	0,800	0,788	0,995	0,927	T_OVARY
0,500	0,013	0,833	0,500	0,625	0,614	0,906	0,728	T_UT
0,800	0,012	0,800	0,800	0,800	0,788	0,998	0,967	T_LUNG
1,000	0,025	0,800	1,000	0,889	0,883	1,000	1,000	T_MESO
0,667	0,000	1,000	0,667	0,800	0,812	1,000	1,000	T_MELA
1,000	0,012	0,857	1,000	0,923	0,920	0,996	0,948	T_BRST
W. Avg.	0,753	0,044	0,767	0,753	0,750	0,717	0,956	0,840

=== Confusion Matrix ===

a	b	c	d	e	f	g	h	i	j	k	l	<-- classified as
16	2	2	0	0	1	0	0	0	0	0	0	a = NORMAL
1	5	1	0	0	0	0	0	0	0	0	0	b = T_COLON
2	0	6	0	0	0	0	0	0	0	0	0	c = T_PAN
0	0	0	3	0	0	0	0	0	1	0	0	d = T_KID
1	0	0	0	4	0	1	0	0	0	0	0	e = T_BLDR
1	0	0	0	0	4	0	0	1	0	0	0	f = T_PROST
0	0	0	0	0	0	4	1	0	0	0	0	g = T_OVARY
4	0	0	0	0	0	0	5	0	0	0	1	h = T_UT
0	0	0	0	1	0	0	0	4	0	0	0	i = T_LUNG
0	0	0	0	0	0	0	0	0	8	0	0	j = T_MESO
0	0	0	0	0	0	0	0	0	1	2	0	k = T_MELA
0	0	0	0	0	0	0	0	0	0	0	6	l = T_BRST

Ek 1.2 mRNA bilgisi ile çok kategorili kanser sınıflandırması sonuçları

Ek 1.2.1 Öznitelik seçimi olmadan mRNA bilgisi ve ANN sınıflandırıcı ile çok kategorili kanser sınıflandırması sonuçları

```
Scheme:weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N
500 -V 0 -S 0 -E 20 -H 2
Relation:     mRNA_tek_normal
Instances:    89
Attributes:   16064
Test mode:89-fold cross-validation

=== Classifier model (full training set) ===

Class NORMAL
    Input
    Node 0
Class T_COLON
    Input
    Node 1
Class T_PAN
    Input
    Node 2
Class T_KID
    Input
    Node 3
Class T_BLDR
    Input
    Node 4
Class T_PROST
    Input
    Node 5
Class T_OVARY
    Input
    Node 6
Class T_UT
    Input
    Node 7
Class T_LUNG
    Input
    Node 8
Class T_MESO
    Input
    Node 9
Class T_MELA
    Input
    Node 10
Class T_BRST
    Input
    Node 11
```

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          21               23.5955 %
Incorrectly Classified Instances        68               76.4045 %
Kappa statistic                          0.0039
Mean absolute error                      0.1495
Root mean squared error                  0.2749
Relative absolute error                 99.7713 %
Root relative squared error             100.121 %
Total Number of Instances               89

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0.971      0.241     1         0.389      0.425    NORMAL
                 0         0          0         0         0          0.169    T_COLON
                 0         0          0         0         0          0.091    T_PAN
                 0         0          0         0         0          0.024    T_KID
                 0         0          0         0         0          0.014    T_BLDR
                 0         0          0         0         0          0.028    T_PROST
                 0         0          0         0         0          0.19     T_OVARY
                 0         0.013      0         0         0          0.058    T_UT
                 0         0          0         0         0          0        T_LUNG
                 0         0.012      0         0         0          0.228    T_MESO
                 0         0          0         0         0          0        T_MELA
                 0         0          0         0         0          0.032    T_BRST
Weighted Avg.    0.236     0.232      0.057     0.236     0.092      0.166

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l   <-- classified as
 21  0  0  0  0  0  0  0  0  0  0  0 |  a = NORMAL
  6  0  0  0  0  0  0  0  0  1  0  0 |  b = T_COLON
  8  0  0  0  0  0  0  0  0  0  0  0 |  c = T_PAN
  4  0  0  0  0  0  0  0  0  0  0  0 |  d = T_KID
  6  0  0  0  0  0  0  0  0  0  0  0 |  e = T_BLDR
  6  0  0  0  0  0  0  0  0  0  0  0 |  f = T_PROST
  4  0  0  0  0  0  0  1  0  0  0  0 |  g = T_OVARY
 10  0  0  0  0  0  0  0  0  0  0  0 |  h = T_UT
  5  0  0  0  0  0  0  0  0  0  0  0 |  i = T_LUNG
  8  0  0  0  0  0  0  0  0  0  0  0 |  j = T_MESO
  3  0  0  0  0  0  0  0  0  0  0  0 |  k = T_MELA
  6  0  0  0  0  0  0  0  0  0  0  0 |  l = T_BRST
```

Ek 1.2.2 SVM öznitelik seçimi uygulanarak mRNA bilgisi ve ANN sınıflandırıcı ile çok kategorili kanser sınıflandırması sonuçları

```
Scheme:       weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M
0.2 -N 500 -V 0 -S 0 -E 20 -H a
Relation:     miRNA_mRNA_tek_normal_SVM_Select
Instances:    89
Attributes:   101
Test mode:    89-fold cross-validation

=== Classifier model (full training set) ===

Class NORMAL
    Input
    Node 0
Class T_COLON
    Input
    Node 1
Class T_PAN
    Input
    Node 2
Class T_KID
    Input
    Node 3
Class T_BLDR
    Input
    Node 4
Class T_PROST
    Input
    Node 5
Class T_OVARY
    Input
    Node 6
Class T_UT
    Input
    Node 7
Class T_LUNG
    Input
    Node 8
Class T_MESO
    Input
    Node 9
Class T_MELA
    Input
    Node 10
Class T_BRST
    Input
    Node 11
```

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          86               96.6292 %
Kappa statistic                          0.9619
Mean absolute error                      0.0196
Root mean squared error                  0.0764
Relative absolute error                 13.0574 %
Root relative squared error             27.844  %
Coverage of cases (0.95 level)          98.8764 %
Mean rel. region size (0.95 level)      16.3858 %
Total Number of Instances               89

=== Detailed Accuracy By Class ===

TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
1,000    0,015    0,955      1,000    0,977      0,970    0,999     0,998     NORMAL
1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     T_COLON
0,875    0,000    1,000      0,875    0,933      0,930    0,988     0,938     T_PAN
1,000    0,012    0,800      1,000    0,889      0,889    0,997     0,950     T_KID
1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     T_BLDR
1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     T_PROST
0,800    0,000    1,000      0,800    0,889      0,889    0,998     0,967     T_OVARY
1,000    0,013    0,909      1,000    0,952      0,947    1,000     1,000     T_UT
1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     T_LUNG
1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     T_MESO
1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     T_MELA
0,833    0,000    1,000      0,833    0,909      0,907    1,000     1,000     T_BRST
W. Avg.   0,966    0,005    0,970      0,966    0,966      0,963     0,998     0,990

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l   <-- classified as
 21  0  0  0  0  0  0  0  0  0  0  0 |  a = NORMAL
  0  7  0  0  0  0  0  0  0  0  0  0 |  b = T_COLON
  0  0  7  1  0  0  0  0  0  0  0  0 |  c = T_PAN
  0  0  0  4  0  0  0  0  0  0  0  0 |  d = T_KID
  0  0  0  0  6  0  0  0  0  0  0  0 |  e = T_BLDR
  0  0  0  0  0  6  0  0  0  0  0  0 |  f = T_PROST
  0  0  0  0  0  0  4  1  0  0  0  0 |  g = T_OVARY
  0  0  0  0  0  0  0 10  0  0  0  0 |  h = T_UT
  0  0  0  0  0  0  0  0  5  0  0  0 |  i = T_LUNG
  0  0  0  0  0  0  0  0  0  8  0  0 |  j = T_MESO
  0  0  0  0  0  0  0  0  0  0  3  0 |  k = T_MELA
  1  0  0  0  0  0  0  0  0  0  0  5 |  l = T_BRST
```

Ek 1.2.3 Öznitelik seçimi olmadan mRNA bilgisi ve SVM sınıflandırıcı ile çok kategorili kanser sınıflandırması sonuçları

```
Scheme:weka.classifiers.functions.SMO -C 1.0 -L 0.0010 -P 1.0E-12 -N 2 -V
-1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -C 250007
-E 1.0"
Relation:     mRNA_tek_normal
Instances:    89
Attributes:   16064
Test mode:89-fold cross-validation

=== Classifier model (full training set) ===
SMO
Kernel used:
  Linear Kernel: K(x,y) = <x,y>

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          67               75.2809 %
Incorrectly Classified Instances        22               24.7191 %
Kappa statistic                          0.7193
Mean absolute error                      0.1421
Root mean squared error                  0.2616
Relative absolute error                 94.8603 %
Root relative squared error             95.2698 %
Total Number of Instances               89

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.952     0.029      0.909     0.952     0.93       0.986    NORMAL
                 0.857     0.024      0.75      0.857     0.8        0.941    T_COLON
                 0.875     0.012      0.875     0.875     0.875      0.87     T_PAN
                 0.5       0          1         0.5       0.667      0.81     T_KID
                 0.5       0.012      0.75      0.5       0.6        0.779    T_BLDR
                 0.667     0          1         0.667     0.8        0.826    T_PROST
                 0.4       0.036      0.4       0.4       0.4        0.758    T_OVARY
                 0.9       0.063      0.643     0.9       0.75       0.872    T_UT
                 0.8       0.012      0.8       0.8       0.8        0.95     T_LUNG
                 0.875     0.012      0.875     0.875     0.875      0.953    T_MESO
                 0         0          0         0         0          0.777    T_MELA
                 0.5       0.072      0.333     0.5       0.4        0.896    T_BRST
Weighted Avg.    0.753     0.027      0.756     0.753     0.743      0.896

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l   <-- classified as
 20  0  0  0  0  0  0  0  1  0  0  0 |  a = NORMAL
  0  6  0  0  0  0  0  0  0  0  0  1 |  b = T_COLON
  0  0  7  0  0  0  0  0  0  0  0  1 |  c = T_PAN
  1  0  0  2  0  0  0  0  0  0  0  1 |  d = T_KID
  0  0  0  0  3  0  1  1  0  0  0  1 |  e = T_BLDR
  0  0  1  0  0  4  0  1  0  0  0  0 |  f = T_PROST
  0  0  0  0  1  0  2  2  0  0  0  0 |  g = T_OVARY
  0  0  0  0  0  0  1  9  0  0  0  0 |  h = T_UT
  0  0  0  0  0  0  0  0  4  0  0  1 |  i = T_LUNG
  0  0  0  0  0  0  1  0  0  7  0  0 |  j = T_MESO
  0  0  0  0  0  0  0  1  0  1  0  1 |  k = T_MELA
  1  2  0  0  0  0  0  0  0  0  0  3 |  l = T_BRST
```

Ek 1.2.4 SVM öznitelik seçimi uygulanarak mRNA bilgisi ve SVM sınıflandırıcı ile çok kategorili kanser sınıflandırması sonuçları

```
Scheme:       weka.classifiers.functions.SMO -C 1.0 -L 0.001 -P 1.0E-12 -
N 2 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -C
250007 -E 1.0"
Relation:     miRNA_mRNA_tek_normal_SVM_Select
Instances:    89
Attributes:   101
Test mode:    89-fold cross-validation

=== Classifier model (full training set) ===
SMO
Kernel used:
  Linear Kernel: K(x,y) = <x,y>
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          82               92.1348 %
Kappa statistic                          0.9108
Mean absolute error                      0.1394
Root mean squared error                  0.2562
Relative absolute error                 93.0798 %
Root relative squared error             93.3255 %
Coverage of cases (0.95 level)          98.8764 %
Mean rel. region size (0.95 level)      78.1835 %
Total Number of Instances               89

=== Detailed Accuracy By Class ===

TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
1,000    0,029    0,913      1,000    0,955      0,941    0,993     0,955     NORMAL
1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     T_COLON
0,875    0,000    1,000      0,875    0,933      0,930    0,875     0,886     T_PAN
0,750    0,000    1,000      0,750    0,857      0,861    0,990     0,861     T_KID
0,833    0,024    0,714      0,833    0,769      0,754    0,974     0,738     T_BLDR
0,833    0,012    0,833      0,833    0,833      0,821    0,969     0,869     T_PROST
   0,800    0,000    1,000      0,800    0,889      0,889    0,987     0,863     T_OVARY
1,000    0,013    0,909      1,000    0,952      0,947    0,994     0,909     T_UT
0,800    0,000    1,000      0,800    0,889      0,889    0,986     0,859     T_LUNG
1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     T_MESO
1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     T_MELA
0,833    0,012    0,833      0,833    0,833      0,821    0,991     0,806     T_BRST
W. Avg.    0,921    0,012    0,928      0,921    0,921      0,915    0,980     0,907

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l   <-- classified as
 21  0  0  0  0  0  0  0  0  0  0  0 |  a = NORMAL
  0  7  0  0  0  0  0  0  0  0  0  0 |  b = T_COLON
  0  0  7  0  1  0  0  0  0  0  0  0 |  c = T_PAN
  0  0  0  3  1  0  0  0  0  0  0  0 |  d = T_KID
  0  0  0  0  5  1  0  0  0  0  0  0 |  e = T_BLDR
  1  0  0  0  0  5  0  0  0  0  0  0 |  f = T_PROST
  0  0  0  0  0  0  4  1  0  0  0  0 |  g = T_OVARY
  0  0  0  0  0  0  0 10  0  0  0  0 |  h = T_UT
  0  0  0  0  0  0  0  0  4  0  0  1 |  i = T_LUNG
  0  0  0  0  0  0  0  0  0  8  0  0 |  j = T_MESO
  0  0  0  0  0  0  0  0  0  0  3  0 |  k = T_MELA
  1  0  0  0  0  0  0  0  0  0  0  5 |  l = T_BRST
```

Ek 1.2.5 Öznitelik seçimi olmadan mRNA bilgisi ve karar ağacı sınıflandırıcı ile çok kategorili kanser sınıflandırması sonuçları

```
Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     mRNA_tek_normal
Instances:    89
Attributes:   16064
Test mode:89-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

Hu6800/X04741_at <= 7.183
|   Hu35KsubA/AA055247_at <= 8.5648
|   |   Hu35KsubA/U15197_at <= 8.59992
|   |   |   Hu6800/M59499_at <= 6.6907
|   |   |   |   Hu6800/X58079_at <= 6.8637
|   |   |   |   |   Hu6800/X64728_at <= 6.0766
|   |   |   |   |   |   Hu6800/L07592_at <= 6.9727
|   |   |   |   |   |   |   Hu6800/X03635_at <= 5.15665
|   |   |   |   |   |   |   |   Hu6800/J00306_at <= 5.13684: T_COLON
(7.0)
|   |   |   |   |   |   |   |   Hu6800/J00306_at > 5.13684: T_PAN (8.0)
|   |   |   |   |   |   |   Hu6800/X03635_at > 5.15665: T_UT (10.0)
|   |   |   |   |   |   Hu6800/L07592_at > 6.9727: T_BLDR (6.0)
|   |   |   |   |   Hu6800/X64728_at > 6.0766: T_PROST (2.0)
|   |   |   |   Hu6800/X58079_at > 6.8637
|   |   |   |   |   Hu6800/HG4582-HT4987_at <= 5: T_OVARY (5.0)
|   |   |   |   |   Hu6800/HG4582-HT4987_at > 5: T_MELA (3.0)
|   |   |   Hu6800/M59499_at > 6.6907: T_KID (4.0/1.0)
|   |   Hu35KsubA/U15197_at > 8.59992
|   |   |   Hu35KsubA/RC_AA047876_at <= 5.7819
|   |   |   |   Hu6800/M99487_at <= 8.5783: NORMAL (20.0)
|   |   |   |   Hu6800/M99487_at > 8.5783: T_PROST (4.0)
|   |   |   Hu35KsubA/RC_AA047876_at > 5.7819: T_LUNG (5.0)
|   Hu35KsubA/AA055247_at > 8.5648: T_BRST (7.0/1.0)
Hu6800/X04741_at > 7.183: T_MESO (8.0)

Number of Leaves  :    13

Size of the tree :     25

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          34               38.2022 %
Incorrectly Classified Instances        55               61.7978 %
Kappa statistic                          0.3063
Mean absolute error                      0.101
Root mean squared error                  0.3123
Relative absolute error                 67.3948 %
Root relative squared error            113.7569 %
Total Number of Instances               89
```

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.667	0.059	0.778	0.667	0.718	0.806	NORMAL
0.286	0.073	0.25	0.286	0.267	0.606	T_COLON
0.375	0.099	0.273	0.375	0.316	0.631	T_PAN
0	0.012	0	0	0	0.488	T_KID
0	0.036	0	0	0	0.482	T_BLDR
0.5	0.036	0.5	0.5	0.5	0.808	T_PROST
0.2	0.048	0.2	0.2	0.2	0.576	T_OVARY
0.4	0.114	0.308	0.4	0.348	0.737	T_UT
0	0.036	0	0	0	0.476	T_LUNG
0.875	0.012	0.875	0.875	0.875	0.931	T_MESO
0	0.105	0	0	0	0.593	T_MELA
0	0.048	0	0	0	0.476	T_BRST
Weighted Avg. 0.382	0.059	0.386	0.382	0.381	0.681	

=== Confusion Matrix ===

a	b	c	d	e	f	g	h	i	j	k	l	<-- classified as
14	0	0	0	1	2	0	1	2	0	0	1	a = NORMAL
1	2	3	0	0	0	0	0	0	0	0	1	b = T_COLON
0	3	3	0	0	0	0	1	0	0	1	0	c = T_PAN
0	0	0	0	2	0	1	0	0	0	0	1	d = T_KID
0	1	3	0	0	0	1	1	0	0	0	0	e = T_BLDR
1	0	0	0	0	3	0	2	0	0	0	0	f = T_PROST
0	0	0	1	0	0	1	0	0	0	3	0	g = T_OVARY
0	2	0	0	0	0	0	4	0	0	4	0	h = T_UT
2	0	0	0	0	0	0	0	0	1	1	1	i = T_LUNG
0	0	1	0	0	0	0	0	0	7	0	0	j = T_MESO
0	0	1	0	0	0	1	1	0	0	0	0	k = T_MELA
0	0	0	0	0	1	1	3	1	0	0	0	l = T_BRST

Ek 1.2.6 SVM öznitelik seçimi uygulanarak mRNA bilgisi ve karar ağacı sınıflandırıcı ile çok kategorili kanser sınıflandırması sonuçları

```
Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     miRNA_mRNA_tek_normal_SVM_Select
Instances:    89
Attributes:   101
Test mode:    89-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

Hu6800/X04741_at <= 7.183
|   Hu35KsubA/RC_AA164851_at <= 8.63115
|   |   Hu35KsubA/K00627_at <= 6.612
|   |   |   Hu35KsubA/RC_D59675_i_at <= 6.8013
|   |   |   |   Hu6800/M59488_at <= 6.5612
|   |   |   |   |   Hu6800/L23808_at <= 6.41859
|   |   |   |   |   |   Hu35KsubA/K00627_at <= 5.2624: T_PAN (6.0)
|   |   |   |   |   |   Hu35KsubA/K00627_at > 5.2624: T_KID (3.0/1.0)
|   |   |   |   |   Hu6800/L23808_at > 6.41859: T_COLON (6.0)
|   |   |   |   Hu6800/M59488_at > 6.5612: T_MELA (2.0)
|   |   |   Hu35KsubA/RC_D59675_i_at > 6.8013
|   |   |   |   Hu35KsubA/AA253232_at <= 7.2188
|   |   |   |   |   Hu6800/HG3431-HT3616_s_at <= 6.29786
|   |   |   |   |   |   Hu6800/D87258_at <= 7.50424: T_BLDR (3.0)
|   |   |   |   |   |   Hu6800/D87258_at > 7.50424: T_KID (2.0)
|   |   |   |   |   Hu6800/HG3431-HT3616_s_at > 6.29786: T_UT (2.0)
|   |   |   |   Hu35KsubA/AA253232_at > 7.2188
|   |   |   |   |   Hu6800/L23808_at <= 6.61361
|   |   |   |   |   |   Hu35KsubA/AA292809_at <= 7.0408
|   |   |   |   |   |   |   Hu35KsubA/RC_AA235803_f_at <= 9.74128: NORMAL
(22.0/1.0)
|   |   |   |   |   |   |   Hu35KsubA/RC_AA235803_f_at > 9.74128: T_PROST
(4.0)
|   |   |   |   |   |   Hu35KsubA/AA292809_at > 7.0408: T_LUNG (2.0/1.0)
|   |   |   |   |   Hu6800/L23808_at > 6.61361
|   |   |   |   |   |   Hu6800/HG1496-HT1496_s_at <= 5: T_LUNG (4.0)
|   |   |   |   |   |   Hu6800/HG1496-HT1496_s_at > 5: T_PAN (2.0)
|   |   Hu35KsubA/K00627_at > 6.612
|   |   |   Hu6800/X51698_s_at <= 6.6002: T_BRST (6.0)
|   |   |   Hu6800/X51698_s_at > 6.6002: T_BLDR (3.0/1.0)
|   Hu35KsubA/RC_AA164851_at > 8.63115
|   |   Hu6800/X58079_at <= 7.0961: T_UT (9.0/1.0)
|   |   Hu6800/X58079_at > 7.0961: T_OVARY (5.0)
Hu6800/X04741_at > 7.183: T_MESO (8.0)

Number of Leaves  :    17

Size of the tree :     33
```

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          37               41.573  %
Kappa statistic                          0.3401
Mean absolute error                      0.0981
Root mean squared error                  0.296
Relative absolute error                 65.5171 %
Root relative squared error            107.8146 %
Coverage of cases (0.95 level)          50.5618 %
Mean rel. region size (0.95 level)      11.236  %
Total Number of Instances               89

=== Detailed Accuracy By Class ===

TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
0,524    0,132    0,550      0,524    0,537      0,398    0,773     0,456     NORMAL
0,143    0,098    0,111      0,143    0,125      0,040    0,517     0,103     T_COLON
0,375    0,062    0,375      0,375    0,375      0,313    0,653     0,174     T_PAN
0,000    0,059    0,000      0,000    0,000      -0,053   0,588     0,075     T_KID
0,167    0,048    0,200      0,167    0,182      0,129    0,687     0,248     T_BLDR
0,167    0,108    0,100      0,167    0,125      0,046    0,541     0,088     T_PROST
0,600    0,024    0,600      0,600    0,600      0,576    0,792     0,472     T_OVARY
0,600    0,076    0,500      0,600    0,545      0,485    0,744     0,372     T_UT
0,000    0,024    0,000      0,000    0,000      -0,037   0,464     0,056     T_LUNG
1,000    0,012    0,889      1,000    0,941      0,937    0,994     0,889     T_MESO
0,000    0,000    0,000      0,000    0,000      0,000    0,655     0,134     T_MELA
0,500    0,012    0,750      0,500    0,600      0,591    0,742     0,367     T_BRST
W. Avg.    0,416    0,071    0,413      0,416    0,411      0,344    0,706     0,338

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l   <-- classified as
 11  2  2  0  0  3  0  2  1  0  0  0 |  a = NORMAL
  1  1  1  0  1  2  0  1  0  0  0  0 |  b = T_COLON
  2  1  3  2  0  0  0  0  0  0  0  0 |  c = T_PAN
  0  1  1  0  1  0  0  0  1  0  0  0 |  d = T_KID
  2  2  0  0  1  0  0  1  0  0  0  0 |  e = T_BLDR
  2  0  1  1  0  1  0  0  0  0  0  1 |  f = T_PROST
  0  0  0  0  0  0  3  2  0  0  0  0 |  g = T_OVARY
  1  0  0  0  1  0  2  6  0  0  0  0 |  h = T_UT
  1  1  0  0  0  2  0  0  0  1  0  0 |  i = T_LUNG
  0  0  0  0  0  0  0  0  0  8  0  0 |  j = T_MESO
  0  1  0  0  0  2  0  0  0  0  0  0 |  k = T_MELA
  0  0  0  2  1  0  0  0  0  0  0  3 |  l = T_BRST
```

Ek 1.2.7 Öznitelik seçimi olmadan mRNA bilgisi ve KNN sınıflandırıcı ile çok kategorili kanser sınıflandırması sonuçları

```
Scheme:weka.classifiers.lazy.IBk -K 1 -W 0 -A
"weka.core.neighboursearch.LinearNNSearch -A
\"weka.core.EuclideanDistance -R first-last\""
Relation:     mRNA_tek_normal
Instances:    89
Attributes:   16064
Test mode:89-fold cross-validation

=== Classifier model (full training set) ===

IB1 instance-based classifier
using 1 nearest neighbour(s) for classification
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          54               60.6742 %
Incorrectly Classified Instances        35               39.3258 %
Kappa statistic                          0.5603
Mean absolute error                      0.076
Root mean squared error                  0.2424
Relative absolute error                 50.7438 %
Root relative squared error             88.3065 %
Total Number of Instances               89

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.714     0          1         0.714     0.833      0.857    NORMAL
                 0.571     0.012      0.8       0.571     0.667      0.78     T_COLON
                 0.875     0.025      0.778     0.875     0.824      0.925    T_PAN
                 0.5       0.035      0.4       0.5       0.444      0.732    T_KID
                 0         0.036      0         0         0          0.482    T_BLDR
                 0.667     0.06       0.444     0.667     0.533      0.803    T_PROST
                 0         0.083      0         0         0          0.458    T_OVARY
                 0.8       0.063      0.615     0.8       0.696      0.868    T_UT
                 0.8       0.012      0.8       0.8       0.8        0.894    T_LUNG
                 0.875     0.086      0.5       0.875     0.636      0.894    T_MESO
                 0.333     0          1         0.333     0.5        0.667    T_MELA
                 0.333     0.012      0.667     0.333     0.444      0.661    T_BRST
Weighted Avg.    0.607     0.032      0.654     0.607     0.606      0.787

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l   <-- classified as
 15  0  1  0  0  2  0  0  1  2  0  0 |  a = NORMAL
  0  4  1  0  0  1  0  1  0  0  0  0 |  b = T_COLON
  0  0  7  0  1  0  0  0  0  0  0  0 |  c = T_PAN
  0  0  0  2  0  1  0  0  0  1  0  0 |  d = T_KID
  0  0  0  0  0  0  5  1  0  0  0  0 |  e = T_BLDR
  0  0  0  0  0  4  0  1  0  1  0  0 |  f = T_PROST
  0  1  0  2  1  0  0  0  0  1  0  0 |  g = T_OVARY
  0  0  0  0  0  1  0  8  0  1  0  0 |  h = T_UT
  0  0  0  0  0  0  0  0  4  0  0  1 |  i = T_LUNG
  0  0  0  0  0  0  1  0  0  7  0  0 |  j = T_MESO
  0  0  0  0  0  0  0  1  0  1  1  0 |  k = T_MELA
  0  0  0  1  1  0  1  1  0  0  0  2 |  l = T_BRST
```

Ek 1.2.8 SVM öznitelik seçimi uygulanarak mRNA bilgisi ve KNN sınıflandırıcı ile çok kategorili kanser sınıflandırması sonuçları

```
Scheme:       weka.classifiers.lazy.IBk -K 1 -W 0 -A
"weka.core.neighboursearch.LinearNNSearch -A
\"weka.core.EuclideanDistance -R first-last\""
Relation:     miRNA_mRNA_tek_normal_SVM_Select
Instances:    89
Attributes:   101
Test mode:    89-fold cross-validation

=== Classifier model (full training set) ===
IB1 instance-based classifier
using 1 nearest neighbour(s) for classification

=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances          79               88.764  %
Kappa statistic                          0.8724
Mean absolute error                      0.0348
Root mean squared error                  0.1326
Relative absolute error                 23.2404 %
Root relative squared error             48.2934 %
Coverage of cases (0.95 level)          93.2584 %
Mean rel. region size (0.95 level)      58.3333 %
Total Number of Instances               89

=== Detailed Accuracy By Class ===

TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
1,000    0,029    0,913      1,000    0,955      0,941    0,985     0,913     NORMAL
0,857    0,000    1,000      0,857    0,923      0,920    0,929     0,868     T_COLON
0,875    0,025    0,778      0,875    0,824      0,807    0,925     0,692     T_PAN
0,750    0,000    1,000      0,750    0,857      0,861    0,875     0,761     T_KID
0,500    0,024    0,600      0,500    0,545      0,518    0,738     0,334     T_BLDR
0,833    0,000    1,000      0,833    0,909      0,907    0,917     0,845     T_PROST
0,800    0,000    1,000      0,800    0,889      0,889    0,900     0,811     T_OVARY
1,000    0,038    0,769      1,000    0,870      0,860    0,981     0,769     T_UT
0,800    0,012    0,800      0,800    0,800      0,788    0,894     0,651     T_LUNG
1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     T_MESO
1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     T_MELA
0,833    0,000    1,000      0,833    0,909      0,907    0,917     0,845     T_BRST
W. Avg.   0,888    0,016    0,895      0,888    0,886      0,877    0,936     0,809

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l   <-- classified as
 21  0  0  0  0  0  0  0  0  0  0  0 |  a = NORMAL
  0  6  1  0  0  0  0  0  0  0  0  0 |  b = T_COLON
  0  0  7  0  1  0  0  0  0  0  0  0 |  c = T_PAN
  0  0  0  3  0  0  0  1  0  0  0  0 |  d = T_KID
  1  0  1  0  3  0  0  1  0  0  0  0 |  e = T_BLDR
  0  0  0  0  0  5  0  0  1  0  0  0 |  f = T_PROST
  0  0  0  0  0  0  4  1  0  0  0  0 |  g = T_OVARY
  0  0  0  0  0  0  0 10  0  0  0  0 |  h = T_UT
  1  0  0  0  0  0  0  0  4  0  0  0 |  i = T_LUNG
  0  0  0  0  0  0  0  0  0  8  0  0 |  j = T_MESO
  0  0  0  0  0  0  0  0  0  0  3  0 |  k = T_MELA
  0  0  0  0  1  0  0  0  0  0  0  5 |  l = T_BRST
```

Ek 1.2.9 Öznitelik seçimi olmadan mRNA bilgisi ve NBM sınıflandırıcı ile çok kategorili kanser sınıflandırması sonuçları

```
Scheme:weka.classifiers.bayes.NaiveBayesMultinomial
Relation:     mRNA_tek_normal
Instances:    89
Attributes:   16064
Test mode:89-fold cross-validation


=== Classifier model (full training set) ===


The independent probability of a class

--------------------------------------

NORMAL      0.21782178217821782
T_COLON     0.07920792079207921
T_PAN       0.0891089108910891
T_KID       0.04950495049504951
T_BLDR      0.06930693069306931
T_PROST     0.06930693069306931
T_OVARY     0.0594059405940594
T_UT        0.10891089108910891
T_LUNG      0.0594059405940594
T_MESO      0.0891089108910891
T_MELA      0.039603960396039604
T_BRST      0.06930693069306931



=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          49               55.0562 %
Incorrectly Classified Instances        40               44.9438 %
Kappa statistic                          0.4964
Mean absolute error                      0.0753
Root mean squared error                  0.2724
Relative absolute error                 50.3007 %
Root relative squared error             99.224  %
Total Number of Instances               89
```

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.619	0.044	0.813	0.619	0.703	0.849	NORMAL
	0.571	0.061	0.444	0.571	0.5	0.837	T_COLON
	0.75	0.012	0.857	0.75	0.8	0.876	T_PAN
	0	0	0	0	0	0.838	T_KID
	0.333	0.036	0.4	0.333	0.364	0.729	T_BLDR
	0.667	0.06	0.444	0.667	0.533	0.833	T_PROST
	0.4	0.131	0.154	0.4	0.222	0.721	T_OVARY
	0.8	0.063	0.615	0.8	0.696	0.876	T_UT
	0.6	0.012	0.75	0.6	0.667	0.837	T_LUNG
	0.875	0.037	0.7	0.875	0.778	0.906	T_MESO
	0	0	0	0	0	0.899	T_MELA
	0	0.036	0	0	0	0.863	T_BRST
Weighted Avg.	0.551	0.044	0.543	0.551	0.536	0.844	

=== Confusion Matrix ===

a	b	c	d	e	f	g	h	i	j	k	l		<-- classified as
13	0	0	0	1	5	1	0	1	0	0	0	|	a = NORMAL
1	4	1	0	0	0	0	0	0	0	0	1	|	b = T_COLON
0	1	6	0	0	0	1	0	0	0	0	0	|	c = T_PAN
1	0	0	0	0	0	2	0	0	0	0	1	|	d = T_KID
0	0	0	0	2	0	2	2	0	0	0	0	|	e = T_BLDR
0	0	0	0	0	4	0	1	0	1	0	0	|	f = T_PROST
0	0	0	0	1	0	2	1	0	1	0	0	|	g = T_OVARY
0	1	0	0	0	0	1	8	0	0	0	0	|	h = T_UT
1	0	0	0	0	0	1	0	3	0	0	0	|	i = T_LUNG
0	0	0	0	0	0	1	0	0	7	0	0	|	j = T_MESO
0	0	0	0	0	0	0	1	0	1	0	1	|	k = T_MELA
0	3	0	0	1	0	2	0	0	0	0	0	|	l = T_BRST

Ek 1.2.10 SVM öznitelik seçimi uygulanarak mRNA bilgisi ve NBM sınıflandırıcı ile çok kategorili kanser sınıflandırması sonuçları

```
Scheme:       weka.classifiers.bayes.NaiveBayesMultinomial
Relation:     miRNA_mRNA_tek_normal_SVM_Select
Instances:    89
Attributes:   101
Test mode:    89-fold cross-validation

=== Classifier model (full training set) ===

The independent probability of a class
--------------------------------------
NORMAL      0.21782178217821782
T_COLON     0.07920792079207921
T_PAN       0.0891089108910891
T_KID       0.04950495049504951
T_BLDR      0.06930693069306931
T_PROST     0.06930693069306931
T_OVARY     0.0594059405940594
T_UT        0.10891089108910891
T_LUNG      0.0594059405940594
T_MESO      0.0891089108910891
T_MELA      0.039603960396039604
T_BRST      0.06930693069306931


=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          76               85.3933 %
Kappa statistic                          0.8359
Mean absolute error                      0.0306
Root mean squared error                  0.1296
Relative absolute error                 20.4229 %
Root relative squared error             47.1932 %
Coverage of cases (0.95 level)          96.6292 %
Mean rel. region size (0.95 level)      14.7004 %
Total Number of Instances               89
```

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0,857	0,015	0,947	0,857	0,900	0,873	0,994	0,983	NORMAL
1,000	0,000	1,000	1,000	1,000	1,000	1,000	1,000	T_COLON
0,875	0,000	1,000	0,875	0,933	0,930	0,881	0,887	T_PAN
0,750	0,000	1,000	0,750	0,857	0,861	0,997	0,950	T_KID
0,667	0,012	0,800	0,667	0,727	0,713	0,962	0,737	T_BLDR
0,833	0,024	0,714	0,833	0,769	0,754	0,978	0,868	T_PROST
0,800	0,060	0,444	0,800	0,571	0,566	0,952	0,777	T_OVARY
0,900	0,025	0,818	0,900	0,857	0,839	0,994	0,967	T_UT
0,800	0,012	0,800	0,800	0,800	0,788	0,995	0,943	T_LUNG
0,875	0,000	1,000	0,875	0,933	0,930	1,000	1,000	T_MESO
1,000	0,000	1,000	1,000	1,000	1,000	1,000	1,000	T_MELA
0,833	0,012	0,833	0,833	0,833	0,821	0,992	0,917	T_BRST
W. Avg.	0,854	0,014	0,881	0,854	0,861	0,849	0,980	0,932

=== Confusion Matrix ===

a	b	c	d	e	f	g	h	i	j	k	l	<-- classified as
18	0	0	0	0	2	0	0	1	0	0	0	a = NORMAL
0	7	0	0	0	0	0	0	0	0	0	0	b = T_COLON
0	0	7	0	0	0	1	0	0	0	0	0	c = T_PAN
0	0	0	3	0	0	1	0	0	0	0	0	d = T_KID
0	0	0	0	4	0	1	1	0	0	0	0	e = T_BLDR
1	0	0	0	0	5	0	0	0	0	0	0	f = T_PROST
0	0	0	0	0	0	4	1	0	0	0	0	g = T_OVARY
0	0	0	0	0	0	1	9	0	0	0	0	h = T_UT
0	0	0	0	0	0	0	0	4	0	0	1	i = T_LUNG
0	0	0	0	0	0	1	0	0	7	0	0	j = T_MESO
0	0	0	0	0	0	0	0	0	0	3	0	k = T_MELA
0	0	0	0	1	0	0	0	0	0	0	5	l = T_BRST

Ek 1.3 mikroRNA + mRNA bilgisi ile çok kategorili kanser sınıflandırması sonuçları

Ek 1.3.1 Öznitelik seçimi olmadan mikroRNA + mRNA bilgisi ve ANN sınıflandırıcı ile çok kategorili kanser sınıflandırması sonuçları

```
Scheme:weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N
500 -V 0 -S 0 -E 20 -H 2
Relation:     miRNA_mRNA_tek_normal
Instances:    89
Attributes:   16281
Test mode:89-fold cross-validation

=== Classifier model (full training set) ===

Class NORMAL
    Input
    Node 0
Class T_COLON
    Input
    Node 1
Class T_PAN
    Input
    Node 2
Class T_KID
    Input
    Node 3
Class T_BLDR
    Input
    Node 4
Class T_PROST
    Input
    Node 5
Class T_OVARY
    Input
    Node 6
Class T_UT
    Input
    Node 7
Class T_LUNG
    Input
    Node 8
Class T_MESO
    Input
    Node 9
Class T_MELA
    Input
    Node 10
Class T_BRST
    Input
    Node 11
```

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          21               23.5955 %
Incorrectly Classified Instances        68               76.4045 %
Kappa statistic                          0
Mean absolute error                      0.1493
Root mean squared error                  0.2748
Relative absolute error                 99.6833 %
Root relative squared error            100.0928 %
Total Number of Instances               89

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         1          0.236     1         0.382      0.394    NORMAL
                 0         0          0         0         0          0.071    T_COLON
                 0         0          0         0         0          0.103    T_PAN
                 0         0          0         0         0          0.012    T_KID
                 0         0          0         0         0          0.102    T_BLDR
                 0         0          0         0         0          0.002    T_PROST
                 0         0          0         0         0          0.024    T_OVARY
                 0         0          0         0         0          0.058    T_UT
                 0         0          0         0         0          0        T_LUNG
                 0         0          0         0         0          0.151    T_MESO
                 0         0          0         0         0          0        T_MELA
                 0         0          0         0         0          0.046    T_BRST
Weighted Avg.    0.236     0.236      0.056     0.236     0.09       0.14

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l   <-- classified as
 21  0  0  0  0  0  0  0  0  0  0  0 |  a = NORMAL
  7  0  0  0  0  0  0  0  0  0  0  0 |  b = T_COLON
  8  0  0  0  0  0  0  0  0  0  0  0 |  c = T_PAN
  4  0  0  0  0  0  0  0  0  0  0  0 |  d = T_KID
  6  0  0  0  0  0  0  0  0  0  0  0 |  e = T_BLDR
  6  0  0  0  0  0  0  0  0  0  0  0 |  f = T_PROST
  5  0  0  0  0  0  0  0  0  0  0  0 |  g = T_OVARY
 10  0  0  0  0  0  0  0  0  0  0  0 |  h = T_UT
  5  0  0  0  0  0  0  0  0  0  0  0 |  i = T_LUNG
  8  0  0  0  0  0  0  0  0  0  0  0 |  j = T_MESO
  3  0  0  0  0  0  0  0  0  0  0  0 |  k = T_MELA
  6  0  0  0  0  0  0  0  0  0  0  0 |  l = T_BRST
```

Ek 1.3.2 SVM öznitelik seçimi uygulanarak mikroRNA + mRNA bilgisi ve ANN sınıflandırıcı ile çok kategorili kanser sınıflandırması sonuçları

```
Scheme:       weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M
0.2 -N 500 -V 0 -S 0 -E 20 -H a
Relation:     miRNA_mRNA_tek_normal_SVM_Select
Instances:    89
Attributes:   101
Test mode:    89-fold cross-validation

=== Classifier model (full training set) ===

Class NORMAL
    Input
    Node 0
Class T_COLON
    Input
    Node 1
Class T_PAN
    Input
    Node 2
Class T_KID
    Input
    Node 3
Class T_BLDR
    Input
    Node 4
Class T_PROST
    Input
    Node 5
Class T_OVARY
    Input
    Node 6
Class T_UT
    Input
    Node 7
Class T_LUNG
    Input
    Node 8
Class T_MESO
    Input
    Node 9
Class T_MELA
    Input
    Node 10
Class T_BRST
    Input
    Node 11
```

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          84               94.382  %
Kappa statistic                          0.9364
Mean absolute error                      0.0208
Root mean squared error                  0.0839
Relative absolute error                 13.8728 %
Root relative squared error             30.5544 %
Coverage of cases (0.95 level)          98.8764 %
Mean rel. region size (0.95 level)      14.9813 %
Total Number of Instances               89

=== Detailed Accuracy By Class ===

TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
1,000    0,015    0,955      1,000    0,977      0,970    1,000     1,000     NORMAL
1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     T_COLON
0,875    0,000    1,000      0,875    0,933      0,930    1,000     1,000     T_PAN
1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     T_KID
0,667    0,012    0,800      0,667    0,727      0,713    0,994     0,915     T_BLDR
0,667    0,000    1,000      0,667    0,800      0,807    0,974     0,861     T_PROST
1,000    0,012    0,833      1,000    0,909      0,907    1,000     1,000     T_OVARY
1,000    0,013    0,909      1,000    0,952      0,947    0,999     0,991     T_UT
1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     T_LUNG
1,000    0,012    0,889      1,000    0,941      0,937    1,000     1,000     T_MESO
1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     T_MELA
1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     T_BRST
W. Avg.   0,944    0,007    0,946      0,944    0,941      0,937    0,998     0,984

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l   <-- classified as
 21  0  0  0  0  0  0  0  0  0  0  0 |  a = NORMAL
  0  7  0  0  0  0  0  0  0  0  0  0 |  b = T_COLON
  0  0  7  0  1  0  0  0  0  0  0  0 |  c = T_PAN
  0  0  0  4  0  0  0  0  0  0  0  0 |  d = T_KID
  1  0  0  0  4  0  1  0  0  0  0  0 |  e = T_BLDR
  0  0  0  0  0  4  0  1  0  1  0  0 |  f = T_PROST
  0  0  0  0  0  0  5  0  0  0  0  0 |  g = T_OVARY
  0  0  0  0  0  0  0 10  0  0  0  0 |  h = T_UT
  0  0  0  0  0  0  0  0  5  0  0  0 |  i = T_LUNG
  0  0  0  0  0  0  0  0  0  8  0  0 |  j = T_MESO
  0  0  0  0  0  0  0  0  0  0  3  0 |  k = T_MELA
  0  0  0  0  0  0  0  0  0  0  0  6 |  l = T_BRST
```

Ek 1.3.3 Öznitelik seçimi olmadan mikroRNA + mRNA bilgisi ve SVM sınıflandırıcı ile çok kategorili kanser sınıflandırması sonuçları

```
Scheme:weka.classifiers.functions.SMO -C 1.0 -L 0.0010 -P 1.0E-12 -N 2 -V
-1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -C 250007
-E 1.0"
Relation:     miRNA_mRNA_tek_normal
Instances:    89
Attributes:   16281
Test mode:89-fold cross-validation

=== Classifier model (full training set) ===
SMO
Kernel used:
  Linear Kernel: K(x,y) = <x,y>
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          69               77.5281 %
Incorrectly Classified Instances        20               22.4719 %
Kappa statistic                          0.7448
Mean absolute error                      0.1418
Root mean squared error                  0.261
Relative absolute error                 94.6709 %
Root relative squared error             95.0684 %
Total Number of Instances               89

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0.015      0.955     1         0.977      0.993    NORMAL
                 0.857     0.012      0.857     0.857     0.857      0.96     T_COLON
                 0.875     0.012      0.875     0.875     0.875      0.87     T_PAN
                 0.5       0          1         0.5       0.667      0.838    T_KID
                 0.5       0.012      0.75      0.5       0.6        0.766    T_BLDR
                 0.667     0          1         0.667     0.8        0.822    T_PROST
                 0.4       0.048      0.333     0.4       0.364      0.757    T_OVARY
                 0.9       0.063      0.643     0.9       0.75       0.872    T_UT
                 0.8       0          1         0.8       0.889      0.955    T_LUNG
                 0.875     0.012      0.875     0.875     0.875      0.991    T_MESO
                 0         0          0         0         0          0.897    T_MELA
                 0.667     0.072      0.4       0.667     0.5        0.906    T_BRST
Weighted Avg.    0.775     0.022      0.787     0.775     0.768      0.907

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l   <-- classified as
 21  0  0  0  0  0  0  0  0  0  0  0 |  a = NORMAL
  0  6  0  0  0  0  0  0  0  0  0  1 |  b = T_COLON
  0  0  7  0  0  0  0  0  0  0  0  1 |  c = T_PAN
  0  0  0  2  0  0  1  0  0  0  0  1 |  d = T_KID
  0  0  0  0  3  0  1  1  0  0  0  1 |  e = T_BLDR
  0  0  1  0  0  4  0  1  0  0  0  0 |  f = T_PROST
  0  0  0  0  1  0  2  2  0  0  0  0 |  g = T_OVARY
  0  0  0  0  0  0  1  9  0  0  0  0 |  h = T_UT
  0  0  0  0  0  0  0  0  4  0  0  1 |  i = T_LUNG
  0  0  0  0  0  0  1  0  0  7  0  0 |  j = T_MESO
  0  0  0  0  0  0  0  1  0  1  0  1 |  k = T_MELA
  1  1  0  0  0  0  0  0  0  0  0  4 |  l = T_BRST
```

Ek 1.3.4 SVM öznitelik seçimi uygulanarak mikroRNA + mRNA bilgisi ve SVM sınıflandırıcı ile çok kategorili kanser sınıflandırması sonuçları

```
Scheme:       weka.classifiers.functions.SMO -C 1.0 -L 0.001 -P 1.0E-12 -
N 2 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -C
250007 -E 1.0"
Relation:     miRNA_mRNA_tek_normal_SVM_Select
Instances:    89
Attributes:   101
Test mode:    89-fold cross-validation

=== Classifier model (full training set) ===
SMO
Kernel used:
  Linear Kernel: K(x,y) = <x,y>
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          83               93.2584 %
Kappa statistic                          0.9236
Mean absolute error                      0.1392
Root mean squared error                  0.2558
Relative absolute error                 92.9093 %
Root relative squared error             93.1576 %
Coverage of cases (0.95 level)         100      %
Mean rel. region size (0.95 level)      77.1536 %
Total Number of Instances               89

=== Detailed Accuracy By Class ===

TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
1,000    0,029    0,913      1,000    0,955      0,941    0,985     0,913     NORMAL
1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     T_COLON
0,875    0,000    1,000      0,875    0,933      0,930    0,948     0,897     T_PAN
0,750    0,012    0,750      0,750    0,750      0,738    0,987     0,674     T_KID
0,833    0,012    0,833      0,833    0,833      0,821    0,985     0,761     T_BLDR
0,667    0,000    1,000      0,667    0,800      0,807    0,978     0,784     T_PROST
1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     T_OVARY
0,900    0,013    0,900      0,900    0,900      0,887    0,988     0,863     T_UT
1,000    0,012    0,833      1,000    0,909      0,907    0,994     0,833     T_LUNG
1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     T_MESO
1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     T_MELA
1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     T_BRST
W. Avg.   0,933    0,010    0,936      0,933    0,931      0,925    0,987     0,900

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l   <-- classified as
 21  0  0  0  0  0  0  0  0  0  0  0 |  a = NORMAL
  0  7  0  0  0  0  0  0  0  0  0  0 |  b = T_COLON
  0  0  7  0  1  0  0  0  0  0  0  0 |  c = T_PAN
  0  0  0  3  0  0  0  1  0  0  0  0 |  d = T_KID
  1  0  0  0  5  0  0  0  0  0  0  0 |  e = T_BLDR
  1  0  0  0  0  4  0  0  1  0  0  0 |  f = T_PROST
  0  0  0  0  0  0  5  0  0  0  0  0 |  g = T_OVARY
  0  0  0  1  0  0  0  9  0  0  0  0 |  h = T_UT
  0  0  0  0  0  0  0  0  5  0  0  0 |  i = T_LUNG
  0  0  0  0  0  0  0  0  0  8  0  0 |  j = T_MESO
  0  0  0  0  0  0  0  0  0  0  3  0 |  k = T_MELA
  0  0  0  0  0  0  0  0  0  0  0  6 |  l = T_BRST
```

Ek 1.3.5 Öznitelik seçimi olmadan mikroRNA + mRNA bilgisi ve karar ağacı sınıflandırıcı ile çok kategorili kanser sınıflandırması sonuçları

```
Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     miRNA_mRNA_tek_normal
Instances:    89
Attributes:   16281
Test mode:89-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

EAM208 <= 5.24708
|   EAM225 <= 5.0387
|   |   Hu6800/D00003_s_at <= 5: T_MESO (8.0)
|   |   Hu6800/D00003_s_at > 5: T_KID (4.0)
|   EAM225 > 5.0387: T_MELA (3.0)
EAM208 > 5.24708
|   Hu35KsubA/AA055247_at <= 8.5648
|   |   Hu35KsubA/U15197_at <= 8.59992
|   |   |   EAM250 <= 8.30111
|   |   |   |   EAM276 <= 5.6665
|   |   |   |   |   Hu6800/U26174_at <= 5.0495
|   |   |   |   |   |   Hu6800/L25286_s_at <= 5: T_BLDR (6.0)
|   |   |   |   |   |   Hu6800/L25286_s_at > 5: T_OVARY (5.0)
|   |   |   |   |   Hu6800/U26174_at > 5.0495: T_PROST (3.0/1.0)
|   |   |   |   EAM276 > 5.6665: T_UT (10.0)
|   |   |   EAM250 > 8.30111
|   |   |   |   Hu6800/J00306_at <= 5.13684: T_COLON (7.0)
|   |   |   |   Hu6800/J00306_at > 5.13684: T_PAN (8.0)
|   |   Hu35KsubA/U15197_at > 8.59992
|   |   |   Hu35KsubA/RC_AA047876_at <= 5.7819
|   |   |   |   Hu6800/M99487_at <= 8.5783: NORMAL (20.0)
|   |   |   |   Hu6800/M99487_at > 8.5783: T_PROST (4.0)
|   |   |   Hu35KsubA/RC_AA047876_at > 5.7819: T_LUNG (5.0)
|   Hu35KsubA/AA055247_at > 8.5648: T_BRST (6.0)

Number of Leaves  :    13

Size of the tree :     25



=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          47               52.809  %
Incorrectly Classified Instances        42               47.191  %
Kappa statistic                          0.4723
Mean absolute error                      0.0787
Root mean squared error                  0.277
Relative absolute error                 52.5066 %
Root relative squared error            100.878  %
Total Number of Instances               89
```

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.571	0.074	0.706	0.571	0.632	0.745	NORMAL
	0.286	0.049	0.333	0.286	0.308	0.614	T_COLON
	0.5	0.049	0.5	0.5	0.5	0.728	T_PAN
	0	0.012	0	0	0	0.494	T_KID
	0.167	0.012	0.5	0.167	0.25	0.577	T_BLDR
	0.5	0.096	0.273	0.5	0.353	0.708	T_PROST
	0.6	0.119	0.231	0.6	0.333	0.74	T_OVARY
	0.6	0.038	0.667	0.6	0.632	0.778	T_UT
	0.4	0.036	0.4	0.4	0.4	0.682	T_LUNG
	0.875	0	1	0.875	0.933	0.938	T_MESO
	0.333	0	1	0.333	0.5	0.667	T_MELA
	1	0.036	0.667	1	0.8	0.982	T_BRST
Weighted Avg.	0.528	0.049	0.569	0.528	0.526	0.739	

=== Confusion Matrix ===

a	b	c	d	e	f	g	h	i	j	k	l		<-- classified as
12	0	0	0	0	3	1	1	3	0	0	1	|	a = NORMAL
1	2	3	0	0	0	1	0	0	0	0	0	|	b = T_COLON
0	3	4	0	1	0	0	0	0	0	0	0	|	c = T_PAN
0	0	0	0	0	1	2	0	0	0	0	1	|	d = T_KID
0	0	1	1	1	0	3	0	0	0	0	0	|	e = T_BLDR
1	0	0	0	0	3	1	1	0	0	0	0	|	f = T_PROST
0	0	0	0	0	2	3	0	0	0	0	0	|	g = T_OVARY
1	1	0	0	0	1	1	6	0	0	0	0	|	h = T_UT
2	0	0	0	0	0	0	0	2	0	0	1	|	i = T_LUNG
0	0	0	0	0	0	1	0	0	7	0	0	|	j = T_MESO
0	0	0	0	0	1	0	1	0	0	1	0	|	k = T_MELA
0	0	0	0	0	0	0	0	0	0	0	6	|	l = T_BRST

Ek 1.3.6 SVM öznitelik seçimi uygulanarak mikroRNA + mRNA bilgisi ve karar ağacı sınıflandırıcı ile çok kategorili kanser sınıflandırması sonuçları

```
Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     miRNA_mRNA_tek_normal-SVM_Select
Instances:    89
Attributes:   101
Test mode:    89-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

EAM208 <= 5.24708
|   Hu6800/X84707_rna1_at <= 7.14317
|   |   Hu6800/X66533_at <= 5.7535: T_MESO (8.0)
|   |   Hu6800/X66533_at > 5.7535: T_KID (4.0)
|   Hu6800/X84707_rna1_at > 7.14317: T_MELA (3.0)
EAM208 > 5.24708
|   Hu35KsubA/R82528_at <= 6.7609
|   |   Hu35KsubA/U15197_at <= 8.59992
|   |   |   Hu6800/X58079_at <= 7.0961
|   |   |   |   Hu6800/M18728_at <= 7.20469
|   |   |   |   |   Hu6800/X83618_at <= 5.296: T_UT (10.0)
|   |   |   |   |   Hu6800/X83618_at > 5.296: T_BLDR (7.0/1.0)
|   |   |   |   Hu6800/M18728_at > 7.20469
|   |   |   |   |   Hu6800/X51698_s_at <= 6.9169
|   |   |   |   |   |   Hu6800/M24461_at <= 6.10166: T_COLON (7.0)
|   |   |   |   |   |   Hu6800/M24461_at > 6.10166: T_PROST (3.0/1.0)
|   |   |   |   |   Hu6800/X51698_s_at > 6.9169: T_PAN (7.0)
|   |   |   Hu6800/X58079_at > 7.0961: T_OVARY (5.0)
|   |   Hu35KsubA/U15197_at > 8.59992
|   |   |   EAM159 <= 8.23445
|   |   |   |   Hu6800/M18728_at <= 5.03788: T_PROST (4.0)
|   |   |   |   Hu6800/M18728_at > 5.03788: T_LUNG (4.0)
|   |   |   EAM159 > 8.23445: NORMAL (20.0)
|   Hu35KsubA/R82528_at > 6.7609: T_BRST (7.0/1.0)

Number of Leaves  :    13

Size of the tree :     25


=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          63               70.7865 %
Kappa statistic                          0.6715
Mean absolute error                      0.0478
Root mean squared error                  0.2102
Relative absolute error                 31.879  %
Root relative squared error             76.5479 %
Coverage of cases (0.95 level)          73.0337 %
Mean rel. region size (0.95 level)       9.3633 %
Total Number of Instances               89
```

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0,857	0,000	1,000	0,857	0,923	0,906	0,922	0,891	NORMAL
0,429	0,049	0,429	0,429	0,429	0,380	0,689	0,259	T_COLON
0,625	0,049	0,556	0,625	0,588	0,546	0,797	0,555	T_PAN
0,750	0,000	1,000	0,750	0,857	0,861	0,875	0,761	T_KID
0,500	0,024	0,600	0,500	0,545	0,518	0,732	0,334	T_BLDR
0,667	0,060	0,444	0,667	0,533	0,504	0,811	0,403	T_PROST
0,800	0,024	0,667	0,800	0,727	0,713	0,888	0,545	T_OVARY
0,800	0,051	0,667	0,800	0,727	0,693	0,878	0,604	T_UT
0,600	0,024	0,600	0,600	0,600	0,576	0,788	0,382	T_LUNG
1,000	0,000	1,000	1,000	1,000	1,000	1,000	1,000	T_MESO
0,000	0,000	0,000	0,000	0,000	0,000	0,833	0,678	T_MELA
0,667	0,036	0,571	0,667	0,615	0,587	0,815	0,403	T_BRST
W. Avg.	0,708	0,025	0,709	0,708	0,703	0,680	0,853	0,624

=== Confusion Matrix ===

a	b	c	d	e	f	g	h	i	j	k	l	<-- classified as
18	0	0	0	0	1	1	0	1	0	0	0	a = NORMAL
0	3	1	0	1	1	0	0	1	0	0	0	b = T_COLON
0	1	5	0	0	0	0	2	0	0	0	0	c = T_PAN
0	0	0	3	0	0	1	0	0	0	0	0	d = T_KID
0	1	0	0	3	0	0	1	0	0	0	1	e = T_BLDR
0	2	0	0	0	4	0	0	0	0	0	0	f = T_PROST
0	0	0	0	1	0	4	0	0	0	0	0	g = T_OVARY
0	0	1	0	0	1	0	8	0	0	0	0	h = T_UT
0	0	0	0	0	1	0	0	3	0	0	1	i = T_LUNG
0	0	0	0	0	0	0	0	0	8	0	0	j = T_MESO
0	0	2	0	0	0	0	0	0	0	0	1	k = T_MELA
0	0	0	0	0	1	0	1	0	0	0	4	l = T_BRST

Ek 1.3.7 Öznitelik seçimi olmadan mikroRNA + mRNA bilgisi ve KNN sınıflandırıcı ile çok kategorili kanser sınıflandırması sonuçları

```
Scheme:weka.classifiers.lazy.IBk -K 1 -W 0 -A
"weka.core.neighboursearch.LinearNNSearch -A
\"weka.core.EuclideanDistance -R first-last\""
Relation:     miRNA_mRNA_tek_normal
Instances:    89
Attributes:   16281
Test mode:89-fold cross-validation

=== Classifier model (full training set) ===

IB1 instance-based classifier
using 1 nearest neighbour(s) for classification

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          54               60.6742 %
Incorrectly Classified Instances        35               39.3258 %
Kappa statistic                          0.5596
Mean absolute error                      0.076
Root mean squared error                  0.2424
Relative absolute error                 50.7438 %
Root relative squared error             88.3065 %
Total Number of Instances               89

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
               0.762     0          1         0.762     0.865      0.881    NORMAL
               0.429     0          1         0.429     0.6        0.714    T_COLON
               0.875     0.037      0.7       0.875     0.778      0.919    T_PAN
               0.5       0.047      0.333     0.5       0.4        0.726    T_KID
               0         0.036      0         0         0          0.482    T_BLDR
               0.667     0.06       0.444     0.667     0.533      0.803    T_PROST
               0         0.083      0         0         0          0.458    T_OVARY
               0.8       0.063      0.615     0.8       0.696      0.868    T_UT
               0.8       0.012      0.8       0.8       0.8        0.894    T_LUNG
               0.875     0.074      0.538     0.875     0.667      0.9      T_MESO
               0.333     0          1         0.333     0.5        0.667    T_MELA
               0.333     0.012      0.667     0.333     0.444      0.661    T_BRST
Weighted Avg.  0.607     0.032      0.664     0.607     0.605      0.787

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l   <-- classified as
 16  0  1  0  0  1  0  0  1  2  0  0 |  a = NORMAL
  0  3  2  0  0  1  0  1  0  0  0  0 |  b = T_COLON
  0  0  7  0  1  0  0  0  0  0  0  0 |  c = T_PAN
  0  0  0  2  0  1  0  0  0  1  0  0 |  d = T_KID
  0  0  0  0  0  0  5  1  0  0  0  0 |  e = T_BLDR
  0  0  0  0  0  4  0  1  0  1  0  0 |  f = T_PROST
  0  0  0  2  1  1  0  0  0  1  0  0 |  g = T_OVARY
  0  0  0  1  0  1  0  8  0  0  0  0 |  h = T_UT
  0  0  0  0  0  0  0  0  4  0  0  1 |  i = T_LUNG
  0  0  0  0  0  0  1  0  0  7  0  0 |  j = T_MESO
  0  0  0  0  0  0  0  1  0  1  1  0 |  k = T_MELA
  0  0  0  1  1  0  1  1  0  0  0  2 |  l = T_BRST
```

Ek 1.3.8 SVM öznitelik seçimi uygulanarak mikroRNA + mRNA bilgisi ve KNN sınıflandırıcı ile çok kategorili kanser sınıflandırması sonuçları

```
Scheme:       weka.classifiers.lazy.IBk -K 1 -W 0 -A
"weka.core.neighboursearch.LinearNNSearch -A
\"weka.core.EuclideanDistance -R first-last\""
Relation:     miRNA_mRNA_tek_normal_SVM_Select
Instances:    89
Attributes:   101
Test mode:    89-fold cross-validation

=== Classifier model (full training set) ===
IB1 instance-based classifier
using 1 nearest neighbour(s) for classification

=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances          82               92.1348 %
Kappa statistic                          0.9108
Mean absolute error                      0.0299
Root mean squared error                  0.1124
Relative absolute error                 19.94   %
Root relative squared error             40.9433 %
Coverage of cases (0.95 level)          93.2584 %
Mean rel. region size (0.95 level)      58.3333 %
Total Number of Instances               89

=== Detailed Accuracy By Class ===

TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
1,000    0,029    0,913      1,000    0,955      0,941    0,985     0,913     NORMAL
1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     T_COLON
0,875    0,012    0,875      0,875    0,875      0,863    0,931     0,777     T_PAN
 0,750    0,000    1,000      0,750    0,857      0,861    0,875     0,761     T_KID
0,500    0,000    1,000      0,500    0,667      0,695    0,750     0,534     T_BLDR
0,667    0,000    1,000      0,667    0,800      0,807    0,833     0,689     T_PROST
1,000    0,024    0,714      1,000    0,833      0,835    0,988     0,714     T_OVARY
1,000    0,013    0,909      1,000    0,952      0,947    0,994     0,909     T_UT
1,000    0,012    0,833      1,000    0,909      0,907    0,994     0,833     T_LUNG
1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     T_MESO
1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     T_MELA
1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     T_BRST
W. Avg.   0,921    0,011    0,933      0,921    0,916      0,914    0,955     0,861

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k  l   <-- classified as
 21  0  0  0  0  0  0  0  0  0  0  0 |  a = NORMAL
  0  7  0  0  0  0  0  0  0  0  0  0 |  b = T_COLON
  0  0  7  0  0  0  1  0  0  0  0  0 |  c = T_PAN
  0  0  0  3  0  0  0  1  0  0  0  0 |  d = T_KID
  1  0  1  0  3  0  1  0  0  0  0  0 |  e = T_BLDR
  1  0  0  0  0  4  0  0  1  0  0  0 |  f = T_PROST
  0  0  0  0  0  0  5  0  0  0  0  0 |  g = T_OVARY
  0  0  0  0  0  0  0 10  0  0  0  0 |  h = T_UT
  0  0  0  0  0  0  0  0  5  0  0  0 |  i = T_LUNG
  0  0  0  0  0  0  0  0  0  8  0  0 |  j = T_MESO
  0  0  0  0  0  0  0  0  0  0  3  0 |  k = T_MELA
  0  0  0  0  0  0  0  0  0  0  0  6 |  l = T_BRST
```

Ek 1.3.9 Öznitelik seçimi olmadan mikroRNA + mRNA bilgisi ve NBM sınıflandırıcı ile çok kategorili kanser sınıflandırması sonuçları

```
Scheme:weka.classifiers.bayes.NaiveBayesMultinomial
Relation:     miRNA_mRNA_tek_normal
Instances:    89
Attributes:   16281
Test mode:89-fold cross-validation

=== Classifier model (full training set) ===

The independent probability of a class
--------------------------------------
NORMAL      0.21782178217821782
T_COLON     0.07920792079207921
T_PAN 0.0891089108910891
T_KID 0.04950495049504951
T_BLDR      0.06930693069306931
T_PROST     0.06930693069306931
T_OVARY     0.0594059405940594
T_UT  0.10891089108910891
T_LUNG      0.0594059405940594
T_MESO      0.0891089108910891
T_MELA      0.039603960396039604
T_BRST      0.06930693069306931


=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          51               57.3034 %
Incorrectly Classified Instances        38               42.6966 %
Kappa statistic                          0.5196
Mean absolute error                      0.0709
Root mean squared error                  0.2653
Relative absolute error                 47.3252 %
Root relative squared error             96.6445 %
Total Number of Instances               89
```

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.714	0.044	0.833	0.714	0.769	0.862	NORMAL
	0.571	0.073	0.4	0.571	0.471	0.85	T_COLON
	0.75	0.012	0.857	0.75	0.8	0.876	T_PAN
	0	0	0	0	0	0.847	T_KID
	0.333	0.036	0.4	0.333	0.364	0.729	T_BLDR
	0.667	0.036	0.571	0.667	0.615	0.835	T_PROST
	0.4	0.131	0.154	0.4	0.222	0.725	T_OVARY
	0.8	0.063	0.615	0.8	0.696	0.881	T_UT
	0.6	0.012	0.75	0.6	0.667	0.842	T_LUNG
	0.875	0.025	0.778	0.875	0.824	0.921	T_MESO
	0	0	0	0	0	0.911	T_MELA
	0	0.036	0	0	0	0.871	T_BRST
Weighted Avg.	0.573	0.042	0.56	0.573	0.559	0.852	

=== Confusion Matrix ===

a	b	c	d	e	f	g	h	i	j	k	l		<-- classified as
15	0	0	0	1	3	1	0	1	0	0	0	|	a = NORMAL
1	4	1	0	0	0	0	0	0	0	0	1	|	b = T_COLON
0	1	6	0	0	0	1	0	0	0	0	0	|	c = T_PAN
1	0	0	0	0	0	2	0	0	0	0	1	|	d = T_KID
0	0	0	0	2	0	2	2	0	0	0	0	|	e = T_BLDR
0	0	0	0	0	4	0	1	0	1	0	0	|	f = T_PROST
0	1	0	0	1	0	2	1	0	0	0	0	|	g = T_OVARY
0	1	0	0	0	0	1	8	0	0	0	0	|	h = T_UT
1	0	0	0	0	0	1	0	3	0	0	0	|	i = T_LUNG
0	0	0	0	0	0	1	0	0	7	0	0	|	j = T_MESO
0	0	0	0	0	0	0	1	0	1	0	1	|	k = T_MELA
0	3	0	0	1	0	2	0	0	0	0	0	|	l = T_BRST

Ek 1.3.10 SVM öznitelik seçimi uygulanarak mikroRNA + mRNA bilgisi ve NBM sınıflandırıcı ile çok kategorili kanser sınıflandırması sonuçları

```
Scheme:       weka.classifiers.bayes.NaiveBayesMultinomial
Relation:     miRNA_mRNA_tek_normal_SVM_Select
Instances:    89
Attributes:   101
Test mode:    89-fold cross-validation

=== Classifier model (full training set) ===

The independent probability of a class
--------------------------------------
NORMAL     0.21782178217821782
T_COLON    0.07920792079207921
T_PAN      0.0891089108910891
T_KID      0.04950495049504951
T_BLDR     0.06930693069306931
T_PROST    0.06930693069306931
T_OVARY    0.0594059405940594
T_UT       0.10891089108910891
T_LUNG     0.0594059405940594
T_MESO     0.0891089108910891
T_MELA     0.039603960396039604
T_BRST     0.06930693069306931



=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          81               91.0112 %
Kappa statistic                          0.8984
Mean absolute error                      0.0198
Root mean squared error                  0.1088
Relative absolute error                 13.2161 %
Root relative squared error             39.6304 %
Coverage of cases (0.95 level)          97.7528 %
Mean rel. region size (0.95 level)      11.7978 %
Total Number of Instances               89
```

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0,952	0,029	0,909	0,952	0,930	0,908	0,997	0,992	NORMAL
1,000	0,000	1,000	1,000	1,000	1,000	1,000	1,000	T_COLON
0,875	0,000	1,000	0,875	0,933	0,930	0,978	0,920	T_PAN
0,750	0,000	1,000	0,750	0,857	0,861	0,997	0,950	T_KID
0,667	0,000	1,000	0,667	0,800	0,807	0,986	0,877	T_BLDR
0,667	0,000	1,000	0,667	0,800	0,807	0,962	0,847	T_PROST
0,800	0,036	0,571	0,800	0,667	0,654	0,988	0,871	T_OVARY
1,000	0,013	0,909	1,000	0,952	0,947	0,999	0,991	T_UT
1,000	0,024	0,714	1,000	0,833	0,835	0,998	0,967	T_LUNG
1,000	0,000	1,000	1,000	1,000	1,000	1,000	1,000	T_MESO
1,000	0,000	1,000	1,000	1,000	1,000	1,000	1,000	T_MELA
1,000	0,000	1,000	1,000	1,000	1,000	1,000	1,000	T_BRST
W. Avg.	0,910	0,012	0,928	0,910	0,911	0,905	0,993	0,960

=== Confusion Matrix ===

a	b	c	d	e	f	g	h	i	j	k	l	<-- classified as
20	0	0	0	0	0	0	0	1	0	0	0	a = NORMAL
0	7	0	0	0	0	0	0	0	0	0	0	b = T_COLON
0	0	7	0	0	0	1	0	0	0	0	0	c = T_PAN
0	0	0	3	0	0	1	0	0	0	0	0	d = T_KID
1	0	0	0	4	0	1	0	0	0	0	0	e = T_BLDR
1	0	0	0	0	4	0	0	1	0	0	0	f = T_PROST
0	0	0	0	0	0	4	1	0	0	0	0	g = T_OVARY
0	0	0	0	0	0	0	10	0	0	0	0	h = T_UT
0	0	0	0	0	0	0	0	5	0	0	0	i = T_LUNG
0	0	0	0	0	0	0	0	0	8	0	0	j = T_MESO
0	0	0	0	0	0	0	0	0	0	3	0	k = T_MELA
0	0	0	0	0	0	0	0	0	0	0	6	l = T_BRS

Ek 2 Göğüs kanseri alt kategori sınıflandırma sonuçları

Ek 2.1 mikroRNA bilgisi ile göğüs kanseri alt kategori sınıflandırma sonuçları

Ek 2.1.1 Öznitelik seçimi olmadan mikroRNA bilgisi ve ANN sınıflandırıcı ile göğüs kanseri alt kategori sınıflandırma sonuçları

```
Scheme:weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N
500 -V 0 -S 0 -E 20 -H a
Relation:     BreastCancerSubtypes_miRNA
Instances:    94
Attributes:   490
Test mode:94-fold cross-validation

=== Classifier model (full training set) ===

Class Lum A
    Input
    Node 0
Class basal
    Input
    Node 1
Class Normal-like
    Input
    Node 2
Class Lum B
    Input
    Node 3
Class ERBB2
    Input
    Node 4

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          72               76.5957 %
Incorrectly Classified Instances        22               23.4043 %
Kappa statistic                          0.6711
Mean absolute error                      0.1213
Root mean squared error                  0.2919
Relative absolute error                 41.0303 %
Root relative squared error             75.7368 %
Total Number of Instances               94
```

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.951	0.113	0.867	0.951	0.907	0.974	Lum A
	0.933	0.013	0.933	0.933	0.933	0.982	basal
	0.2	0.06	0.286	0.2	0.235	0.649	Normal-like
	0.333	0.024	0.667	0.333	0.444	0.88	Lum B
	0.813	0.103	0.619	0.813	0.703	0.898	ERBB2
Weighted Avg.	0.766	0.078	0.748	0.766	0.746	0.916	

=== Confusion Matrix ===

a	b	c	d	e	<-- classified as
39	0	0	1	1	a = Lum A
0	14	1	0	0	b = basal
4	0	2	1	3	c = Normal-like
2	1	1	4	4	d = Lum B
0	0	3	0	13	e = ERBB2

Ek 2.1.2 SVM öznitelik seçimi uygulanarak mikroRNA bilgisi ve ANN sınıflandırıcı ile göğüs kanseri alt kategori sınıflandırma sonuçları

```
Scheme:       weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M
0.2 -N 500 -V 0 -S 0 -E 20 -H a
Relation:     BreastCancerSubtypes_miRNA_SVM_Select
Instances:    94
Attributes:   101
Test mode:    94-fold cross-validation

=== Classifier model (full training set) ===

Class Lum A
    Input
    Node 0
Class basal
    Input
    Node 1
Class Normal-like
    Input
    Node 2
Class Lum B
    Input
    Node 3
Class ERBB2
    Input
    Node 4

=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances          83               88.2979 %
Kappa statistic                          0.8365
Mean absolute error                      0.0556
Root mean squared error                  0.1605
Relative absolute error                 18.8131 %
Root relative squared error             41.6357 %
Coverage of cases (0.95 level)          98.9362 %
Mean rel. region size (0.95 level)      30.4255 %
Total Number of Instances               94

=== Detailed Accuracy By Class ===

TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
0,976    0,075    0,909      0,976    0,941      0,895    0,997     0,996     Lum A
0,933    0,013    0,933      0,933    0,933      0,921    0,999     0,996     basal
0,500    0,012    0,833      0,500    0,625      0,616    0,970     0,837     Normal-like
0,667    0,024    0,800      0,667    0,727      0,695    0,988     0,936     Lum B
1,000    0,038    0,842      1,000    0,914      0,900    0,994     0,971     ERBB2
W. Avg.0,883    0,046    0,880      0,883    0,874      0,845    0,993     0,967

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 40  0  0  1  0 |  a = Lum A
  0 14  1  0  0 |  b = basal
  2  1  5  1  1 |  c = Normal-like
  2  0  0  8  2 |  d = Lum B
  0  0  0  0 16 |  e = ERBB2
```

Ek 2.1.3 Öznitelik seçimi olmadan mikroRNA bilgisi ve SVM sınıflandırıcı ile göğüs kanseri alt kategori sınıflandırma sonuçları

```
Scheme:       weka.classifiers.functions.SMO -C 1.0 -L 0.001 -P 1.0E-12 -
N 2 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -C
250007 -E 1.0"
Relation:     BreastCancerSubtypes_miRNA
Instances:    94
Attributes:   490
Test mode:    94-fold cross-validation

=== Classifier model (full training set) ===

SMO

Kernel used:
  Linear Kernel: K(x,y) = <x,y>

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          64               68.0851 %
Kappa statistic                          0.5574
Mean absolute error                      0.2613
Root mean squared error                  0.347
Relative absolute error                 88.4102 %
Root relative squared error             90.0173 %
Coverage of cases (0.95 level)          100     %
Mean rel. region size (0.95 level)      80.6383 %
Total Number of Instances               94

=== Detailed Accuracy By Class ===

TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
0,854    0,151    0,814      0,854    0,833      0,700    0,910     0,813     Lum A
0,933    0,025    0,875      0,933    0,903      0,885    0,981     0,844     basal
0,100    0,095    0,111      0,100    0,105      0,005    0,488     0,106     Normal-like
0,500    0,061    0,545      0,500    0,522      0,456    0,767     0,365     Lum B
0,500    0,090    0,533      0,500    0,516      0,421    0,841     0,456     ERBB2
W.Avg.0,681    0,103    0,667      0,681    0,673      0,577    0,846     0,625

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 35  0  3  1  2 |  a = Lum A
  0 14  1  0  0 |  b = basal
  4  1  1  1  3 |  c = Normal-like
  2  1  1  6  2 |  d = Lum B
  2  0  3  3  8 |  e = ERBB2
```

Ek 2.1.4 SVM öznitelik seçimi uygulanarak mikroRNA bilgisi ve SVM sınıflandırıcı ile göğüs kanseri alt kategori sınıflandırma sonuçları

```
Scheme:       weka.classifiers.functions.SMO -C 1.0 -L 0.001 -P 1.0E-12 -
N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -C
250007 -E 1.0"
Relation:     BreastCancerSubtypes_miRNA_SVM_Select
Instances:    94
Attributes:   101
Test mode:    94-fold cross-validation

=== Classifier model (full training set) ===

SMO

Kernel used:
  Linear Kernel: K(x,y) = <x,y>

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          82               87.234  %
Kappa statistic                         0.8208
Mean absolute error                     0.2472
Root mean squared error                 0.3267
Relative absolute error                 83.6585 %
Root relative squared error             84.7565 %
Coverage of cases (0.95 level)          100     %
Mean rel. region size (0.95 level)      80.8511 %
Total Number of Instances               94

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall   F-Measure   MCC      ROC Area  PRC Area
Class
               0,976     0,094     0,889       0,976    0,930       0,875    0,978     0,948
Lum A
               1,000     0,013     0,938       1,000    0,968       0,962    1,000     1,000
basal
               0,500     0,012     0,833       0,500    0,625       0,616    0,750     0,568
Normal-like
               0,667     0,012     0,889       0,667    0,762       0,742    0,939     0,695
Lum B
               0,875     0,051     0,778       0,875    0,824       0,787    0,939     0,722
ERBB2
Weighted Avg.  0,872     0,055     0,872       0,872    0,864       0,829    0,946     0,845

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 40  0  0  1  0 |  a = Lum A
  0 15  0  0  0 |  b = basal
  2  1  5  0  2 |  c = Normal-like
  1  0  1  8  2 |  d = Lum B
  2  0  0  0 14 |  e = ERBB2
```

Ek 2.1.5 Öznitelik seçimi olmadan mikroRNA bilgisi ve karar ağacı sınıflandırıcı ile göğüs kanseri alt kategori sınıflandırma sonuçları

```
Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     BreastCancerSubtypes_miRNA
Instances:    94
Attributes:   490
Test mode:    94-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------


hsa-miR-29c* <= 6.047352
|   hsa-miR-522 <= 1.602398
|   |   ebv-miR-BART16 <= 3.73585: Lum B (2.0)
|   |   ebv-miR-BART16 > 3.73585: ERBB2 (2.0)
|   hsa-miR-522 > 1.602398: basal (14.0)
hsa-miR-29c* > 6.047352
|   hsa-miR-190b <= 1.825671
|   |   hsa-miR-146a <= 10.021354
|   |   |   hsa-miR-101 <= 9.895896
|   |   |   |   hsa-miR-183* <= 2.053115: Normal-like (5.0/1.0)
|   |   |   |   hsa-miR-183* > 2.053115: ERBB2 (13.0)
|   |   |   hsa-miR-101 > 9.895896: Lum A (2.0)
|   |   hsa-miR-146a > 10.021354: Lum B (3.0/1.0)
|   hsa-miR-190b > 1.825671
|   |   hsa-miR-204 <= 5.073896
|   |   |   hsa-miR-142-3p <= 11.690557
|   |   |   |   hsa-miR-18a <= 6.725002
|   |   |   |   |   hsa-miR-25 <= 9.060418: Normal-like (2.0/1.0)
|   |   |   |   |   hsa-miR-25 > 9.060418: Lum A (40.0/1.0)
|   |   |   |   hsa-miR-18a > 6.725002: Lum B (3.0)
|   |   |   hsa-miR-142-3p > 11.690557: Lum B (4.0)
|   |   hsa-miR-204 > 5.073896: Normal-like (4.0)


Number of Leaves  :    12

Size of the tree :     23


=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          48               51.0638 %
Kappa statistic                          0.3256
Mean absolute error                      0.1975
Root mean squared error                  0.4318
Relative absolute error                 66.8383 %
Root relative squared error            112.0372 %
Coverage of cases (0.95 level)          52.1277 %
Mean rel. region size (0.95 level)      22.3404 %
Total Number of Instances               94
```

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0,683	0,264	0,667	0,683	0,675	0,418	0,673	0,622	Lum A
0,533	0,076	0,571	0,533	0,552	0,470	0,730	0,430	basal
0,300	0,107	0,250	0,300	0,273	0,178	0,506	0,172	Normal-like
0,000	0,134	0,000	0,000	0,000	-0,139	0,471	0,121	Lum B
0,563	0,077	0,600	0,563	0,581	0,498	0,747	0,464	ERBB2
W. Avg.	0,511	0,169	0,511	0,511	0,510	0,343	0,651	0,453

=== Confusion Matrix ===

a	b	c	d	e	<-- classified as
28	1	6	5	1	a = Lum A
1	8	1	3	2	b = basal
4	0	3	1	2	c = Normal-like
9	2	0	0	1	d = Lum B
0	3	2	2	9	e = ERBB2

Ek 2.1.6 SVM öznitelik seçimi uygulanarak mikroRNA bilgisi ve karar ağacı sınıflandırıcı ile göğüs kanseri alt kategori sınıflandırma sonuçları

```
Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     BreastCancerSubtypes_miRNA_SVM_Select
Instances:    94
Attributes:   101
Test mode:    94-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------


hsa-miR-18a <= 6.722492
|   hsa-miR-190b <= 1.92837
|   |   hsa-miR-342-3p <= 11.05032
|   |   |   hsa-miR-34b* <= 6.505558: basal (2.0)
|   |   |   hsa-miR-34b* > 6.505558
|   |   |   |   hsa-miR-183* <= 2.053115: Normal-like (4.0/1.0)
|   |   |   |   hsa-miR-183* > 2.053115: ERBB2 (14.0)
|   |   hsa-miR-342-3p > 11.05032
|   |   |   hcmv-miR-UL112 <= 1.136414: Lum A (3.0)
|   |   |   hcmv-miR-UL112 > 1.136414: Normal-like (2.0/1.0)
|   hsa-miR-190b > 1.92837
|   |   hsa-miR-142-3p <= 11.690557
|   |   |   hsa-miR-424 <= 10.301772
|   |   |   |   hsa-miR-185 <= 7.200661: Normal-like (2.0)
|   |   |   |   hsa-miR-185 > 7.200661: Lum A (39.0/1.0)
|   |   |   hsa-miR-424 > 10.301772: Normal-like (4.0)
|   |   hsa-miR-142-3p > 11.690557: Lum B (3.0)
hsa-miR-18a > 6.722492
|   hsa-miR-135b <= 5.993922
|   |   hsa-miR-298 <= 1.031447: Lum B (7.0)
|   |   hsa-miR-298 > 1.031447: basal (2.0/1.0)
|   hsa-miR-135b > 5.993922: basal (12.0)


Number of Leaves  :    12

Size of the tree :     23


=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          63               67.0213 %
Kappa statistic                          0.5418
Mean absolute error                      0.1357
Root mean squared error                  0.3536
Relative absolute error                 45.9254 %
Root relative squared error             91.7422 %
Coverage of cases (0.95 level)          67.0213 %
Mean rel. region size (0.95 level)      22.5532 %
Total Number of Instances               94
```

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0,829	0,170	0,791	0,829	0,810	0,656	0,791	0,762	Lum A
0,733	0,063	0,688	0,733	0,710	0,653	0,844	0,663	basal
0,400	0,060	0,444	0,400	0,421	0,357	0,668	0,292	Normal-like
0,250	0,061	0,375	0,250	0,300	0,226	0,425	0,246	Lum B
0,688	0,090	0,611	0,688	0,647	0,571	0,800	0,526	ERBB2
W.Avg.0,670	0,114	0,654	0,670	0,660	0,555	0,741	0,590	

=== Confusion Matrix ===

a	b	c	d	e	<-- classified as
34	1	3	2	1	a = Lum A
1	11	0	2	1	b = basal
3	0	4	0	3	c = Normal-like
4	3	0	3	2	d = Lum B
1	1	2	1	11	e = ERBB2

Ek 2.1.7 Öznitelik seçimi olmadan mikroRNA bilgisi ve KNN sınıflandırıcı ile göğüs kanseri alt kategori sınıflandırma sonuçları

```
Scheme:       weka.classifiers.lazy.IBk -K 1 -W 0 -A
"weka.core.neighboursearch.LinearNNSearch -A
\"weka.core.EuclideanDistance -R first-last\""
Relation:     BreastCancerSubtypes_miRNA
Instances:    94
Attributes:   490
Test mode:    94-fold cross-validation

=== Classifier model (full training set) ===

IB1 instance-based classifier
using 1 nearest neighbour(s) for classification



=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          47               50 %
Kappa statistic                          0.308
Mean absolute error                      0.2061
Root mean squared error                  0.4361
Relative absolute error                 69.7473 %
Root relative squared error            113.1521 %
Coverage of cases (0.95 level)          50      %
Mean rel. region size (0.95 level)      20      %
Total Number of Instances               94

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
      0,732    0,245    0,698      0,732    0,714      0,484    0,743     0,628     Lum A
0,733    0,051    0,733      0,733    0,733      0,683    0,841     0,580     basal
0,100    0,083    0,125      0,100    0,111      0,018    0,508     0,108     Normal-like
      0,167    0,171    0,125      0,167    0,143      -0,004   0,498     0,127     Lum B
      0,188    0,115    0,250      0,188    0,214      0,081    0,536     0,185     ERBB2
W.Avg.0,500    0,165    0,493      0,500    0,495      0,335    0,667     0,426

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 30  0  4  4  3 |  a = Lum A
  1 11  1  2  0 |  b = basal
  5  1  1  1  2 |  c = Normal-like
  3  3  0  2  4 |  d = Lum B
  4  0  2  7  3 |  e = ERBB2
```

Ek 2.1.8 SVM öznitelik seçimi uygulanarak mikroRNA bilgisi ve KNN sınıflandırıcı ile göğüs kanseri alt kategori sınıflandırma sonuçları

```
Scheme:       weka.classifiers.lazy.IBk -K 1 -W 0 -A
"weka.core.neighboursearch.LinearNNSearch -A
\"weka.core.EuclideanDistance -R first-last\""
Relation:     BreastCancerSubtypes_miRNA_SVM_Select
Instances:    94
Attributes:   101
Test mode:    94-fold cross-validation

=== Classifier model (full training set) ===

IB1 instance-based classifier
using 1 nearest neighbour(s) for classification



=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          60               63.8298 %
Kappa statistic                          0.4935
Mean absolute error                      0.1536
Root mean squared error                  0.3711
Relative absolute error                 51.9835 %
Root relative squared error             96.2797 %
Coverage of cases (0.95 level)          63.8298 %
Mean rel. region size (0.95 level)      20      %
Total Number of Instances               94

=== Detailed Accuracy By Class ===

TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
0,829    0,208    0,756      0,829    0,791      0,617    0,811     0,701     Lum A
0,867    0,025    0,867      0,867    0,867      0,841    0,921     0,772     basal
0,200    0,048    0,333      0,200    0,250      0,192    0,576     0,152     Normal-like
0,333    0,110    0,308      0,333    0,320      0,216    0,612     0,188     Lum B
0,438    0,103    0,467      0,438    0,452      0,344    0,667     0,300     ERBB2
W. Avg.0,638    0,131    0,622      0,638    0,627      0,510    0,754     0,520

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 34  0  3  2  2 |  a = Lum A
  0 13  0  2  0 |  b = basal
  4  1  2  0  3 |  c = Normal-like
  4  1  0  4  3 |  d = Lum B
  3  0  1  5  7 |  e = ERBB2
```

Ek 2.2 mRNA bilgisi ile göğüs kanseri alt kategori sınıflandırma sonuçları

Ek 2.2.1 Öznitelik seçimi olmadan mRNA bilgisi ve ANN sınıflandırıcı ile göğüs kanseri alt kategori sınıflandırma sonuçları

```
Scheme:weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N
500 -V 0 -S 0 -E 20 -H a
Relation:     BreastCancerSubtypes_mRNA
Instances:    94
Attributes:   491
Test mode:94-fold cross-validation

=== Classifier model (full training set) ===
Class Lum A
    Input
    Node 0
Class basal
    Input
    Node 1
Class Normal-like
    Input
    Node 2
Class Lum B
    Input
    Node 3
Class ERBB2
    Input
    Node 4

=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances          92               97.8723 %
Incorrectly Classified Instances         2                2.1277 %
Kappa statistic                          0.9708
Mean absolute error                      0.0215
Root mean squared error                  0.0885
Relative absolute error                  7.2887 %
Root relative squared error             22.9651 %
Total Number of Instances               94

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0          1         1         1          1        Lum A
                 1         0          1         1         1          1        basal
                 0.9       0.012      0.9       0.9       0.9        0.996    Normal-like
                 1         0          1         1         1          1        Lum B
                 0.938     0.013      0.938     0.938     0.938      0.998    ERBB2
Weighted Avg.    0.979     0.003      0.979     0.979     0.979      0.999

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 41  0  0  0  0 |  a = Lum A
  0 15  0  0  0 |  b = basal
  0  0  9  0  1 |  c = Normal-like
  0  0  0 12  0 |  d = Lum B
  0  0  1  0 15 |  e = ERBB2
```

Ek 2.2.2 SVM öznitelik seçimi uygulanarak mRNA bilgisi ve ANN sınıflandırıcı ile göğüs kanseri alt kategori sınıflandırma sonuçları

```
Scheme:weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N
500 -V 0 -S 0 -E 20 -H a
Relation:     BreastCancerSubtypes_mRNA_SVM_Select
Instances:    94
Attributes:   101
Test mode:94-fold cross-validation

=== Classifier model (full training set) ===

Class Lum A
    Input
    Node 0
Class basal
    Input
    Node 1
Class Normal-like
    Input
    Node 2
Class Lum B
    Input
    Node 3
Class ERBB2
    Input
    Node 4

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          93               98.9362 %
Incorrectly Classified Instances         1                1.0638 %
Kappa statistic                          0.9854
Mean absolute error                      0.022
Root mean squared error                  0.0767
Relative absolute error                  7.4465 %
Root relative squared error             19.8888 %
Total Number of Instances               94

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.976     0          1         0.976     0.988      0.999    Lum A
                 1         0          1         1         1          1        basal
                 1         0          1         1         1          1        Normal-like
                 1         0          1         1         1          1        Lum B
                 1         0.013      0.941     1         0.97       0.999    ERBB2
Weighted Avg.    0.989     0.002      0.99      0.989     0.989      0.999

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 40  0  0  0  1 |  a = Lum A
  0 15  0  0  0 |  b = basal
  0  0 10  0  0 |  c = Normal-like
  0  0  0 12  0 |  d = Lum B
  0  0  0  0 16 |  e = ERBB2
```

Ek 2.2.3 Öznitelik seçimi olmadan mRNA bilgisi ve SVM sınıflandırıcı ile göğüs kanseri alt kategori sınıflandırma sonuçları

```
Scheme:       weka.classifiers.functions.SMO -C 1.0 -L 0.001 -P 1.0E-12 -
N 2 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -C
250007 -E 1.0"
Relation:     BreastCancerSubtypes_mRNA
Instances:    94
Attributes:   40493
Test mode:    94-fold cross-validation

=== Classifier model (full training set) ===

SMO

Kernel used:
  Linear Kernel: K(x,y) = <x,y>

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          68               72.3404 %
Kappa statistic                          0.6023
Mean absolute error                      0.2587
Root mean squared error                  0.3442
Relative absolute error                 87.5463 %
Root relative squared error             89.2984 %
Coverage of cases (0.95 level)          98.9362 %
Mean rel. region size (0.95 level)      80      %
Total Number of Instances               94

=== Detailed Accuracy By Class ===

TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
0,927    0,245    0,745      0,927    0,826      0,678    0,877     0,771     Lum A
0,933    0,000    1,000      0,933    0,966      0,960    0,965     0,944     basal
0,000    0,083    0,000      0,000    0,000      -0,098   0,508     0,112     Normal-like
0,583    0,012    0,875      0,583    0,700      0,683    0,892     0,648     Lum B
0,563    0,064    0,643      0,563    0,600      0,526    0,826     0,483     ERBB2
W.Avg.0,723    0,128    0,706      0,723    0,706      0,615    0,845     0,664

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 38  0  2  0  1 |  a = Lum A
  0 14  0  0  1 |  b = basal
  7  0  0  0  3 |  c = Normal-like
  5  0  0  7  0 |  d = Lum B
  1  0  5  1  9 |  e = ERBB2
```

Ek 2.2.4 SVM öznitelik seçimi uygulanarak mRNA bilgisi ve SVM sınıflandırıcı ile göğüs kanseri alt kategori sınıflandırma sonuçları

```
Scheme:weka.classifiers.functions.SMO -C 1.0 -L 0.0010 -P 1.0E-12 -N 2 -V
-1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -C 250007
-E 1.0"
Relation:     BreastCancerSubtypes_mRNA_SVM_Select
Instances:    94
Attributes:   101
Test mode:94-fold cross-validation

=== Classifier model (full training set) ===

SMO

Kernel used:
  Linear Kernel: K(x,y) = <x,y>

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          87               92.5532 %
Incorrectly Classified Instances         7                7.4468 %
Kappa statistic                          0.8976
Mean absolute error                      0.2438
Root mean squared error                  0.322
Relative absolute error                 82.5066 %
Root relative squared error             83.5311 %
Total Number of Instances               94

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.951     0.038      0.951     0.951     0.951      0.982    Lum A
                 1         0          1         1         1          1        basal
                 0.8       0.012      0.889     0.8       0.842      0.905    Normal-like
                 0.917     0.012      0.917     0.917     0.917      0.987    Lum B
                 0.875     0.038      0.824     0.875     0.848      0.947    ERBB2
Weighted Avg.    0.926     0.026      0.926     0.926     0.925      0.971

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 39  0  0  1  1 |  a = Lum A
  0 15  0  0  0 |  b = basal
  1  0  8  0  1 |  c = Normal-like
  0  0  0 11  1 |  d = Lum B
  1  0  1  0 14 |  e = ERBB2
```

Ek 2.2.5 Öznitelik seçimi olmadan mRNA bilgisi ve karar ağacı sınıflandırıcı ile göğüs kanseri alt kategori sınıflandırma sonuçları

```
Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     BreastCancerSubtypes_mRNA
Instances:    94
Attributes:   40493
Test mode:    94-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

A_23_P165778 <= 12.606773: basal (15.0)
A_23_P165778 > 12.606773
|   A_23_P309739 <= 10.120992
|   |   A_23_P101960 <= 12.263164: Lum B (4.0)
|   |   A_23_P101960 > 12.263164
|   |   |   A_23_P101992 <= 7.882635: Normal-like (5.0/1.0)
|   |   |   A_23_P101992 > 7.882635: ERBB2 (17.0/1.0)
|   A_23_P309739 > 10.120992
|   |   A_23_P210731 <= 7.662148
|   |   |   A_23_P55270 <= 9.777469: Lum A (39.0/1.0)
|   |   |   A_23_P55270 > 9.777469: Lum B (9.0/1.0)
|   |   A_23_P210731 > 7.662148: Normal-like (5.0)

Number of Leaves  :    7
Size of the tree :     13
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances          62               65.9574 %
Kappa statistic                          0.5286
Mean absolute error                      0.1413
Root mean squared error                  0.3631
Relative absolute error                 47.8074 %
Root relative squared error             94.2076 %
Coverage of cases (0.95 level)          68.0851 %
Mean rel. region size (0.95 level)      22.766  %
Total Number of Instances               94

=== Detailed Accuracy By Class ===

TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
0,805    0,189    0,767      0,805    0,786      0,613    0,712     0,703     Lum A
0,933    0,000    1,000      0,933    0,966      0,960    0,967     0,944     basal
0,100    0,095    0,111      0,100    0,105      0,005    0,317     0,108     Normal-like
0,500    0,098    0,429      0,500    0,462      0,377    0,728     0,274     Lum B
0,500    0,077    0,571      0,500    0,533      0,447    0,726     0,358     ERBB2
W.Avg.0,660    0,118    0,658      0,660    0,658      0,545    0,715     0,565

=== Confusion Matrix ===
  a  b  c  d  e   <-- classified as
 33  0  4  2  2 |  a = Lum A
  0 14  0  1  0 |  b = basal
  4  0  1  2  3 |  c = Normal-like
  3  0  2  6  1 |  d = Lum B
  3  0  2  3  8 |  e = ERBB2
```

Ek 2.2.6 SVM öznitelik seçimi uygulanarak mRNA bilgisi ve karar ağacı sınıflandırıcı ile göğüs kanseri alt kategori sınıflandırma sonuçları

```
Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     BreastCancerSubtypes_mRNA_SVM_Select
Instances:    94
Attributes:   101
Test mode:94-fold cross-validation

=== Classifier model (full training set) ===
J48 pruned tree
------------------
A_23_P37127 <= 9.025836: basal (15.0)
A_23_P37127 > 9.025836
|   A_23_P55270 <= 9.723034
|   |   A_23_P163992 <= 11.407546
|   |   |   A_23_P152949 <= 7.206625
|   |   |   |   A_23_P135568 <= 9.141778: Normal-like (4.0)
|   |   |   |   A_23_P135568 > 9.141778: ERBB2 (2.0)
|   |   |   A_23_P152949 > 7.206625
|   |   |   |   A_23_P396765 <= 10.174733: Lum A (40.0/1.0)
|   |   |   |   A_23_P396765 > 10.174733: Normal-like (4.0/1.0)
|   |   A_23_P163992 > 11.407546: ERBB2 (5.0/1.0)
|   A_23_P55270 > 9.723034
|   |   A_23_P45011 <= 7.322676: Lum B (12.0/1.0)
|   |   A_23_P45011 > 7.322676
|   |   |   A_23_P256033 <= 9.36892: ERBB2 (9.0)
|   |   |   A_23_P256033 > 9.36892: Normal-like (3.0/1.0)

Number of Leaves  :    9
Size of the tree :     17
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances          64               68.0851 %
Incorrectly Classified Instances        30               31.9149 %
Kappa statistic                          0.5671
Mean absolute error                      0.1347
Root mean squared error                  0.3457
Relative absolute error                 45.5627 %
Root relative squared error             89.6825 %
Total Number of Instances               94

=== Detailed Accuracy By Class ===
               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.805     0.094      0.868     0.805     0.835      0.83     Lum A
                 0.933     0          1         0.933     0.966      0.967    basal
                 0.1       0.119      0.091     0.1       0.095      0.361    Normal-like
                 0.667     0.061      0.615     0.667     0.64       0.818    Lum B
                 0.5       0.128      0.444     0.5       0.471      0.678    ERBB2
Weighted Avg.    0.681     0.083      0.702     0.681     0.69       0.775

=== Confusion Matrix ===
  a  b  c  d  e   <-- classified as
 33  0  3  3  2 |  a = Lum A
  0 14  1  0  0 |  b = basal
  4  0  1  0  5 |  c = Normal-like
  0  0  1  8  3 |  d = Lum B
  1  0  5  2  8 |  e = ERBB2
```

Ek 2.2.7 Öznitelik seçimi olmadan mRNA bilgisi ve KNN sınıflandırıcı ile göğüs kanseri alt kategori sınıflandırma sonuçları

```
Scheme:       weka.classifiers.lazy.IBk -K 1 -W 0 -A
"weka.core.neighboursearch.LinearNNSearch -A
\"weka.core.EuclideanDistance -R first-last\""
Relation:     BreastCancerSubtypes_mRNA
Instances:    94
Attributes:   40493
Test mode:    94-fold cross-validation

=== Classifier model (full training set) ===

IB1 instance-based classifier
using 1 nearest neighbour(s) for classification



=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          54               57.4468 %
Kappa statistic                          0.3754
Mean absolute error                      0.1779
Root mean squared error                  0.4024
Relative absolute error                 60.1822 %
Root relative squared error            104.4064 %
Coverage of cases (0.95 level)          57.4468 %
Mean rel. region size (0.95 level)      20      %
Total Number of Instances               94

=== Detailed Accuracy By Class ===

TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
0,878    0,377    0,643      0,878    0,742      0,506    0,750     0,618     Lum A
0,667    0,000    1,000      0,667    0,800      0,792    0,833     0,720     basal
0,100    0,095    0,111      0,100    0,105      0,005    0,502     0,107     Normal-like
0,167    0,049    0,333      0,167    0,222      0,161    0,559     0,162     Lum B
0,313    0,103    0,385      0,313    0,345      0,229    0,605     0,237     ERBB2
W.Avg.0,574    0,198    0,560      0,574    0,550      0,407    0,688     0,457

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 36  0  3  1  1 |  a = Lum A
  1 10  0  2  2 |  b = basal
  6  0  1  0  3 |  c = Normal-like
  6  0  2  2  2 |  d = Lum B
  7  0  3  1  5 |  e = ERBB2
```

Ek 2.2.8 SVM öznitelik seçimi uygulanarak mRNA bilgisi ve KNN sınıflandırıcı ile göğüs kanseri alt kategori sınıflandırma sonuçları

```
Scheme:weka.classifiers.lazy.IBk -K 1 -W 0 -X -A
"weka.core.neighboursearch.LinearNNSearch -A
\"weka.core.EuclideanDistance -R first-last\""
Relation:     BreastCancerSubtypes_mRNA_SVM_Select
Instances:    94
Attributes:   101
Test mode:94-fold cross-validation

=== Classifier model (full training set) ===

IB1 instance-based classifier
using 1 nearest neighbour(s) for classification



=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          81               86.1702 %
Incorrectly Classified Instances        13               13.8298 %
Kappa statistic                          0.8032
Mean absolute error                      0.0688
Root mean squared error                  0.23
Relative absolute error                 23.2883 %
Root relative squared error             59.6795 %
Total Number of Instances               94

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.951     0.189      0.796     0.951     0.867      0.881    Lum A
                 1         0          1         1         1          1        basal
                 0.5       0.024      0.714     0.5       0.588      0.738    Normal-like
                 0.917     0          1         0.917     0.957      0.958    Lum B
                 0.688     0.013      0.917     0.688     0.786      0.837    ERBB2
Weighted Avg.    0.862     0.087      0.866     0.862     0.856      0.887

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 39  0  1  0  1 |  a = Lum A
  0 15  0  0  0 |  b = basal
  5  0  5  0  0 |  c = Normal-like
  1  0  0 11  0 |  d = Lum B
  4  0  1  0 11 |  e = ERBB2
```

Ek 2.3 mikroRNA + mRNA bilgisi ile göğüs kanseri alt kategori sınıflandırma sonuçları

Ek 2.3.1 Öznitelik seçimi olmadan mikroRNA + mRNA bilgisi ve ANN sınıflandırıcı ile göğüs kanseri alt kategori sınıflandırma sonuçları

```
Scheme:weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N
500 -V 0 -S 0 -E 20 -H a
Relation:     BreastCancerSubtypes
Instances:    94
Attributes:   491
Test mode:94-fold cross-validation

=== Classifier model (full training set) ===
Class Lum A
    Input
    Node 0
Class basal
    Input
    Node 1
Class Normal-like
    Input
    Node 2
Class Lum B
    Input
    Node 3
Class ERBB2
    Input
    Node 4

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          88               93.617  %
Incorrectly Classified Instances         6                6.383  %
Kappa statistic                          0.9114
Mean absolute error                      0.0369
Root mean squared error                  0.1394
Relative absolute error                 12.4955 %
Root relative squared error             36.1686 %
Total Number of Instances               94

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.976     0.057      0.93      0.976     0.952      0.99     Lum A
                 1         0          1         1         1          1        basal
                 0.7       0.012      0.875     0.7       0.778      0.986    Normal-like
                 0.917     0          1         0.917     0.957      1        Lum B
                 0.938     0.026      0.882     0.938     0.909      0.983    ERBB2
Weighted Avg.    0.936     0.03       0.936     0.936     0.935      0.991

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 40  0  0  0  1 |  a = Lum A
  0 15  0  0  0 |  b = basal
  2  0  7  0  1 |  c = Normal-like
  1  0  0 11  0 |  d = Lum B
  0  0  1  0 15 |  e = ERBB2
```

Ek 2.3.2 SVM öznitelik seçimi uygulanarak mikroRNA + mRNA bilgisi ve ANN sınıflandırıcı ile göğüs kanseri alt kategori sınıflandırma sonuçları

```
Scheme:weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N
500 -V 0 -S 0 -E 20 -H a
Relation:     BreastCancerSubtypes_SVM_Select
Instances:    94
Attributes:   101
Test mode:94-fold cross-validation

=== Classifier model (full training set) ===

Class Lum A
    Input
    Node 0
Class basal
    Input
    Node 1
Class Normal-like
    Input
    Node 2
Class Lum B
    Input
    Node 3
Class ERBB2
    Input
    Node 4

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          93               98.9362 %
Incorrectly Classified Instances         1                1.0638 %
Kappa statistic                          0.9854
Mean absolute error                      0.0271
Root mean squared error                  0.098
Relative absolute error                  9.1726 %
Root relative squared error             25.4379 %
Total Number of Instances               94

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.976     0          1         0.976     0.988      0.988    Lum A
                 1         0          1         1         1          1        basal
                 1         0          1         1         1          0.999    Normal-like
                 1         0          1         1         1          1        Lum B
                 1         0.013      0.941     1         0.97       0.992    ERBB2
Weighted Avg.    0.989     0.002      0.99      0.989     0.989      0.993

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 40  0  0  0  1 |  a = Lum A
  0 15  0  0  0 |  b = basal
  0  0 10  0  0 |  c = Normal-like
  0  0  0 12  0 |  d = Lum B
  0  0  0  0 16 |  e = ERBB2
```

Ek 2.3.3 Öznitelik seçimi olmadan mikroRNA + mRNA bilgisi ve SVM sınıflandırıcı ile göğüs kanseri alt kategori sınıflandırma sonuçları

```
Scheme:weka.classifiers.functions.SMO -C 1.0 -L 0.0010 -P 1.0E-12 -N 2 -V
-1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -C 250007
-E 1.0"
Relation:     BreastCancerSubtypes
Instances:    94
Attributes:   40982
Test mode:94-fold cross-validation

=== Classifier model (full training set) ===

SMO

Kernel used:
  Linear Kernel: K(x,y) = <x,y>

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          70               74.4681 %
Incorrectly Classified Instances        24               25.5319 %
Kappa statistic                          0.6342
Mean absolute error                      0.2579
Root mean squared error                  0.3429
Relative absolute error                 87.2583 %
Root relative squared error             88.961  %
Total Number of Instances               94

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                0.927     0.226      0.76      0.927     0.835      0.882    Lum A
                0.933     0          1         0.933     0.966      0.966    basal
                0         0.071      0         0         0          0.518    Normal-like
                0.667     0.012      0.889     0.667     0.762      0.891    Lum B
                0.625     0.064      0.667     0.625     0.645      0.837    ERBB2
Weighted Avg.   0.745     0.119      0.718     0.745     0.725      0.85

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 38  0  2  0  1 |  a = Lum A
  0 14  0  0  1 |  b = basal
  7  0  0  0  3 |  c = Normal-like
  4  0  0  8  0 |  d = Lum B
  1  0  4  1 10 |  e = ERBB2
```

Ek 2.3.4 SVM öznitelik seçimi uygulanarak mikroRNA + mRNA bilgisi ve SVM sınıflandırıcı ile göğüs kanseri alt kategori sınıflandırma sonuçları

```
Scheme:weka.classifiers.functions.SMO -C 1.0 -L 0.0010 -P 1.0E-12 -N 0 -V
-1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -C 250007
-E 1.0"
Relation:     BreastCancerSubtypes_SVM_Select
Instances:    94
Attributes:   101
Test mode:94-fold cross-validation

=== Classifier model (full training set) ===

SMO

Kernel used:
  Linear Kernel: K(x,y) = <x,y>

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          92               97.8723 %
Incorrectly Classified Instances         2                2.1277 %
Kappa statistic                          0.9709
Mean absolute error                      0.2421
Root mean squared error                  0.3194
Relative absolute error                 81.9306 %
Root relative squared error             82.8597 %
Total Number of Instances               94

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
               0.976     0          1         0.976     0.988      0.983    Lum A
               1         0          1         1         1          1        basal
               1         0          1         1         1          0.973    Normal-like
               0.917     0          1         0.917     0.957      0.995    Lum B
               1         0.026      0.889     1         0.941      0.987    ERBB2
Weighted Avg.  0.979     0.004      0.981     0.979     0.979      0.987

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 40  0  0  0  1 |  a = Lum A
  0 15  0  0  0 |  b = basal
  0  0 10  0  0 |  c = Normal-like
  0  0  0 11  1 |  d = Lum B
  0  0  0  0 16 |  e = ERBB2
```

Ek 2.3.5 Öznitelik seçimi olmadan mikroRNA + mRNA bilgisi ve karar ağacı sınıflandırıcı ile göğüs kanseri alt kategori sınıflandırma sonuçları

```
Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     BreastCancerSubtypes
Instances:    94
Attributes:   40982
Test mode:94-fold cross-validation
=== Classifier model (full training set) ===
J48 pruned tree
------------------
A_23_P165778 <= 12.606773: basal (15.0)
A_23_P165778 > 12.606773
|   hsa-miR-190b <= 1.825671
|   |   A_23_P258931 <= 7.466581
|   |   |   hsa-miR-183* <= 2.053115: Normal-like (5.0/1.0)
|   |   |   hsa-miR-183* > 2.053115: ERBB2 (15.0)
|   |   A_23_P258931 > 7.466581
|   |   |   A_23_P56150 <= 15.136319: Lum B (4.0)
|   |   |   A_23_P56150 > 15.136319: Lum A (2.0)
|   hsa-miR-190b > 1.825671
|   |   A_23_P55270 <= 9.777469
|   |   |   hsa-miR-424 <= 10.301772: Lum A (39.0)
|   |   |   hsa-miR-424 > 10.301772: Normal-like (4.0)
|   |   A_23_P55270 > 9.777469
|   |   |   A_32_P104448 <= 7.272844: Lum B (8.0)
|   |   |   A_32_P104448 > 7.272844: Normal-like (2.0)

Number of Leaves  :    9
Size of the tree :     17
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances          61               64.8936 %
Incorrectly Classified Instances        33               35.1064 %
Kappa statistic                          0.5261
Mean absolute error                      0.1383
Root mean squared error                  0.3655
Relative absolute error                 46.7889 %
Root relative squared error             94.8254 %
Total Number of Instances               94
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.805     0.075      0.892     0.805     0.846      0.869    Lum A
                 0.933     0          1         0.933     0.966      0.967    basal
                 0.3       0.107      0.25      0.3       0.273      0.691    Normal-like
                 0.333     0.11       0.308     0.333     0.32       0.622    Lum B
                 0.438     0.141      0.389     0.438     0.412      0.698    ERBB2
Weighted Avg.    0.649     0.082      0.681     0.649     0.663      0.805

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 33  0  3  2  3 |  a = Lum A
  1 14  0  0  0 |  b = basal
  0  0  3  3  4 |  c = Normal-like
  3  0  1  4  4 |  d = Lum B
  0  0  5  4  7 |  e = ERBB2
```

Ek 2.3.6 SVM öznitelik seçimi uygulanarak mikroRNA + mRNA bilgisi ve karar ağacı sınıflandırıcı ile göğüs kanseri alt kategori sınıflandırma sonuçları

```
Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     BreastCancerSubtypes_SVM_Select
Instances:    94
Attributes:   101
Test mode:94-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

A_23_P165778 <= 12.606773: basal (15.0)
A_23_P165778 > 12.606773
|   hsa-miR-190b <= 1.825671
|   |   hsa-miR-149 <= 7.285571
|   |   |   A_23_P350396 <= 8.303366
|   |   |   |   A_23_P101992 <= 7.882635: Normal-like (4.0/1.0)
|   |   |   |   A_23_P101992 > 7.882635: ERBB2 (17.0/1.0)
|   |   |   A_23_P350396 > 8.303366: Lum B (3.0)
|   |   hsa-miR-149 > 7.285571: Lum A (2.0)
|   hsa-miR-190b > 1.825671
|   |   A_23_P137665 <= 9.7198
|   |   |   A_23_P401055 <= 7.464031: Lum A (36.0)
|   |   |   A_23_P401055 > 7.464031: Lum B (3.0)
|   |   A_23_P137665 > 9.7198
|   |   |   A_24_P224488 <= 11.427867
|   |   |   |   hsa-miR-18a <= 5.689179: Lum A (3.0)
|   |   |   |   hsa-miR-18a > 5.689179: Lum B (5.0)
|   |   |   A_24_P224488 > 11.427867: Normal-like (6.0)

Number of Leaves  :    10

Size of the tree :     19


Time taken to build model: 0.06 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          65               69.1489 %
Incorrectly Classified Instances        29               30.8511 %
Kappa statistic                          0.5813
Mean absolute error                      0.1289
Root mean squared error                  0.3459
Relative absolute error                 43.6211 %
Root relative squared error             89.7337 %
Total Number of Instances               94
```

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.756	0.132	0.816	0.756	0.785	0.851	Lum A
	0.933	0	1	0.933	0.966	0.967	basal
	0.5	0.071	0.455	0.5	0.476	0.686	Normal-like
	0.25	0.11	0.25	0.25	0.25	0.563	Lum B
	0.75	0.09	0.632	0.75	0.686	0.801	ERBB2
Weighted Avg.	0.691	0.094	0.703	0.691	0.696	0.807	

=== Confusion Matrix ===

a	b	c	d	e		<-- classified as
31	0	3	5	2	|	a = Lum A
0	14	0	0	1	|	b = basal
4	0	5	0	1	|	c = Normal-like
3	0	3	3	3	|	d = Lum B
0	0	0	4	12	|	e = ERBB2

Ek 2.3.7 Öznitelik seçimi olmadan mikroRNA + mRNA bilgisi ve KNN sınıflandırıcı ile göğüs kanseri alt kategori sınıflandırma sonuçları

```
Scheme:weka.classifiers.lazy.IBk -K 1 -W 0 -A
"weka.core.neighboursearch.LinearNNSearch -A
\"weka.core.EuclideanDistance -R first-last\""
Relation:     BreastCancerSubtypes
Instances:    94
Attributes:   40982
Test mode:94-fold cross-validation

=== Classifier model (full training set) ===

IB1 instance-based classifier
using 1 nearest neighbour(s) for classification



=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          56               59.5745 %
Incorrectly Classified Instances        38               40.4255 %
Kappa statistic                          0.4005
Mean absolute error                      0.1698
Root mean squared error                  0.3923
Relative absolute error                 57.4493 %
Root relative squared error            101.7696 %
Total Number of Instances               94

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.927     0.377      0.655     0.927     0.768      0.775    Lum A
                 0.667     0          1         0.667     0.8        0.833    basal
                 0.1       0.071      0.143     0.1       0.118      0.514    Normal-like
                 0.167     0.049      0.333     0.167     0.222      0.559    Lum B
                 0.313     0.103      0.385     0.313     0.345      0.605    ERBB2
Weighted Avg.    0.596     0.196      0.569     0.596     0.562      0.7

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 38  0  1  1  1 |  a = Lum A
  1 10  0  2  2 |  b = basal
  6  0  1  0  3 |  c = Normal-like
  6  0  2  2  2 |  d = Lum B
  7  0  3  1  5 |  e = ERBB2
```

Ek 2.3.8 SVM öznitelik seçimi uygulanarak mikroRNA + mRNA bilgisi ve KNN sınıflandırıcı ile göğüs kanseri alt kategori sınıflandırma sonuçları

```
Scheme:weka.classifiers.lazy.IBk -K 1 -W 0 -A
"weka.core.neighboursearch.LinearNNSearch -A
\"weka.core.EuclideanDistance -R first-last\""
Relation:     BreastCancerSubtypes_SVM_Select
Instances:    94
Attributes:   101
Test mode:94-fold cross-validation

=== Classifier model (full training set) ===

IB1 instance-based classifier
using 1 nearest neighbour(s) for classification



=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          79               84.0426 %
Incorrectly Classified Instances        15               15.9574 %
Kappa statistic                          0.7757
Mean absolute error                      0.0769
Root mean squared error                  0.247
Relative absolute error                 26.0212 %
Root relative squared error             64.0725 %
Total Number of Instances               94

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.951     0.132      0.848     0.951     0.897      0.91     Lum A
                 1         0          1         1         1          1        basal
                 0.5       0.048      0.556     0.5       0.526      0.726    Normal-like
                 0.75      0          1         0.75      0.857      0.875    Lum B
                 0.688     0.051      0.733     0.688     0.71       0.818    ERBB2
Weighted Avg.    0.84      0.071      0.841     0.84      0.837      0.885

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 39  0  1  0  1 |  a = Lum A
  0 15  0  0  0 |  b = basal
  3  0  5  0  2 |  c = Normal-like
  2  0  0  9  1 |  d = Lum B
  2  0  3  0 11 |  e = ERBB2
```