İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

KÜTLE SPEKTROMETRESİ VERİLERİNİN ANALİZİYLE PROSTAT ve YUMURTALIK KANSERLERİNİN BELİRLENMESİ

YÜKSEK LİSANS TEZİ

Vedat TAŞKIN

504101400

Elektronik ve Haberleşme Mühendisliği Anabilim Dalı

Biyomedikal Mühendisliği Programı

Tez Danışmanı: Prof. Dr. Tamer ÖLMEZ

25 OCAK 2013

İTÜ, Fen Bilimleri Enstitüsü'nün 504101400 numaralı Yüksek Lisans Öğrencisi **Vedat TAŞKIN**, ilgili yönetmeliklerin belirlediği gerekli tüm şartları yerine getirdikten sonra hazırladığı "KÜTLE SPEKTROMETRESİ VERİLERİNİN ANALİZİYLE PROSTAT ve YUMURTALIK KANSERLERİNİN BELİRLENMESİ" başlıklı tezini aşağıda imzaları olan jüri önünde başarı ile sunmuştur.

Tez Danışmanı : **Prof. Dr. Tamer ÖLMEZ** .....
İstanbul Teknik Üniversitesi

Jüri Üyeleri : **Prof. Dr. Mehmet KORÜREK** .....
İstanbul Teknik Üniversitesi

Prof. Dr. Nizamettin AYDIN .....
Yıldız Teknik Üniversitesi

Teslim Tarihi : 17 Aralık 2012

Savunma Tarihi : 24 Ocak 2013

Eşime ve oğluma,

ÖNSÖZ

Bu yüksek lisans tezinin hazırlanmasında başta tez danışmanım Prof. Dr. Tamer Ölmez olmak üzere, yardımlarını esirgemeyen Yük. Müh. Berat Doğan'a teşekkürü bir borç bilirim. Yüksek lisans boyunca desteklerini esirgemeyen araştırma görevlisi arkadaşlarıma minnettarım.
Tezimi hiçbir şekilde hakkını ödeyemeyeceğim sevgili eşim Esma'ya ithaf ediyorum.

Aralık 2012

Vedat TAŞKIN
(Elektrik-Elektronik Mühendisi)

İÇİNDEKİLER

Sayfa

KISALTMALAR

ACO : Karınca Koloni Optimizasyonu
ADD : Ayrık Dalgacık Dönüşümü
API : Atmospheric Pressure Ionizaiton
C : Centroid
CA : Cancer Antigen
CSV : Comma-Separated Variables
ÇKEKK : Çekirdek Kısmi En Küçük Kareler
EEG : Electroencephalography
EW : Equivalent Width
ESI : Electrosprey Ionizaiton
FAB : Fast Atom Bombardment
IP : Instant at Which Peak Occurs
ICA : Independent Component Analysis
KNN : K- Nearest Neihgbor
KPLS : Kernel Partial Least Squares
KS : Kütle Spektrometresi
LC : Sıvı Kramatografisi
LDA : Linear Discriminant Analaysis
MALDI : Matrix Assisted Laser Desorption Ionization
MS : Mass Spectrometrey
MSA : Mean Square Abscissa
NN : Neural Network
PV : Peak Value
PCA : Principal Component Analaysis
PSA : Prostat Specific Agent
PSO : Parçacık Sürü Optimizasyonu
SELDI : Surface Enhanced Laser Desorption Ionization
SVM : Support Vector Machines
TBA : Temel Bileşenler Analizi
TOF : Time Of Flight

ÇİZELGE LİSTESİ

Sayfa

ŞEKİL LİSTESİ

Sayfa

KÜTLE SPEKTROMETRESİ VERİLERİNİN ANALİZİYLE PROSTAT ve YUMURTALIK KANSERLERİNİN BELİRLENMESİ

ÖZET

Kanser hastalıkları ölüm nedeni olarak kalp ve damar hastalıklarının hemen ardından gelmektedir. Kanser hastalıkları genellikle yaşlılık dönemlerinde görülse de bu hastalıkların her yaştan insanı etkilediği de bilinen bir gerçektir. Kanser tedavisi üzerinde yapılan çalışmalar son yıllarda oldukça artmıştır. Bunun yanında erken teşhis konulması tedavi olasılığını arttırmaktadır. Yumurtalık ve prostat kanseri gibi bazı kanser türlerinde belirtiler kuvvetli olmadığı için erken teşhis oldukça zordur. Bu zorluğun üstesinden gelmek için erken tanı yöntemleri üzerinde hala çalışılmaktadır. Proteomların kütle spektrometresi ile analiz edilmesi sonucu elde edilen verilerin erken tanı için kullanılabileceği görülmüştür. Biyoinformatiğin bu alt kolu son yıllarda oldukça popülerlik kazanmış ve bu alanda bir çok çalışmanın yapılmasını sağlamıştır. Geliştirilen yöntemler ile kütle spektrometresi verileri üzerinde tümörün varlığını gösteren biyobelirteçler aranmaktadır. Bu alanda yapılan çalışmalarda genellikle bulunan biyobelirteçler veri kümesine bağımlı olmaktadır. Dolayısıyla veri kümesinden bağımsız yöntemlerin geliştirilmesi önem kazanmaktadır. Bu yöntemlerin kanser teşhisinde daha faydalı olacağı aşikardır. Bu çalışmada da veri kümesi ve kanser türünden bağımsız olarak kanser teşhisi yapılmaya çalışılmıştır. Bu amaçla ve yöntemlerin geçerliliğini sınamak üzere farklı makine öğrenmesi ve örüntü tanıma yöntemleri iki ayrı veri kümesi üzerinde test edilerek karşılaştırılmıştır.

Kanser teşhisi için kullanılan proteomik verileri yüksek boyutlu olup verilerin ilk haliyle kullanılması oldukça zordur. Yüksek boyutun getirdiği problemlerden kurtulamak ve öznitelik çıkarmak için literatürde kullanılan bir çok çalışma incelenerek kullanılan yöntemlerden başarılı olanlar bu çalışmada denendi. İlk olarak ön işleme işlemlerinin devamı olacak şekilde t-test kullanıldı ve istatistiksel olarak anlamsız veriler ayıklandı. Böylece veri boyutunun biraz daha azalması sağlandı. Öznitelik çıkarımı aşamasında literatürde daha önce de kullanılan dalgacık analizi ve istatistiksel hesaplamalar bu çalışmada kullanıldı. Bu yöntemler ile karşılaştırmak üzere ilk defa bu çalışmada önerilen birtakım istatistiksel hesaplamalar kullanıldı ve yüksek başarımlar elde edildi. Öznitelik çıkarım yöntemlerinden sonra öznitelik seçimi için literatürde kullanılan ve başarılı sonuçlar veren çekirdek kısmi en küçük kareler yöntemi kullanıldı.

Özellikle öznitelik çıkarımı ve seçimi üzerinde durulan bu çalışmada farklı sınıflayıcılar kullanılarak öznitelik çıkarma yöntemleri karşılaştırıldı. Bu amaçla çekirdek kısmi en küçük kareler yöntemi ile farklı boyutlarda öznitelik alınarak k-en yakın komşuluk (kNN), destek vektör makineleri (DVM) ve lineer diskriminant analiz (LDA) ile sınıflandırıldı. Yapılan denemeler sonucu yumurtalık kanserinin belirlenmesinde en iyi sonuç %95,3 ile dalgacık dönüşümü ve LDA'nın birlikte kullanımı ile elde edildi. Prostat kanserinde, kullanılan veri kümesindeki farklılıktan

dolayı ilk önce tümörun (malign veya benign) var olup olmadığı belirlenmeye çalışıldı. Bu kısımda en yüksek başarı %97,3 ile bu çalışmada önerilen istatistiksel öznitelik çıkarımı ve kNN ile elde edildi. Daha sonra da tümörlü örnekler malign veya benign olarak sınıflandırıldı. Bu kısımda da en yüksek başarım %88,9 ile önerilen istatistiksel öznitelik çıkarımı ve kNN ile elde edildi.

PROSTATE AND OVARIAN CANCER IDENTIFICATION BY ANALYZING MASS SPECTROMETRY DATA

SUMMARY

Cells are the basic structural and functional units of the living organisms. All cells have the ability of proliferating under some control mechanism. When this control mechanism loses its function, cells start to divide and grow uncontrollably which leads the formation of tumors. Tumors can be categorized into two groups. The first group is called as benign tumors, which do not invade neighboring tissues and do not spread throughout the body. While the second group is called as malign tumors that can spread by the lymphatic system or bloodstream and thus can affect more distant parts of the body. These kinds of tumors are called as cancerous tumors.

The early diagnosis of cancerous tumors has vital importance for a successful treatment process. For instance five year survival rate is 92.1% for stage-1 ovarian cancer, whereas this rate decrease until 11.6% for stage-5. As well as in some cancer types the symptoms cannot be strong (ovarian cancer, prostate etc.), in case of that early diagnosis may not be possible. Generally, imaging systems are used for this purpose by performing an inner body scan, but the low specificity and sensitivity results of these methods are not still reliable enough to decide whether a cancerous tumor in its early stage exists or not. So, in most cases it is not possible to diagnose tumors, until they have already invaded surrounding tissues and metastasized throughout the body. This necessitates the need of different techniques for early diagnosis of cancer.

In recent years different methods have concentrated upon early diagnosis. Searching tumors in sputum and bronchoscopy for breast cancer, endoscopy for gastric cancer, looking a substance in blood which is called CA-125 for ovarian cancer and PSA (prostate specific agent) for prostate cancer are some of these methods. However all of these methods cannot give a satisfactory result. Another method is analyzing proteomic patterns with mass spectrometry which can be used for many types of cancer. Furthermore this technique excels the mentioned methods with its high accuracy and easily applicability.

Recently, mass spectrometry (MS) analysis of proteomics patterns has emerged as a new technology for the early diagnosis of cancer. In this method, a serum proteome (entire set of proteins in a serum sample) is first cleaved into small peptides, whose absolute masses are then measured by the mass spectrometer. These masses are then compared to the databases which are containing the known protein sequences. Thus, a mass spectrometry profile of the related sample is created. But note that, mass spectrometry in itself is not a diagnostic tool. In order to diagnose a disease, the obtained mass spectrometry profile must be analyzed by several computational methods. After an analysis, disease related biomarkers (proteins) are identified.

Mass spectra, is a high dimensional data which consist of tens of thousands of m/z ratios and an intensity level for each m/z ratio. Currently, a low resolution SELDI-TOF MS (Surface Enhanced Laser Desorption/Ionization Time of Flight Mass Spectrometry) can measure up to 15500 data points that record data between 500 and 20000 m/z ratios. With a high resolution MS, the data points could be 400000.

The high dimensionality of the MS data brings some difficulties for computational methods which are known as the "curse of dimensionality" and the "curse of data sparsity". To address these problems before analyzing the MS data a dimensionality reduction stage should be performed. Three methods are used for this purpose: filtering, wrapper and embedded methods. Filtering methods use some statistical tests to evaluate features, such as the t-test, Wilcoxon test, Mann -Whitley test and Kolmogorov-Smirnov test. After applying one of these statistical tests to the data, a score is obtained for each point (feature). According to the obtained scores, statistically insignificant points are extracted from the data by setting a threshold value. One of the weaknesses of filtering methods is that, they consider all features individually and ignore the interactions between the features. Therefore, after a filtering process generally the obtained data will have highly correlated and thus redundant features, which will worsen the classification performance. Even though the filtering methods have the above mentioned disadvantage, they are still preferred as an initial dimension reduction step in many studies. In wrapper methods, dimension reduction process is integrated into the classification stage. In these methods, a subset of features are first selected with an algorithm and then classified with a classification method. According to the obtained classification error, the feature selection algorithm updates its parameters until the optimum subset of features is found. Since the dimensionality is high, usually a stochastic algorithm such as, genetic algorithm, particle swarm optimization and ant colony optimization is used for this purpose. The main disadvantage of this method is the computational load of the search algorithms. As in the wrapper methods, embedded methods also integrate the feature selection process with the classification stage. Moreover, their computational load is less, when compared to the wrapper methods. Therefore, they are sometimes preferred to the wrapper methods.

Dimension reduction (feature selection and extraction) methods are not restricted to the above mentioned traditional methods for the MS data. Recently, wavelet analysis and statistical methods are used for this purpose. In the former one, the discrete wavelet transform (DWT) is applied to the MS data and approximation coefficients are obtained. Since the approximation coefficients represent the low frequency components, the obtained signal has a smoother form of the MS data with a low dimensionality. While in the latter one, the MS data is first divided into intervals and some statistical moments are then computed for the segments represented by these intervals. Both the wavelet analysis and interval based methods mentioned above, use filtering methods (such as t-test) as an initial dimension reduction step.

In this study, a three stage dimension reduction strategy is proposed for prostate and ovarian cancer classification from the MS data. The initial stage consists of a filtering method (t-testing), while in the second stage four different methods, wavelet analysis, statistical method-1, statistical method-2 and statistical method-3 are used for comparison. First method, wavelet analysis is very effective tool for feature reduction process where it is commonly used in the literature for this aim. The

second method, statistical method-1 is based on some statistical features which are used in the literature. The third and fourth methods, statistical method-2 and statistical method-3 are firstly used in this work in the view of feature extraction from proteomic patterns. The statistical method-2 is used before in literature for Electroencephalograpy (EEG) signals classification. The statistical method-3 is evolved version of statistical method-2 where a feature extraction method is changed with better one. In the last stage, a feature selection (transformation) method, kernel partial least square (KPLS) is used. KPLS is preferential method for feature selection due to its high speed and performance. KPLS is kernel based, iterative and supervised method so it can provide better performance and speed according to unsupervised methods, such as Principal Component Analysis (PCA).

After the three stage dimension reduction process, the MS data are classified with k-nearest neighbor classifier (k-NN), support vector machines (SVM) and linear discriminant analysis (LDA). For the high-resolution ovarian cancer dataset, an accuracy of 95.3% is obtained with a combination of wavelet analysis and KPLS methods. The prostate cancer classification is handled in two phases. In the first phase, the low resolution prostate MS data are classified as normal and cancerous samples with an accuracy of %97.3. While in the second phase, the data are classified whether the samples are benign or malign with an accuracy of 88.9%. Here, the best results are obtained with a combination of statistical method-3 and KPLS.

1. GİRİŞ

Vücudumuzda tüm organlar hücrelerden oluşmaktadır ve hücreler vücudumuzun en küçük yapıtaşlarıdır. Değişik etkiler altında hücreler değişime uğrayabilmekte ve kontrolsüz bir şekilde çoğalabilmektedir. Oluşan bu yeni hücrelere tümör adı verilir. Tümörler iyi huylu (bening) ve kötü huylu (malign) olmak üzere iki sınıfa ayrılır. İyi huylu tümörler genellikle kanser olarak adlandırılmazlar. Komşu bölgelere ve vücuda yayılmazlar, sınırları belirgindir. Orijinlerini tahmin etmek mümkündür ve vücuttan tamamen çıkarıldıklarında genellikle tekrarlamazlar. Kötü huylu tümörler ise kanser olarak adlandırılır. Lenf ve kan yoluyla vücuda yayılabilirler (metastaz).

Kanser teşhisinin konulabilmesi için ilk başvurulan işlem, görüntüleme yöntemleri (ultrasonografi, tomografi…) ile doku farklılığının araştırılmasıdır. Daha sonra kesin tanı için çoğunlukla biyopsi yapılmaktadır. Bu yöntem ile doku örneğinin incelenmesi sonucunda hastanın kanser olup olmadığı kesinlik kazanır. Ne var ki bu işlem tümörün iyi huylu olup olmadığını anlamak için de gereklidir. Bazen görüntüleme yöntemleri ile elde edilen verilerin net olmaması sebebiyle tümörün bulunmadığı durumlarda bile tedbir amaçlı biyopsi yapılabilmektedir. Bu işlem hasta için sıkıntı verici bir durum olduğu kadar gereksiz zaman, işgücü ve para harcanmasına sebep olmaktadır.

Kanser hastalığında erken teşhis yaşam süresi ve kalitesini doğrudan etkilemektedir. Bu sebeple tanının henüz hastalığın başlangıç aşamasında konulabilmesi için yoğun çalışmalar yapılmaktadır. Akciğer kanseri için balgamda tümör hücrelerinin aranması, bronkoskopi (ucu ışıklı bir boruyla solunum yollarının incelenmesi) uygulanması, mide kanseri için endoskopi, yumurtalık kanseri için CA-125 ve prostat kanseri için PSA isimli antijenlerin araştırılması, meme ve rahim ağzı kanseri için düzenli kontroller bunlardan bazılarıdır (**url1**). Son yıllarda kullanılmaya başlanan diğer bir yöntem ise kütle spektrometresi (mass spectrometry) ile kanın analiz edilmesi sonucu birtakım biyobelirteçlerin (biomarker) araştırılmasıdır (Ceccarelli ve diğ.,2009). Bu yöntem ile farklı kanser türlerinin teşhis edilmeye çalışıldığı bir çok çalışma bulunmaktadır. Datta ve DePadilla (2006) yumurtalık

kanserini kümeleme algoritmaları ile %87 ve sınıflama algoritmaları ile %98 lere varan başarı oranlarıyla sınıflamıştır. Yine yumurtalık kanserini Neves ve diğ. (2010) ICA ve NN kullanarak %99, Tang ve diğ.(2010) t-test, bazı istatistikler ve KPLS kullanarak %99,3 gibi yüksek oranlarla sınıflayabilmiştir. Ben-Dor ve diğ. (2000) çalışmalarında kNN ve SVM kullanarak yumurtalık (%89.3) , kolon (%88.7) ve kan (%95.8) kanserleri üzerinde yüksek başarımlar elde etmiştir. Li ve diğ. (2006) t-test ve KPLS ile kolon kanserinde %91.9 başarı elde etmişken, Krishnapuram ve diğ. (2004) RVM ve SVM ile %96.8 başarı elde etmiştir. Buna rağmen Li ve diğ. (2006) önerdikleri method ile göğüs kanseri teşhisinde %100 başarı sağlamıştır.

Bazı kanser türlerinde (rahim kanseri, kolon kanseri gibi) belirtiler belirgindir ve erken tanı mümkündür. Prostat ve yumurtalık kanseri gibi bazı kanser türlerinde belirtiler kuvvetli değildir, bu nedenle erken tanı oldukça zordur. Genellikle de erken dönemlerde hiçbir şikâyete yol açmadığından tanı konulduğu zaman bahsi geçen hastalıkların ilerlemiş olduğu görülmektedir. Bu da ölümle sonuçlanan vakaların sayısının artmasına sebep olmaktadır. Örneklendirmek gerekirse, erken tanı ile tedavisi %92.1 oranla mümkün olabilen yumurtalık kanserinin ilerlemesiyle bu tedavi olasılığı %11.6' lara kadar düşmektedir (**url2**). Prostat kanserinde ölüm oranı yumurtalık kanseri kadar yüksek olmasa da etkilediği kişi sayısı bakımından oldukça önemlidir (**url3**). Bu da çok sayıda insanın yaşam kalitesinin düşmesi anlamına gelmektedir. Daha önce de bahsedilen, yumurtalık kanserinde erken teşhis için kanda aranan CA-125 bileşeninin doğru sonuç verme olasılığı %35 kadar olup oldukça düşük bir orandır (Morton, 2006). Yine prostat kanserinde kanda aranan PSA bileşeninin kesin sonuç veremediği de bilinmektedir(Göğüş, 2006). Bu sebeplerle yumurtalık ve prostat kanseri için alternatif erken tanı yöntemlerinin geliştirilmesine de ihtiyaç vardır. Bu nedenle bu çalışmada yumurtalık ve prostat kanserine ait proteomik verileri kullanıldı. Aynı zamanda çalışmada önerilen yöntemin geçerliliğini sınamak üzere iki veri kümesi kullanılması düşünüldü.

Proteomik verilerinin sınıflandırılması için literatürde kullanılan ve başarılı sonuçlar veren iki yöntem bu çalışmada kullanılarak karşılaştırıldı ve bir başka yöntem önerilerek daha yüksek başarımların elde edilmesi sağlandı.

Bölüm 2 de ilk olarak KS cihazları tanıtılarak bu çalışmada kulanılan verilerin elde edilme aşamalarına değinilmiş, daha sonra ham verilere uygulanan ön işleme adımları anlatılmış ve son olarak da proteomik verilerin analizinde kullanılan

yöntemlerden bahsedilmiştir. Bölüm 3 de, öznitelik çıkarma, öznitelik seçme ve sınıflandırma işlemleri tanıtıldıktan sonra kullanılan algoritma verilmiştir. Son olarak, 4. Bölüm de kullanılan veri kümeleri üzerinde elde edilen sonuçlar verilip değerlendirilmiş ve ileride yapılabilecek çalışmalar hakkında fikirler sunulmuştur.

2. KÜTLE SPEKTROMETRESİ ve PROTEOMİK VERİLERİNİN ÖN İŞLENMESİ

2.1 Kütle Spektrometresi

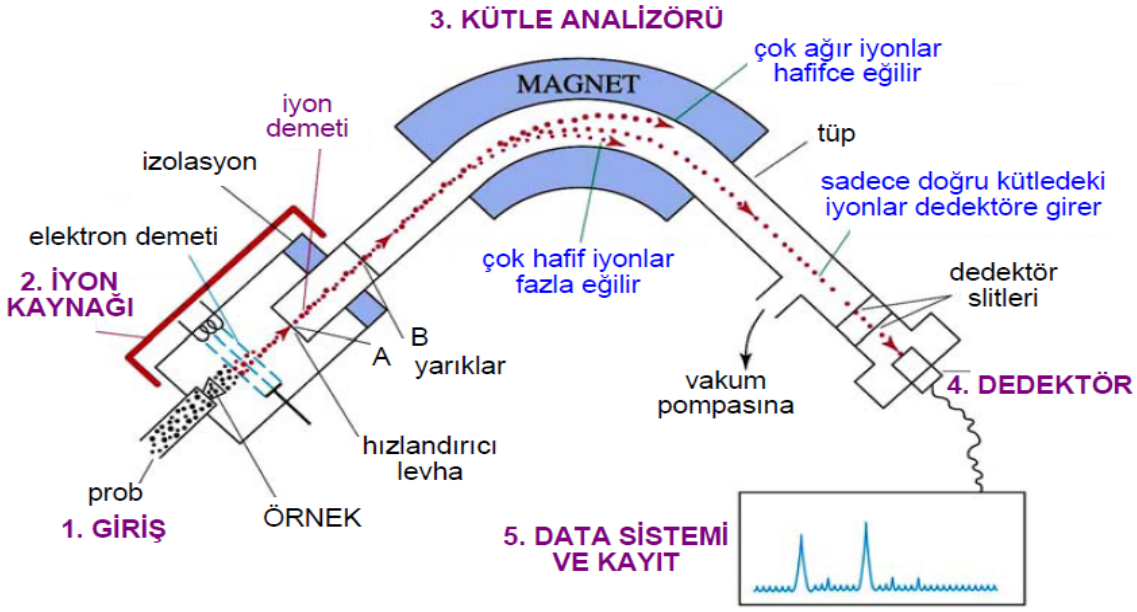
Kütle spektrometreleri ile manyetik veya elektriksel bir alanda hareket eden yüklü parçacılar, kütle/yük (mass/charge, m/z) oranına göre diğer yüklü parçacıklardan ayrılarak analiz edilmektedir. İlk kez 1913 yılında J.J. Thomson tarafından Neonun iki izotopunun olduğunun gösterilmesiyle bu yöntemin temelleri atılmıştır. Teknolojik gelişmeler hassas kütle spektrometre cihazlarının üretilmesini, kimya, fizik vb. alanlardan sonra tıp alanında da kullanılmasını sağlamıştır. Tıp alanında ilk olarak metabolik hastalıkların tanısında (Rashed ve diğ., 1995), protein (Belov ve diğ., 2000), lipit (Vreken ve diğ., 2000) karbonhidrat (Berry ve diğ., 1995) ve DNA (Banks ve diğ., 1994) gibi birçok bileşiğin analizinde kullanılmıştır.

Normal koşullarda herhangi bir yük taşımayan moleküller, kütle spektrometresi ile iyonize edilir ve daha sonra bu moleküller elektriksel bir alandan geçirilir. Bu işlemlerin sonucunda moleküllerin kütle/yük (m/z) oranına karşılık yoğunluk grafikleri elde edilir. Bilinen bir örneğin grafiğinden yola çıkılarak yeni örnekler hakkında fikir sahibi olunur.

Şekil 2.1'de örnek bir KS cihazı için blok diyagram verilmiştir. Giriş kısmından analiz edilecek örnekler sisteme verilir. Bu örnekler 2. kısımda iyon kaynağı ile bombardıman edilerek iyonize hale getirilir. Daha sonra hızlandırılan iyonların manyetik bir alandan geçmesiyle (3. Kısım, Kütle Analizörü) farklı yollar izlemesi sağlanır. Vakumlanmış (İyonların havadaki moleküller ile çarpışmasını engellemek için uygulanmaktadır.) tüp içerisinden ilerleyen iyonlar detektör tarafından algılanır. Son olarak elde edilen veriler uygun bir şekilde kaydedilip saklanır.

Şekil 2.1 : Kütle spektrometresi blok diyagramı (url4).

KS cihazları için alternatif yöntem ve tasarımlar kullanılmaktadır. Örnek bir KS cihazı için detaylı bir gösterim Şekil 2.2'de verilmiştir. 5 temel kısımdan oluşan KS cihazları için çok sık kullanılan teknolojiler sonraki kısımda ayrı ayrı ele alınmıştır (**url5**).



Şekil 2.2 : Kütle spektrometre cihazının bölümleri (url4) .

2.1.1 Örnek girişi

Örneklerin sisteme verildiği bu kısımda örneklerin hangi fazda olduğu (katı, sıvı veya gaz) önemlidir. Sisteme verilen örneklerin fiziksel haline göre uygulanacak yöntemlerle gazlaşması sağlanarak hareket kabiliyeti arttırılır.

2.1.2 İyon kaynağı

Bu kısımda, örnek girişinden gelen moleküller iyonize edilir. Bu işlem için uygulanan bazı iyonizasyon yöntemleri şunlardır (Bramer, 1998):

Uçucu bileşikler için:

Elektron İyonizasyon (Electron Ionizaiton-EI): Basit ve en çok kullanılan iyonizasyon tipidir. Bu yöntemle bir flamentten akım geçirilerek filamentin etrafına elektron saçması sağlanır. Bu elektronların molekülerle çarpışması sonucu iyonizasyon sağlanır.

Kimyasal İyonizasyon (Chemical Ionizaiton-CI) : Bu yöntemde metan gibi bir reaktif gaz, 1 torr kadar basınçta, özel bir elektron demeti kaynağı içine konulur.

Aynı kaynak içine, konsantrasyonu reaktifin 10^{-3} –10^{-4} katı olacak şekilde örnek ilave edilir. Burada gerçekleşen kimyasal olaylar neticesinde örneğin iyonlaşması sağlanır. Burada seçilecek teknik ve kimyasallar örneğe göre değişiklik göstermektedir.

Uçucu olmaya bileşikler için:

Hızlı Atom Bombardımanı (Fast Atom Bombardment-FAB) : Bu yöntemde Ar, Xe gibi inert (kimyasal olarak aktif olmayan madde) gazlara tanecik demeti olarak ihtiyaç duyulmaktadır. Analit/Matriks adı verilen organik moleküller ile örneğin yüzeyinin homojen olması sağlanır ve hızlı tanecik demetiyle bombardıman edilir. Analit/matriks yüzeyi üstüne gelen enerjiyi örneğe transfer eder ve bu şekilde iyonizasyon gerçekleşir.

Atmosferik Basınç İyonizasyon(Atmospheric Pressure Ionizaiton-API): Özellikle Sıvı Kramatografisi-Kütle spectrometresi (LC/MS) sistemi için kullanılan yaygın bir iyonizasyon yöntemdir. Örneğin çözelti halinde olması gereken bu sistemin Elektrosprey İyonizasyon(Electrosprey Ionizaiton -ESI) ve Atmosferik basınç kimyasal iyonizasyon (Atmospheric Pressure Chemical Ionizaiton) şeklinde iki tipi vardır. Temelde her iki sistemde örneğin sprey şeklinde sıkılması (küçük damlalar halinde) ve daha sonra bunların iyonize hale getirilmesi prensibine dayanır.

MALDI (Matrix Assisted Laser Desorption Ionization): Hızlı atomik bombardıman (FAB) iyonizasyonuna benzer. Matriks içinde çözülen örneğe gelen lazer ışınları sayesinde iyonizasyon sağlanır.

SELDI(Surface Enhanced Laser Desorption Ionization): MALDI tekniğine benzeyen bu yöntemde örnekler bir özel yüzeyin üzerine konur ve birtakım işlemler sonrasında yüzeye gelen lazer ışınları ile iyonizasyonun geçekleşmesi sağlanır. Genellikle Uçuş zamanlı analizörler bu sistemle beraber kullanılır.

2.1.3 Kütle analizörü

Farklı prensiplere dayanarak çalışan kütle analizör cihazları bulunmaktadır. Bunlardan bazıları (Biberoğlu G., 2003):

Manyetik analiz metodu (Magnetic Analysis Methods): İyon kaynağında oluşan farklı m/z oranına sahip iyonlar aynı manyetik alanda farklı yollar izler (kütlelerinin

de farklı olması göz önünde bulundurularak) ve detektör üzerinde farklı noktalara düşer.

Quadrupole Analizör: Dört adet paralel metalik çubuğa gerilim uygulanması sayesinde radyo frekansları ile özel bir titreşim alanı oluşturulur ve iyonlar detektöre yönlendirilir.

Uçuş Zamanlı Analizörler (Time of Flight-TOF): Bu yöntem, üretilen iyonların kütlelerindeki farklılıktan dolayı sabit mesafeyi farklı zamanlarda almaları esasına dayanır.

İyon Kapanları (Ion traps): Quadrupole analizörün modifiye edilmiş bir şeklidir. Bu sistem ile iyonizasyon ve kütle analizi aynı yerde gerçekleşmektedir.

2.1.4 Detektör

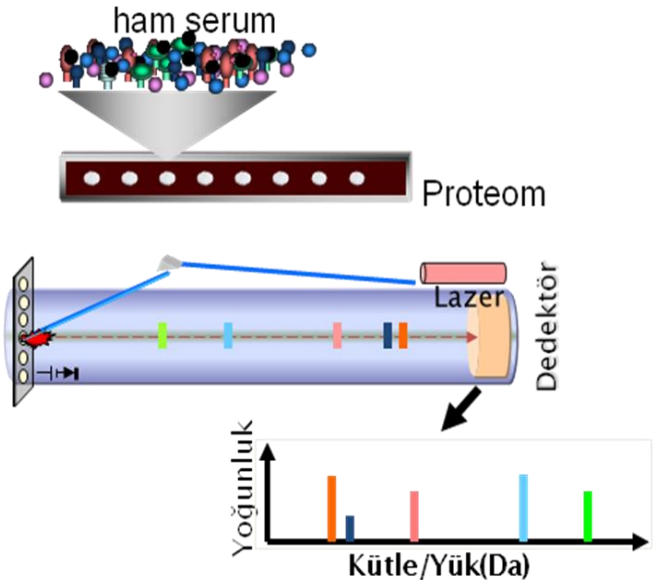
Algılanması istenen bileşenlere bağlı olarak geliştirilmiş çok çeşitli kromatografi detektörü vardır Bu cihazlarda kullanılan algılayıcılar aynı zamanda iyon detektörleridir. İyonizasyon odaları, orantılı sayıcılar, Geiger-Mueller tüpleri, nötron sayıcılar ve sintilasyon sayıcıları gibi detektörler iyon detektörleridir. Ayrıca gaz kramatografisi detektörlerin çoğu (kullanılan iyonizasyon tekniğine bağlı olarak) iyon detektörüdür. Örneğin, argon, helyum, elektron yakalama, alev iyonizasyon gibi detektörler bu tip detektörlerdir (**url6**).

2.1.5 Data sistemi ve kayıt

Ölçülen büyüklüklerin oldukça küçük boyutlarda olması (tek bir iyonun detektörde oluşturduğu akım: 10^{-17} -10^{-9}A mertebelerinde) ve çok sık aralıklarla elde edilmesi (saniyede 60 kadar iyonun detektöre ulaşması) detektörler kadar veri kayıt sistemini de önemli hale getirmektedir. Ayrıca elde edilen sinyaller bir dizi işlemden geçirildikten sonra (iyonları kütlelerinin belirlenmesi, sinyallerin normalize edilmesi ve referans değerler ile karşılaştırılması vs.) grafiksel (yük/kütle değerine karşılık yoğunluk olacak şekilde) hale getirilir. Bu aşamadan sonra veriler yorumlanması ve değerlendirilmesi mümkün olmaktadır.

2.2 Proteomik Verilerin Ön İşlenmesi

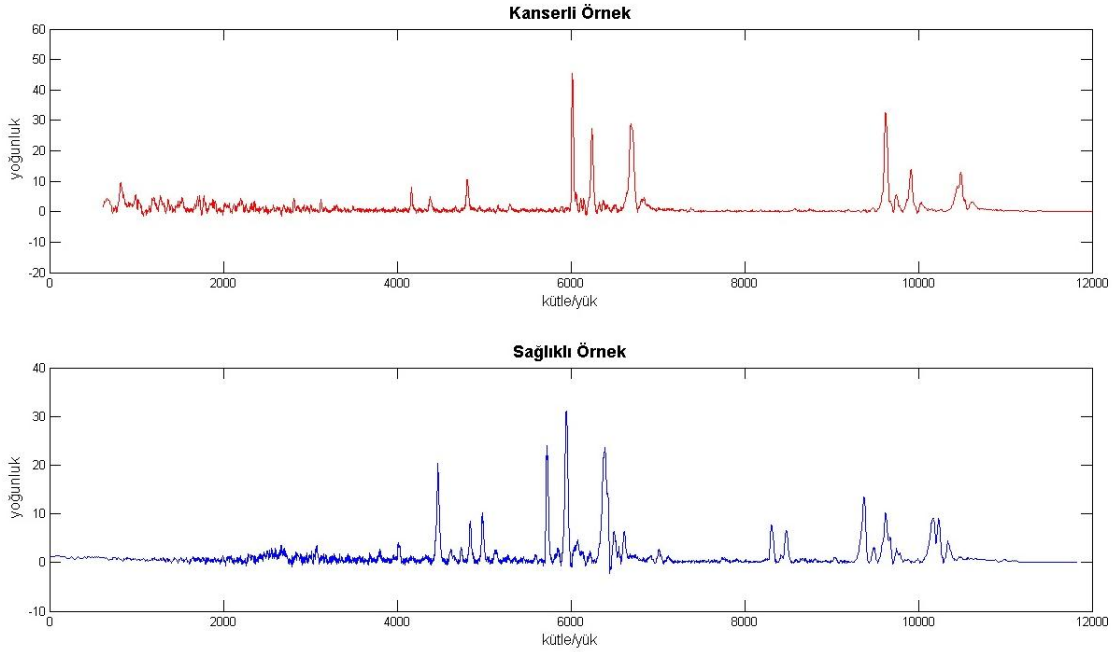
Bu çalışmada elde edilen veriler SELDI-TOF MS (Yüzey Güçlendirmeli Lazer Dezorpsiyon/İyonizasyon Uçuş Zamanlı Kütle Spektrometresi) kullanılarak alınmıştır. Bu KS cihazı için Şekil 2.3'de temsili bir gösterim bulunmaktadır. Bu sistem daha önce açıklanan SELDI ve TOF sistemlerinin birleşimi olarak düşünülebilir. Bu sistem ile veriler yüksek çözünürlükle elde edildiğinden veriler yüksek boyutlu olmakta, bu boyut çoğunlukla yüz binlerle ifade edilmektedir. Elde edilen verilerin bu haliyle analiz edilmesi oldukça zordur.



Şekil 2.3 : SELDI-TOF MS.

Şekil 2.4'de FDA-NCI klinik proteomik programı (FDA-NCI Clinical Proteomics Program Databank) [**url7**] veritabanından alınan prostat kanserli ve sağlıklı bireylere ait birer kütle spektrometresi örneği görülmektedir.

Öznitelik çıkartma işleminden önce kütle spektrometresi verilerinin, sınıflama başarımını arttırmak için bir ön işlemden geçirilmesi gerekmektedir. Bu amaçla MATLAB (Bioinformatics Toolbox)'tan faydalanılmıştır. Ön işleme adımlarına aşağıda kısaca değindikten sonra proteomik verilerinin kanser teşhisi için kullanıldığı çalışmalara ve bu çalışmalarda kullanılan yöntemlere değinilecektir.



Şekil 2.4 : Örnek kütle spektrometre verisi.

2.2.1 Yeniden örnekleme

Yeniden örnekleme işlemi ile spektrum içerisindeki gereksiz bilgi ayıklanmaktadır. Bu işlem kütle/yük oranını homojenize ederek farklı spektrumları aynı çözünürlükte karşılaştırma imkanı tanır. Ayrıca yüksek çözünürlüklü kütle spektrometresinin boyutu oldukça fazla olduğu için (>370000) yeniden örnekleme işlemi ile hesapsal yük de önemli ölçüde azalmaktadır. Öyle ki, yeniden örnekleme işlemi sonucunda spektrum boyutu 15000'e kadar düşmektedir.

2.2.2 Taban hattı kaymasının giderilmesi

Veriler üzerinde, kimyasal gürültülerden veya aşırı iyon yüklemesinden kaynaklanan taban hattı kayması düzeltilmektedir. Bu amaçla, alçak frekanslı taban hattı gürültüsü elde edilerek orijinal spektrumdan çıkartılmaktadır.

2.2.3 Spektrum hizalama

Kalibrasyondan kaynaklanan hatalardan ötürü, elde edilen m/z oranları ile iyonların gerçek uçuş zamanları arasında farklılıklar gözlemlenebilir. Bu ise farklı deneyler arasında spektrumda kaymalara neden olmaktadır. Spektrumda görülen bu kaymaları gidermek için diğer bir ön işlem olarak spektrumlar hizalanmaktadır. Bu amaçla

spektrum içerisinde belirli sayıda referans tepe değerleri belirlenir ve bütün veri kümesi bu referans değerler baz alınarak hizalanır.

2.2.4 Normalizasyon

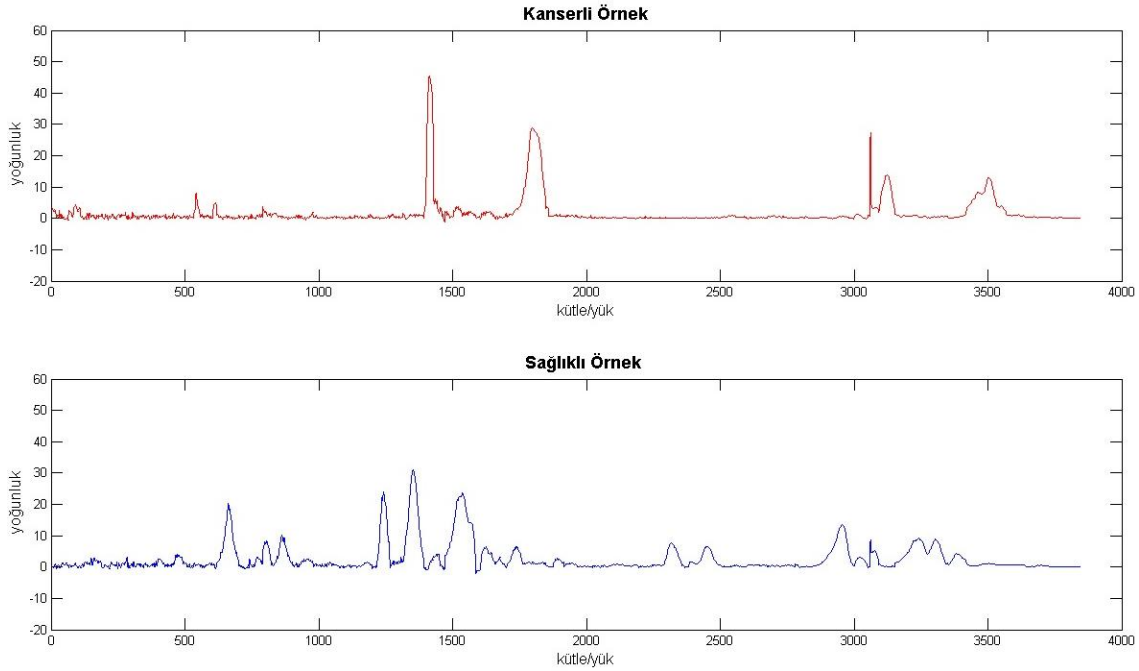
Farklı deneyler arasında, ayrıştırılan ve iyonize olan proteinlerin miktarından kaynaklanan sistematik farklar görülebilir. Bu farklardan kaynaklanan hataları en aza indirgemek için veri kümesi üzerine normalizasyon işlemi uygulanmaktadır. Bu amaçla her spektrum, maksimum yoğunluğu 100 olacak şekilde normalize edilmiştir. m/z < 1000 Da olan düşük kütleli bölgelere genelde sorun teşkil ettiği için dokunulmamıştır.

2.2.5 Boyut indirgeme

Ön işlemden geçirilen verilere t-test uygulanarak istatistiksel olarak gereksiz noktaların veri kümesinden ayıklanması hedeflenmiştir. Student's t-test olarak da bilinen bu yöntem sıklıkla boyut indirgeme amacıyla kullanılmaktadır. Farklı sayıda örnek içeren iki veri kümesini karşılaştırmak için oldukça kolay ve faydalı bir yöntemdir. Her bir öznitelik x_j için, sırasıyla μ_j^+ ve μ_j^- 1.sınıf (kanser, pozitif) ve 2.sınıf (sağlıklı, negatif) ortalamaları hesaplanır. Aynı şekilde δ_j^+ ve δ_j^- standart sapmaları hesaplanır. n^+ ve n^- 1.sınıf ve 2.sınıfa ait örnek sayısını göstermektedir. Bu elde edilen değerler ile T skoru eşitlik **(2.1)** yardımıyla hesaplanmaktadır.

$$T(x_i) = \frac{\left|\mu_i^+ - \mu_i^-\right|}{\sqrt{\frac{(\delta_i^+)^2}{n_+} + \frac{(\delta_i^-)^2}{n_-}}} \qquad \textbf{(2.1)}$$

Kullanılan verideki örnek sayısına ve istenen hata payına göre t-tablosu yardımıyla minimum bir T skoru tanımlanır. Elimizdeki veri göz önüne alınırsa (yumurtlık kanseri için), toplam 216 örnek ve %5 hata payı için bu değer 2 olarak atanabilir. Eşitlik (2.1) yardımı ile hesaplanan T skorları 2 den büyük ise bunun elde edildiği öznitelikler tutulur, değilse gürültü kabul edilip atılır. Bu şekilde elimizdeki verinin boyutu 12000-15000' den 3000-5000'e kadar düşmüştür. Bu aşamada T skoru istenirse daha büyük seçilerek elde edilecek verinin boyutunun daha da düşürülmesi sağlanabilir. Şekil 2.5'de t-Test sonucu boyutu azaltılmış veri görülmektedir.



Şekil 2.5 : t-Test sonrası boyutu indirgenmiş örnek veri kümesi.

2.3 Proteomik Verilerin Analizi

Özellikle moleküler düzeyde hücrelerin incelenmesi ile ortaya çıkan yüksek boyuttaki verilerin analiz edilmesi ihtiyacı biyoinformatik alanın doğmasını sağlamıştır. Biyoinformatik genetik bilgiyi analiz etmek ve anlamak için bilgisayar bilimleri, bilgi teknolojileri ve genetik bilim dallarından oluşan bir kombinasyondur (**url8**). Genomik ve proteomik biyoinformatiğin en temel iki çalışma alanıdır. İnsan genom projesi ile tanınırlık kazanan genomik çalışmalar genomun haritalanması, sekanslanması (DNA dizimininin araştırılması) ve karakterizasyonu gibi konularla ilgilenmektedir. Genomiğin alt dalı veya eşdeğeri olacak şekilde tanımlanabilecek olan proteomik; proteinlerin sentezi, dizilimi, fonksiyonlarının incelenmesi ve bu sistemlerin modellenmesi gibi konularla ilgilenmektedir. Proteom ise bir organizma, bir doku veya bir hücre tarafından salınan tüm proteinler şeklinde tanımlanabilir.

Kanser teşhisi için de kandan elde edilen proteom örnekleri kullanılmaktadır. Proteomik analizlerin en başında hücreden veya organizmadan binlerce proteinin ayrıştırılması ve görüntülenmesi gelmektedir. Bu amaçla başta kütle spektrometresi (KS) olmak üzere proteomik ayrıştırma için birçok yöntem geliştirilmiştir. Bunlardan bazıları şunlardır [**url9**].

* Çip tabanlı (Chip-based) teknolojik analiz

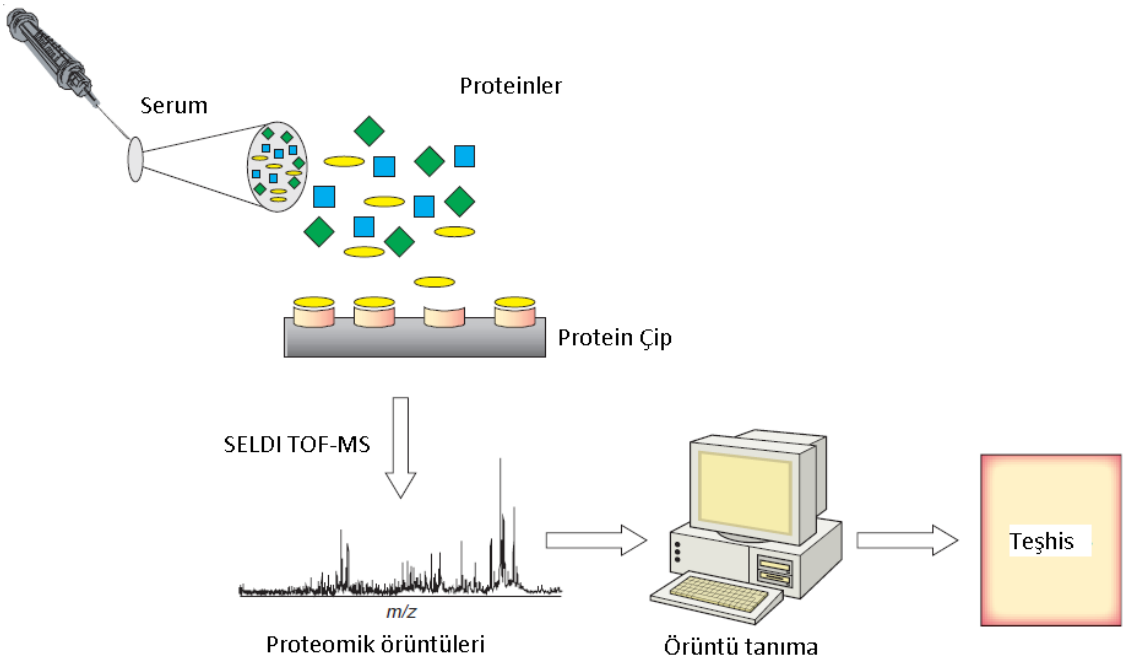
* Kütle (Mass) spektrometresi (MS)

* Afinite kromatografisi

* İkili Hibrit (Two-Hybrit) görüntüleme

* İki boyutlu jel elektroforezi ve izoelektrik odaklama

Bu çalışmada kullanılan veriler proteomik örneklerin kütle spektrometresi (KS) ile analiz edilmesiyle elde edilmiştir. Şekil 2.6'da bu verilerin elde edilişi ve işlenmesi bölümler halinde verilmiştir.



Şekil 2.6 : Proteomik verilerin elde ediliş aşamaları (Conrads ve diğ., 2003).

Proteomik örneklerin KS ile analizi işleminde, ilk olarak kan örneği alınarak bir dizi parçalama ve saflaştırma işlemlerinden geçirilir. Kullanılacak analiz yöntemine bağlı olarak proteomlar hazırlanır. Daha sonra kütle spektrometresinde analiz işlemi gerçekleştirilir. Elde edilen verilerin yüksek boyutlarda olması sebebiyle veriler ilk haliyle kullanılamaz. Ön işleme adı verilen kısımdan sonra biyobelirteç bulunmaya çalışılır. Bu işlemin tamamen otomatik yapılması asıl hedeflerden biridir. Bazı çalışmalarda biyobelirteç tayini belli noktaların seçimi ile yapılırken (Qu ve diğ., 2002; Banez ve diğ., 2003), bazı çalışmalarda biyobelirteclerin hangi noktada olduğundan ziyade teşhisin doğru yapılıp yapılamadığı önemlidir (Tang ve diğ., 2010). İlk yöntemde olduğu gibi biyobelirteclerin hangi noktada olduğundan yola çıkılarak yapılan teşhislerde yöntem sadece kullanılan veri kümesi üzerinde çalışır.

Bu sebeple literatürde farklı farklı biyobelirteçler önerilmiştir. Bunun başlıca sebepleri ise şunlardır (Yurdakul ve diğ., 2009):

- Farklı kütle / yük aralığı seçimi,
- Farklı çip tiplerinin kullanılması,
- Örnek büyüklüğü farklılıkları

Buradan hareketle kullanılan veri kümesinden bağımsız ve genel geçerliliği olan bir teşhis yöntemi ortaya koymak çok daha etkili bir yöntem olacaktır. Biyoinformatik alanına girilen bu kısımda veri analizi için kullanılan yöntemler daha çok makine öğrenmesi ve örüntü tanıma ile alakalıdır. Bunun yanında matematik ve istatistiksel yöntemlere de ihtiyaç duyulabilmektedir.

Elde edilen verilerin genellikle yüksek boyutlu olması ve az sayıda örnekten alınabilmesi modelleme aşamasında çok sayıda problemi de beraberinde getirmektedir (Molinaro ve diğ., 2005). Yine bu sebeple yapılan çalışmalarda genellikle vurgulanılan kısım, elde edilen verilerden öznitelik seçimi ve çıkarımı olmaktadır. Son yıllarda yoğunlaşılan bu alanda farklı boyut azaltma, öznitelik çıkarma, öznitelik seçme ve sınıflama yöntemlerinin kullanıldığı görülmektedir. Bu kısımda kullanılan yöntemler üç başlık altında incelenebilir. Bunlar; filtreleme yöntemi (filter), sargı (wrapper) yöntemi ve gömülü yöntem (embeded) olarak bilinmektedir. Bu yöntemler sınıflayıcı yapısıyla entegre edilme şekillerine bağlı olarak incelenmektedir.

T-test, Wilcoxon test, Mann-Whitley test ve Kolmogorov-Smirnov test gibi istatistiksel yöntemler filtreleme yöntemleri içinde değerlendirilen veri boyutunu indirgeme metodlarıdır. İstatistiksel testler sonucunda her öznitelik elemanına bir skor değeri atanır ve daha sonra bu skora göre bazı öznitelikler veri kümesinden ayıklanır. Oldukça basit ve kullanışlı olan t-test literatürde en sık karşılaşılan yöntemlerden biridir. Student t-Test adıda verilen bu yöntemi Zhu ve diğ. (2003), Wu ve diğ. (2003) ve Tang ve diğ.(2010) çalışmalarında kullanmıştır. Bunun yanında farklı istatistiksel yöntemlerde denenmiştir. F-testini Bhanot ve diğ. (2006), Ki-kare testini Liu ve diğ.(2003), Kolmogorov-Smirnov(Ks) testini Yu ve diğ. (2005) çalışmalarında kullamışlardır. Yöntemlerin bir dezavantajı, her özniteliği bağımsız bir şekilde düşünmesi, yani öznitelikler arasındaki olası ilişkileri göz ardı ediyor olmasıdır. Bu nedenle filtreleme işlemi sonrasında birbirleriyle yüksek korelasyona

sahip öznitelikler elde edilebilir (Liu ve diğ., 2009) . Bunun yanında basit ve hızlı olması, sınıflayıcıdan bağımsız olması (farklı sınıflayıcılar ile test etme imkanı verir) fitreleme yöntemlerinin en önemli avantajlarıdır (Pengyi ve diğ.,2003).

Bir diğer yöntem olan sargı yönteminde, öznitelik seçme işlemi sınıflama işlemi ile tümleşiktir. Bu yöntemde, seçilen öznitelik alt kümeleri bir sınıflayıcı ile sınıflanarak bir hata değeri elde edilir. Bu hata değeri baz alınarak optimum öznitelik elemanları bulunmaya çalışılır. Çözüm uzayı oldukça büyük olduğu için genetik algoritma, parçacık sürü optimizasyonu (PSO) ve karınca koloni optimizasyonu (ACO) gibi stokastik algoritmalar bu yöntemde sıklıkla kullanılmaktadır (Pengyi ve diğ.,2003; Blanco ve diğ., 2004; Inza ve diğ., 2004). Bu yöntemin dezavantajı ise hesapsal yükün oldukça fazla olmasıdır. Gömülü yöntem ise yine sargı yöntemi gibi öznitelik seçme işlemi ile sınıflama işlemini birlikte yapan bir yöntemdir (Geurts ve diğ., 2005; Jong ve diğ., 2004). Bu yöntemin hesap yükü sargı yöntemine kıyasla daha azdır. Dolayısıyla bu yöntem kimi zaman sargı yöntemine tercih edilmektedir.

Literatürde farklı öznitelik çıkarma yöntemleri ile farklı sınıflayıcıların kullanıldığını görmek mümkündür. Bunların bir kısmı gömülü ve sargı yöntemler ile birlikte kullanılan, bir kısmı da filtreleme yöntemi ile birlikte denenen yani öznitelik seçiminden bağımsız sınıflayıcılardır. Datta ve DePadilla (2006) LDA,QDA, Nnet,Svm ve RF'i, Tang ve diğ. (2010) KPLS'i, Wu ve diğ. (2003) kNN,Svm ve RF'i çalışmalarında sınıflayıcı olarak kullanmışlardır.

Yukarıda yapılan öznitelik seçimi yöntemlerini ve sınıflayıcıları danışmalı (supervised) ve danışmasız (unsupervised) yöntemler olarak da incelemek mümküdür. Hem öznitelikler için hem de sınıflayıcılar için her iki yöntemin de çalışmalarda kullanıldığı görülmektedir. LDA ve PLS danışmalı (Hilario ve Kalousis, 2008) , dalgacık analizi (Qu ve diğ.,2003) ve TBA (Taşkın ve diğ.,) kullanılan danışmasız öznitelik çıkarma yöntemlerinden bazılarıdır. K-means, Diana ve Fanny literatürde kullanılan ve kümeleme (clustering) olarak da bilinen danışmasız sınıflama algortimalarından bazılarıdır (Datta ve Datta, 2002). LDA, NN, kNNve SVM kullanılan danışmalı sınıflayıcılardan bazılarıdır (Neves ve diğ., 2010; Li ve diğ., 2006). Datta ve DePadilla (2006) farklı danışmalı ve danışmasız algortimaları birarada kullanmıştır.

Literatürde farklı veri kümeleri kullanarak farklı kanser türlerinin teşhis edildiği çalışmalar bulunmaktadır. Elde edilen başarımların karşılaştırılması her zaman mümkün olamamaktadır. Kanser türlerine göre sonuçlar değişebilmektedir. Bir kanser türünde önemli başarımlar elde edilirken aynı yöntemle farklı bir kanser türünde yeterince iyi sonuçlar alınamayabilmektedir (Li ve diğ., 2006). Kullanılan yöntem ve kanser türü kadar kullanılan veri kümeleri de önem arz etmektedir. Aynı kanser türünde farklı veri kümelerin üzerinde farklı sonuçlar elde edilebilmektedir. Kullanılan KS teknolojisindeki farklılıkların yanında örneklerin kanser dereceleri de bu farklılıkta oldukça etkilidir. Özellikle ilerlemiş kanser vakalarının teşhis edilmesi daha kolay olmaktadır (Petricoin ve diğ., 2002). Sınıflama için eğitim ve test kümeleri belirlenirken örnekler arasındaki farklılıklar (kanserin ileri seviyede olması), kümelerdeki örnek sayısı gibi etkenler aynı veri kümesi üzerinde farklı sonuçlar almaya sebep olabilir.

Bu çalışmada kullanılan veriler, proteomların yüksek çözünürlüğe sahip kütle spektrometresi ile analiz edilmesi ile elde edilmiştir. Bunun sonucunda da yüksek boyutlu veriler elde edilmiştir. Öznitelik seçme işlemi için sınıflayıcıdan bağımsız yöntemler tercih edilmiştir. Bu şekilde farklı öznitelik seçme işlemleri ve sınıflayıcılar ile elde edilen sonuçların karşılaştırılması mümkün oluştur. İki farklı kanser türünün teşhis edilemeye çalışıldığı bu çalışmada yeni bir takım istatistiksel yöntemler de denenmiştir. Bu işlemlerin sonucunda seçilen veri kümelerinde kanser teşhisi yüksek başarım oranlarıyla yapılabilmiştir.

3. PROTEOMİK VERİLERİ İÇİN ÖZNİTELİKLERİN BELİRLENMESİ ve SINIFLAMA SÜREÇLERİ

3.1 Öznitelik Çıkarma Yöntemleri

3.1.1 Dalgacık dönüşümü

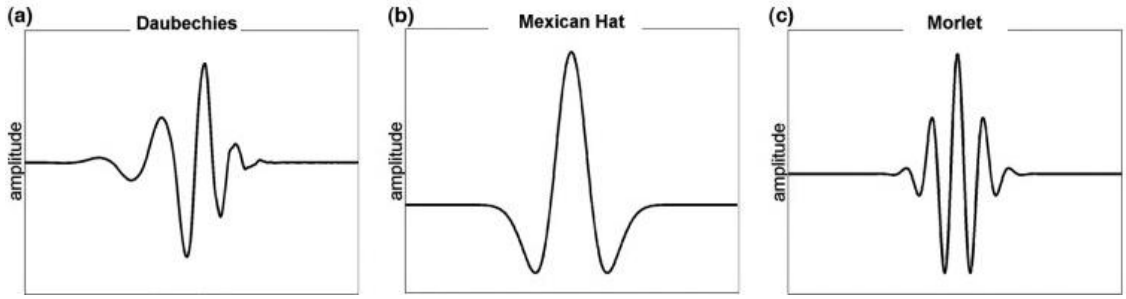
Dalgacık dönüşümü fourier dönüşümü gibi sinyallerin matematiksel bir yolla zaman domeninden frekans domenine geçirilmesi olarak düşünülebilir. Fourier analizinden farklı olarak ise frekans domenine geçişte zaman bilgilerinin de taşınmasıdır. Yani, bir işaretin fourier analizine baktığımızda bir frekans bileşeninin zamanın hangi anında ortaya çıktığını anlamak imkânsızdır. Eğer sinyal çok değişmiyorsa bu eksiklik çok önemli değildir. Bazı sinyallerde ise değişimler yalnızca işaretin bir kısmında oluşabilir ki bu kısımlar genellikle önemli parçalardır ve fourier analiziyle farkedilmesi mümkün değildir. Bu gibi durumlarda bir sinyalin yerel ve genel olarak incelenmesini mümkün kılacak frekans-zaman gösterimi oldukça faydalı olacaktır. KSFD (Kısa zamanlı Fourier Dönüşümü) ile bir ölçüde yapılmaya çalışılan bu gösterimin zayıf tarfı ise kullanılan pencere boyutlarının sabit olmasıdır. Dalgacık analizi ile bu eksiklik giderilerek pencere boyutunun değişken olması sağlanarak esnek bir zaman-frekans gösterimi mümkün hale gelmiştir.

Bir dalgacık sınırlı süreklilikte ortalama değeri 0 olan bir dalga şeklidir. Fourier analizinin ana fonksiyonu olan sinüs ile dalgacık dalgasını karşılaştırırsak şunları görürüz: Sinüs sınırlı sürekli değildir, eksi sonsuz ile artı sonsuz arasında sonsuz salınır, düzgün ve tahmin edilebilirdir. Dalgacık ise sonludur ve genellikle düzgün değildir,asimetrik yapıdadır.

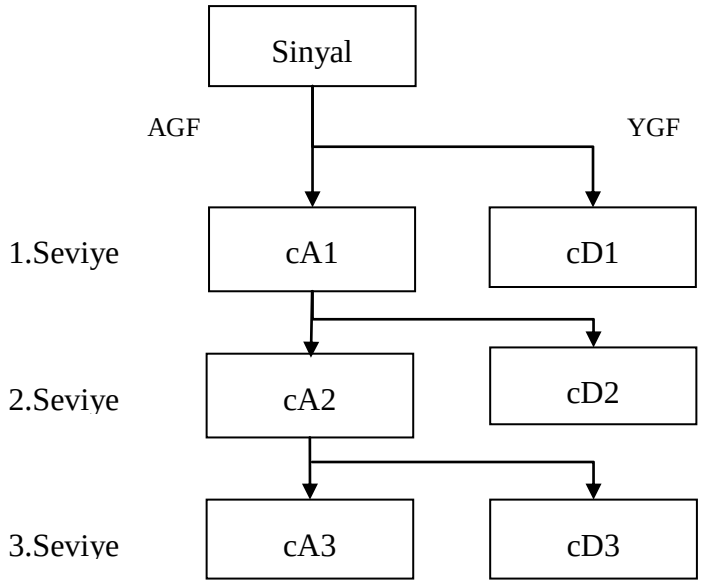
Dalgacık analizinde herhangi bir sinyali işlemek için bir ana dalgacık (mother wavelet) tanımlanmalıdır. Şekil 3.1'de çok sık karşılaşılan üç dalgacık örneği verilmiştir.

Bu çalışmada dalgacık dönüşümü sinyaldeki gürültülerden kurtulmak ve aynı zamanda boyut azaltmak için kullanılmıştır. Bu amaçla bir filtreleme yöntemi olarak

da kullanılabilen Ayrık Dalgacık Dönüşümü (ADD) bu çalışmada kullanılmıştır. Bu işlemin adımları Şekil 3.2'deki gibi temsil edilebilir.

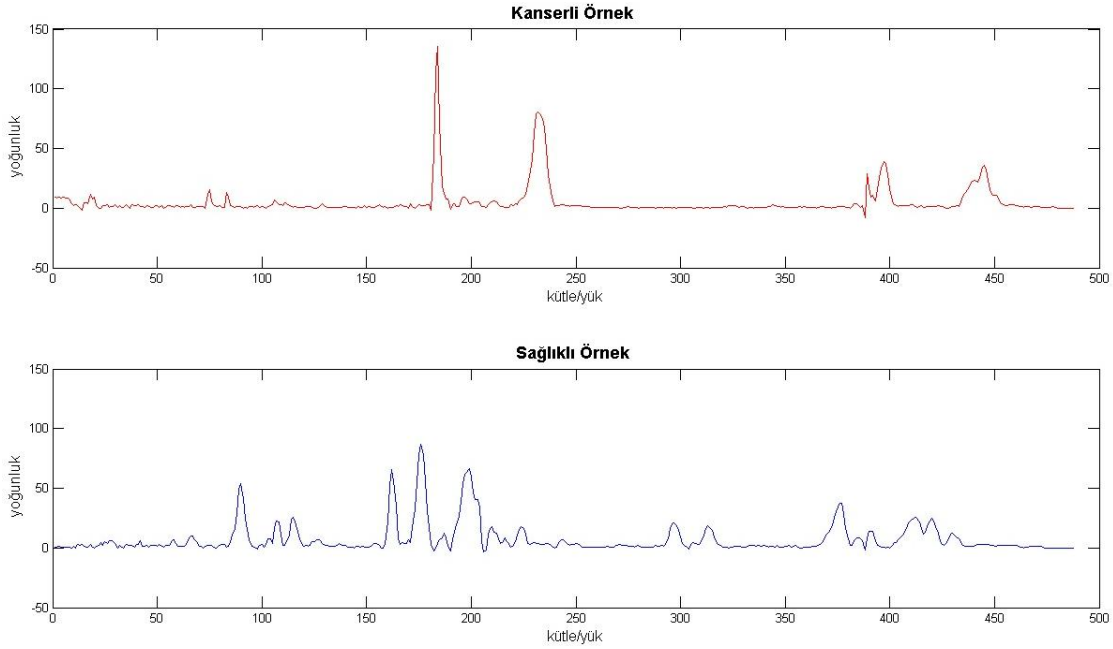


Şekil 3.1 : (a) Daubechies , (b) Mexican Hat ve (c) Morlet dalgacık fonksiyonları.



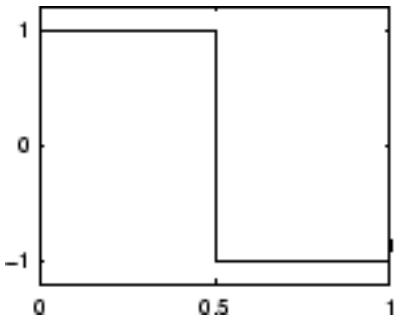
Şekil 3.2 : ADD adımları.

Bir çok sinyalde düşük frekans bileşenleri en önemli bölümdür. Bu kısım işaretin ana fikrini verir. Diğer taraftan yüksek frekans bileşenleri önemli nüansları içerir. Yaklaşımlar Alçak Geçiren Filtre (AGF) ile elde edilen düşük frekans bileşenlerdir. Detaylar ise Yüksek Geçiren Filtre (YGF) ile ile elde edilen yüksek frekanslı bileşenlerdir. Her bir seviyede sinyal 2 ile aşağı yönde örneklenmekte (downsampling) ve bu şekilde sinyalin boyutu yarıya inmektedir. Aynı zamanda burada seçilen ana dalgacık fonksiyonu ile sinyal her bir seviyede filtrelenmiş oluyor. Bu şekilde 3. seviye ye kadar devam edilen filtreleme işlemi sonucunda örnek sinyal Şekil 3.3'deki gibi elde edilmiştir.



Şekil 3.3 : Dalgacık dönüşümü sonrası örnek veri kümesi.

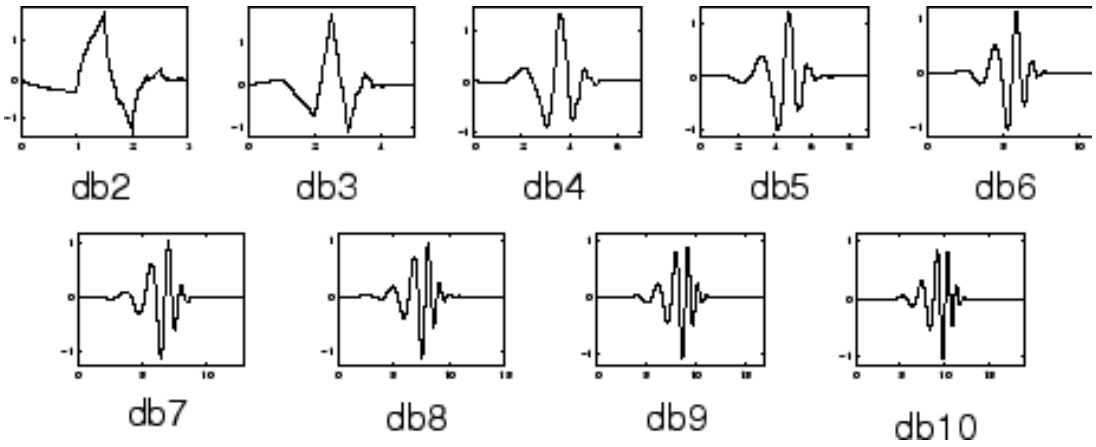
Şekil 3.4'de verilen Haar dalgacığı en kolay ve en basit dalgacık fonksiyonudur. Haar süreksizdir ve basamak fonksiyonuna benzer. Daubechies1 de Haar gibi aynı dalgacığı sunar. Şekil 3.5'de ise db2'den db10'a kadar olan daubechies dalgacıkları gösterilmiştir. Bu çalışmada daubechies ana dalgacık fonksiyonun 10 farklı versiyonu (db1,db2,…..db10) kullanılarak sinyallerin yaklaşım değerleri elde edildi. Bunların içinden en iyi sonucu verenler sonuç tablosunda gösterildi.



Şekil 3.4 : Haar Dalgacığı

3.1.2 İstatistiksel öznitelikler-1

T-testten sonra elde edilen veriler pencerelere bölünerek gerekli öznitelikler hesaplandı. İstatistiksel öznitelik seçimi için farklı boyutlarda (10, 20,…,100) pencere boyutları kullanıldı.



Şekil 3.5 : Kullanılan Daubechies fonksiyonları db2,…,db10 (url10).

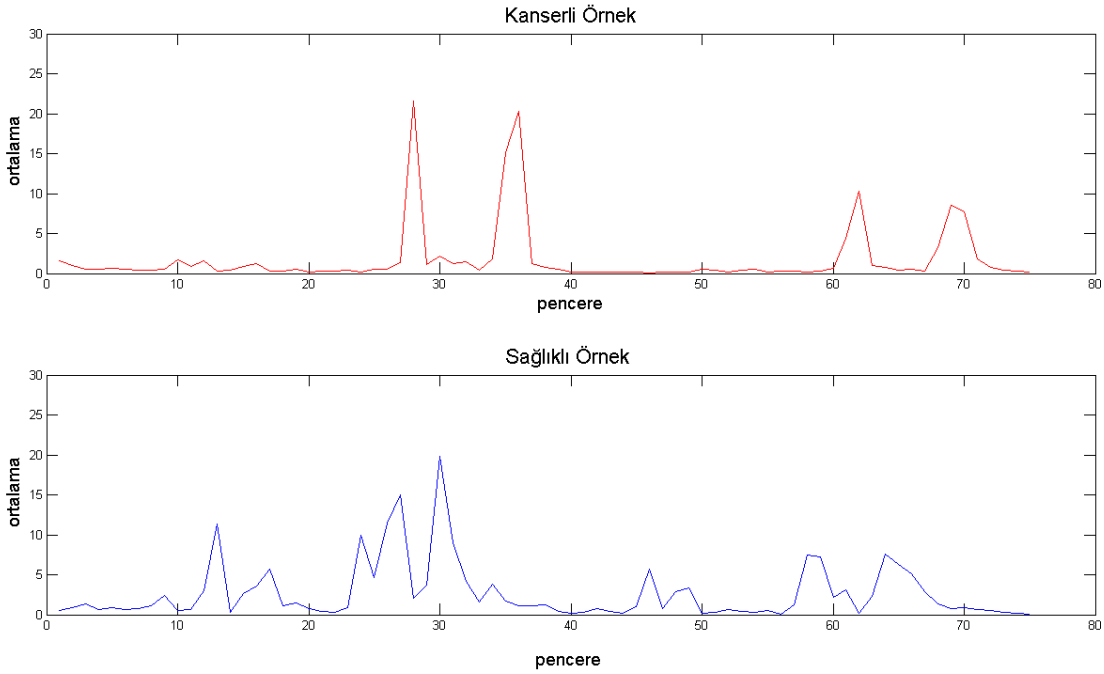
Her pencere için ortalama(mean), varyans(variance), çarpıklık(skewness) ve basıklık (kurtosis) değerleri hesaplandı. Bu dört öznitelik KS verisinin belirlenen pencere için karakteristiğini yansıtacaktır. Bu dört öznitelik Tang ve diğ. (2010) makalesine göre hesaplandı. Basitçe, ortalama gözlemlerin toplamının gözlem sayısına bölünmesiyle elde edilir. Varyans gözlemlerin ortalamadan ne kadar uzakta olduğunun ölçüsüdür. Çarpıklık ise dağılımın ortalamaya göre gösterdiği asimetrinin bir ölçüsüdür. Son olarak basıklık ise dağılımın normal bir dağılıma göre dik veya yatay oluşunun bir ölçüsüdür. Yani yüksek basıklık oranına sahip bir dağılım ortalama yakınında sivri bir tepe oluştururken düşük basıklık değerine sahip dağılım ise daha yatay bir tepe oluşturur.

Seçilen bir veri üzerinde öznitelikler örneklendirilirse; pencere boyutunun 50 olarak seçilmesiyle, 75 pencere elde edilmiş (t-testten sonra veri boyutu ~3800) ve her bir pencere için ortalama, varyans, basıklık ve çarpıklık değerleri hesaplanmıştır.

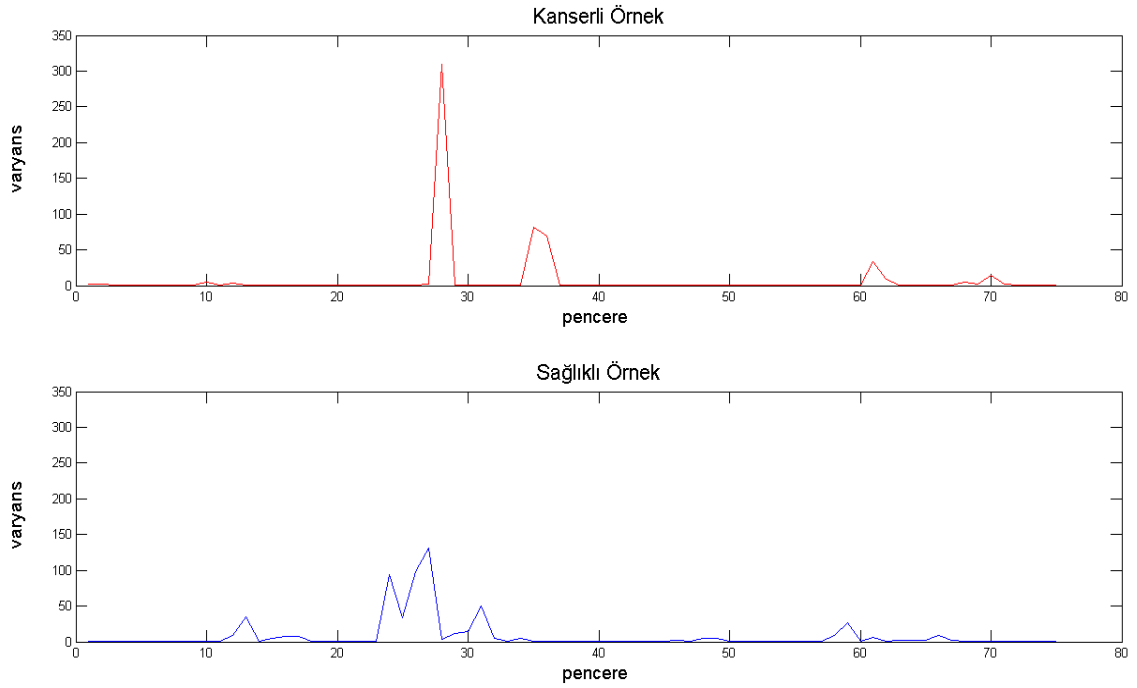
n örnek sayısını ve x_i ise bir örneğe ait vektördeki elemanların her birini temsil etmek üzere, ortalama ($\bar{x}$) eşitlik **(3.1)** de verildiği şekilde hesaplanır. Şekil 3.6'da her bir pencere için hesaplanan ortalama ile elde edilen öznitelikler gösterilmiştir. σ standart sapma olmak üzere, varyans eşitlik **(3.2)** deki gibi hesaplanır. Şekil 3.7 'de elde edilen öznitelikler görülmektedir.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \textbf{(3.1)}$$

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad \textbf{(3.2)}$$



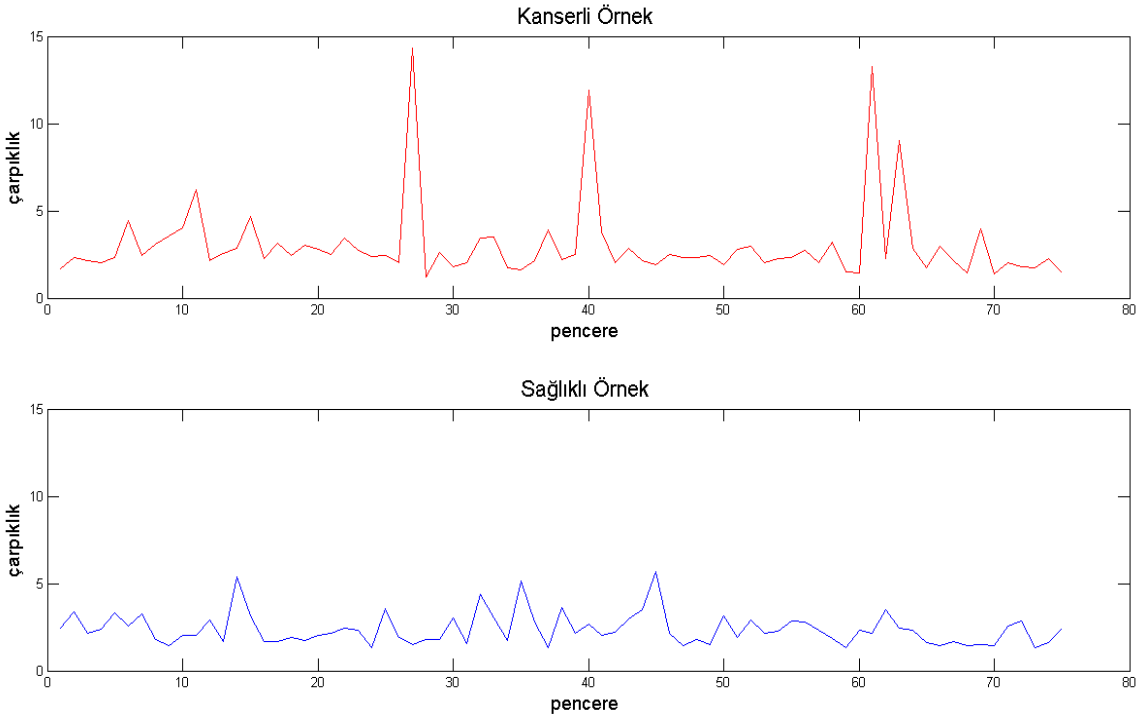
Şekil 3.6 : Ortalama temelli öznitelikler



Şekil 3.7 : Varyans temelli öznitelikler

E(t) t'nin beklenen değeri olmak üzere, çarpıklık eşitlik **(3.3)** de verildiği gibi hesaplanır. Şekil 3.8'de elde edilen öznitelikler görülmektedir.

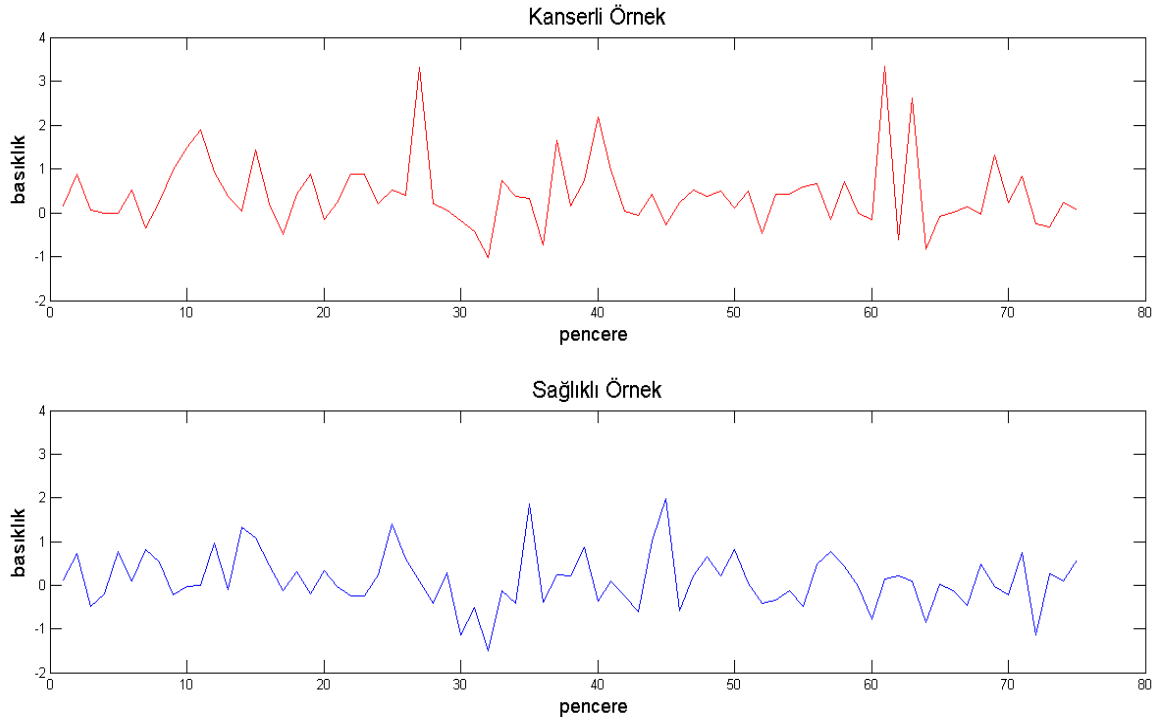
$$y = \frac{E(x-\bar{x})^3}{\sigma^3} \qquad \textbf{(3.3)}$$



Şekil 3.8 : Çarpıklık temelli öznitelikler

Yine **E(t) t**'nin beklenen değeri olmak üzere, basıklık eşitlik **(3.4)** de verildiği gibi hesaplanır. Şekil 3.9'da elde edilen öznitelikler görülmektedir.

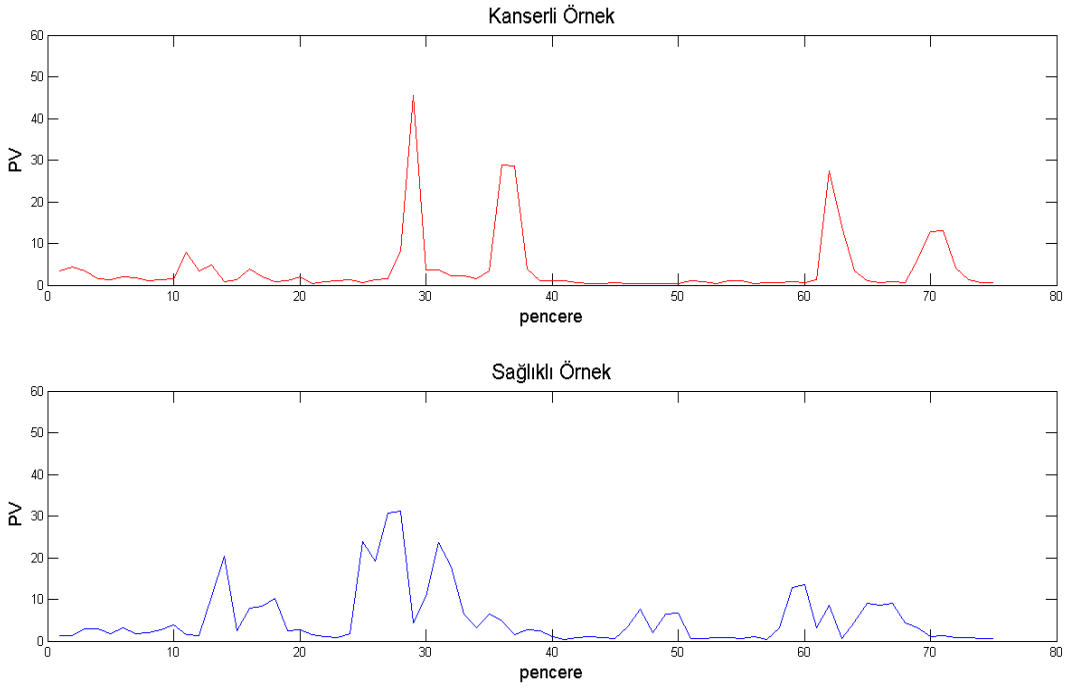
$$k = \frac{E(x-\bar{x})^4}{\sigma^4} \qquad \textbf{(3.4)}$$



Şekil 3.9 : Basıklık temelli öznitelikler

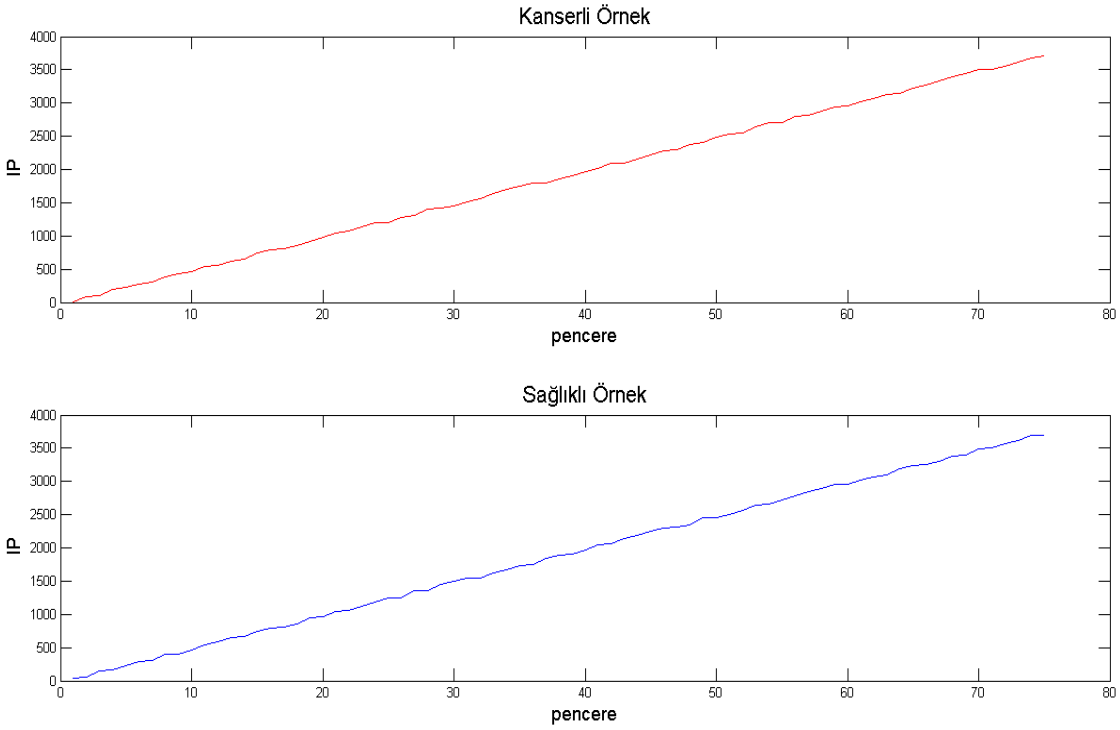
3.1.3 İstatistiksel öznitelikler-2

Chandaka ve diğ.(2009) tarafından EEG sinyallerinden öznitelik çıkarımı için kullanılan istatistiksel öznitelikler-2, bu çalışmada bazı değişiklikler ile kullanılmıştır. Bunlardan en önemlisi hesaplama yöntemleri kullanılırken pencere kullanılmasıdır. İstatistiksel öznitelikler-1 yönteminde olduğu gibi verideki bir örnek için boyutları değişen pencereler (10,20,…,100) üzerinde gerekli öznitelikler hesaplanmıştır. Bu şekilde literatürde daha önce de kullanılan istatistiksel öznitelikler-1 yöntemi ile karşılaştırma yapılması amaçlandı. Kolaylıkla hesaplanabilecek olan her bir pencereye ait tepe değeri (Peak Value-PV) ve bu tepe değerin hangi noktada ortaya çıktığı (instant at which peak occurs- IP) hesaplanacak ilk iki özniteliktir. Şekil 3.10'da elde edilen *PV* öznitelikleri ve Şekil 3.11'de *IP* öznitelikleri görülmektedir.



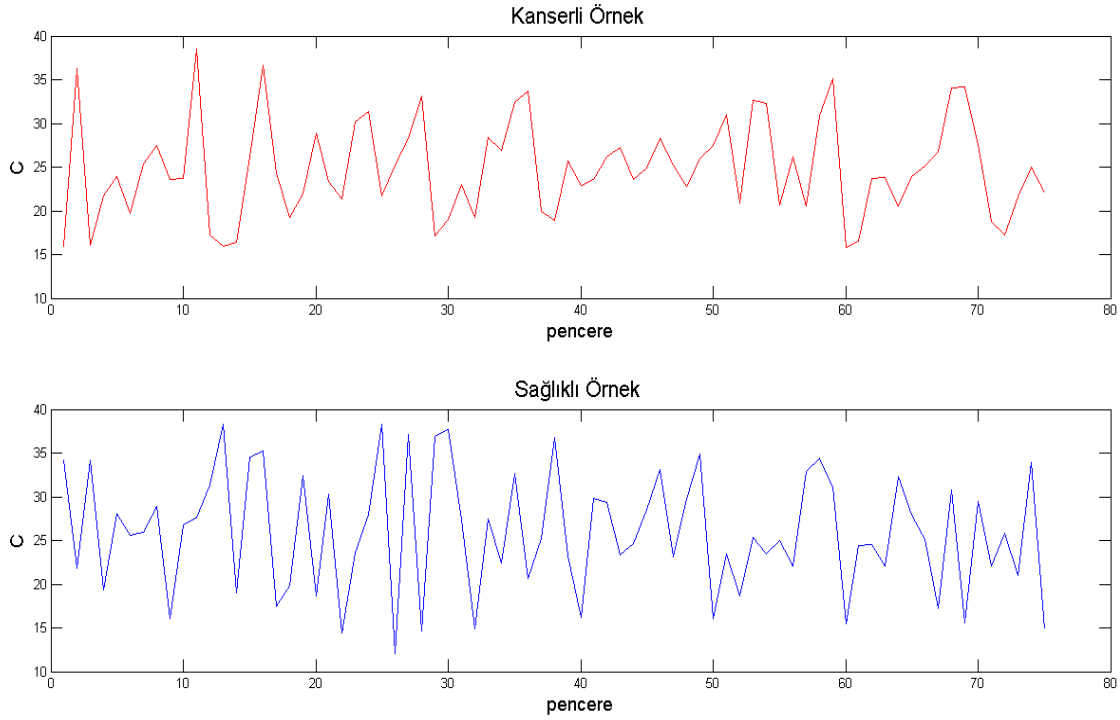
Şekil 3.10 : *PV* temelli öznitelikler

(3.5) eşitliği ile verilen "centroid" aynı zamanda birinci moment olarak adlandırılır. Bir moment eğer sıfır etrafında alınırsa merkezi moment olarak adlandırılır ki bu aynı zamanda beklenen değere karşılık gelir. Şekil 3.12'de elde edilen öznitelikler görülmektedir. **(3.5), (3.6)** ve **(3.7)** de verilen eşitliklerde **m** veri kümesinde ilgilenilen noktanın x ekseni üzerindeki yeridir, yani kütle/yük oranıdır. **R(m)** ise bu noktalara karşılık gelen yoğunluk değerleridir.



Şekil 3.11 : *IP* temelli öznitelikler

$$C = \frac{\sum_{m=-M}^{M} mR(m)}{\sum_{m=-M}^{M} R(m)} \qquad \textbf{(3.5)}$$



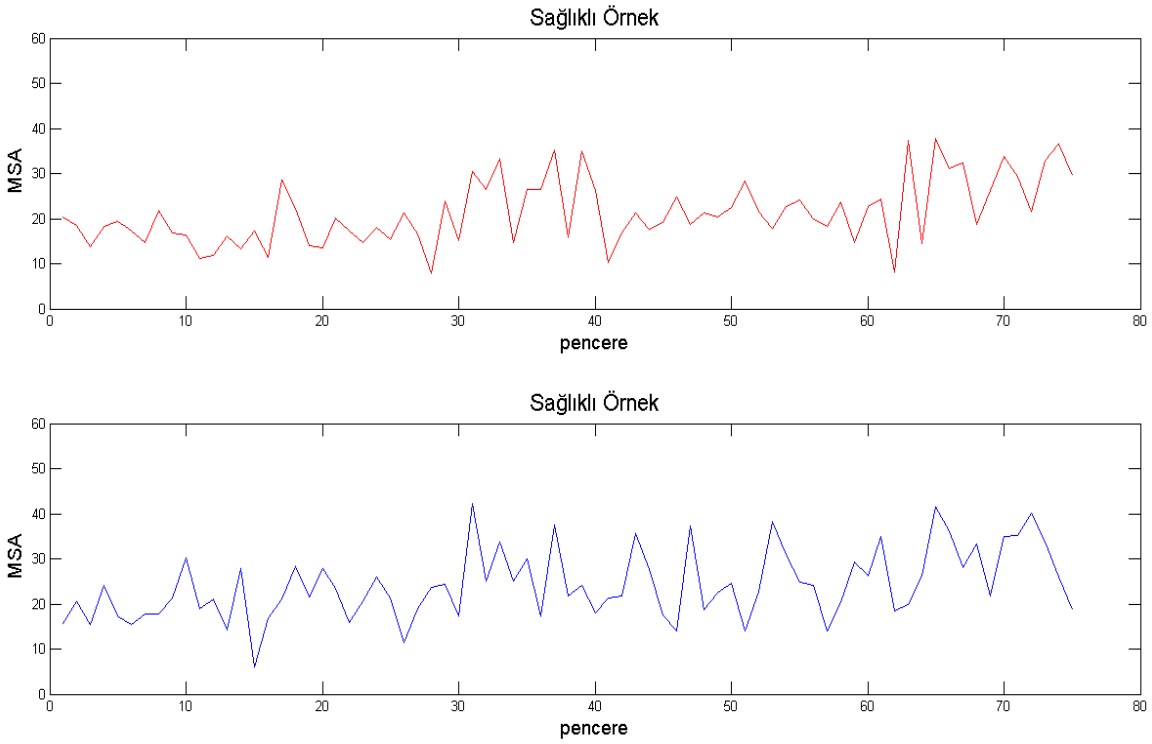
Şekil 3.12 : *Centroid* temelli öznitelikler

(3.6) eşitliği ile verilen "Mean-square abscissa" ise normalize edilmiş ikinci moment olarak adlandırılır. Eğer bu moment sıfır etrafında alınmış ise varyans olarak adlandırılır. Şekil 3.13'de her bir pencere için hesaplanan MSA değeri ile elde edilen öznitelikler görülmektedir.

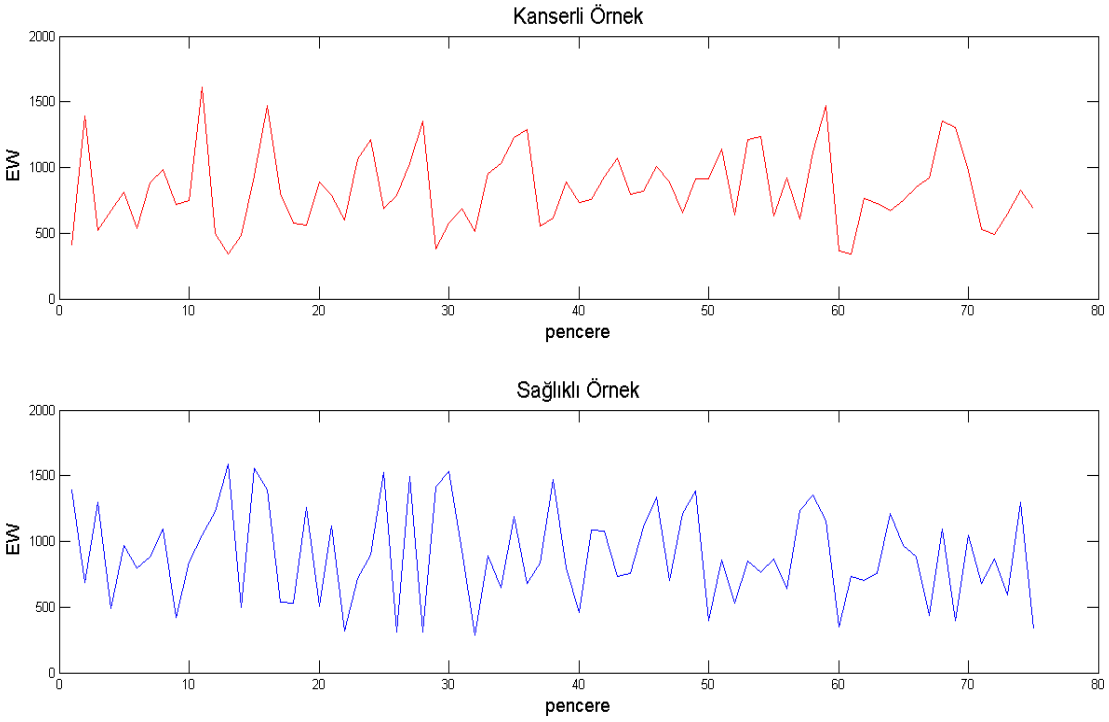
$$MSA = \frac{\sum_{m=-M}^{M} m^2 R(m)}{\sum_{m=-M}^{M} R(m)} \qquad \textbf{(3.6)}$$

(3.7) eşitliği ile verilen "Equivalent width" astronomide sıkça kullanılmaktadır (**url11**). Bir spektrumun karakteristiğini belirlemek için sınırlandırılmış bir bölgede kalan çizginin altındaki alanı veren belirli uzunluktaki dikdörtgenin genişliği olarak tanımlanabilir. Şekil 3.14'de elde edilen öznitelikler görülmektedir

$$EW = \frac{\sum_{m=-M}^{M} R(m)}{Peak.value.of.R(m)} \qquad \textbf{(3.7)}$$



Şekil 3.13 : *MSA* temelli öznitelikler



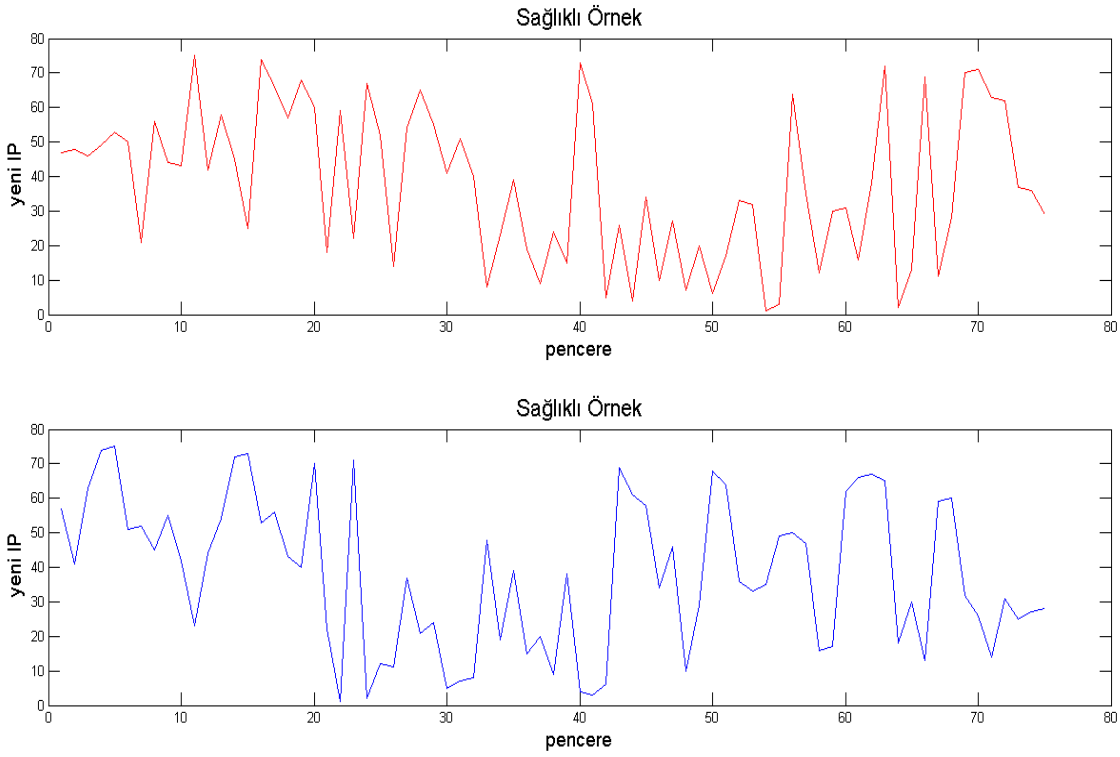
Şekil 3.14 : *EW* temelli öznitelikler

3.1.4 İstatistiksel öznitelikler-3

Bu kısımda kullanılan öznitelikler, istatistiksel öznitelikler-2'den faydalanarak oluşturulmuştur. İstatistiksel öznitelikler-2'de elde edilen beş öznitelikten biri olan *instant at which peak occurs* (IP) üzerinde bir takım değişiklikler yapılmış ve diğerleri olduğu gibi kullanılmıştır. *Yeni-IP'*nin hesaplanması için sırasıyla şu işlemler gerçekleştirilir.Her bir pencere için tepe değerleri bulunur. Daha sonra bu tepe değerleri büyükten küçüğe sıralanır. Bu sıralama sonrasında pencere numaraları öznitelik olarak kabul edilir. Bu şekilde hesaplanan yeni-IP'nin kullanılması ile yapılan işlemler ile daha iyi sonuçlar alındığı görülmüştür. Örnek veri kullanılarak elde edilen yeni-IP öznitelikleri Şekil 3.15'de gösterilmiştir.

3.2 En İyi Özniteliklerin Belirlenmesi

Bu çalışmada en iyi özniteliklerin belirlenmesi için literatürde sıklıkla kulanılan *çekirdek kısmi en küçük kareler yöntemi* (KPLS) kullanılmıştır. Bu yöntem bir anlamda *çekirdek matris* (kernel matrix) ile *kısmi en küçük kareler* (PLS) yönteminin birleşimidir (Rosipal, 2003). Kısmi en küçük kareler (KEKK), bağımlı değişkenler (sınıflar) ile bağımsız değişkenler (öznitelikler) arasında doğrusal bir ilişki kurmak için kullanılan etkili bir yöntemdir.



Şekil 3.15 : İstatistiksel öznitelikler-3 yöntemi ile elde edilen *yeni-IP* temelli öznitelikler.

KEKK, regresyon için katsayıların hesaplanmasında kullanıldığı gibi öznitelik seçimi için de kullanılabilmektedir. Özellikle bağımsız değişken sayısının fazla ve örnek sayısının az olduğu durumlarda oldukça kullanışlı olan bu yöntem 1960'lı yıllarda Helman Wold tarafından geliştirilmiştir (Rosipal ve Trejo, 2001). Doğrusal olmayan iteratif kısmi en küçük kareler yöntemi (NIPALS) olarak da adlandırılan bu yöntemden sonra farklı versiyonları literatürde kullanılmıştır (Baffi ve diğ., 1999, Wold ve diğ., 1989, Wold, 1992).

KEKK, **X** nxN öznitelik matrisi ve **Y** nxL sınıf matrisi arasındaki Y=XB+E şeklindeki ilişkiyi tanımlamak için kullanılır. Burada n örnek sayısını, N özniteliklerin boyutunu ve L ise kaç farklı sınıf değişkeninin olduğunu göstermektedir. Bu çalışmada iki sınıflı bir sınıflama yapılacağı için L=1 seçilmiştir. N değeri bir önceki adımda kullanılan öznitelik çıkarma yöntemine göre değişmekle birlikte 200-500 arasında değişmektedir, n ise kullanılan veri kümesine göre 150-250 arasında değişmektedir.

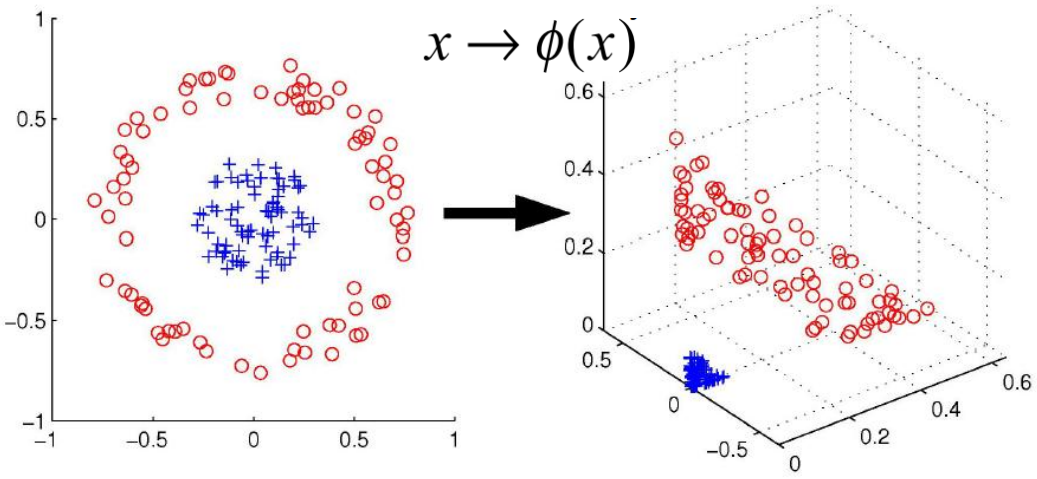
KEKK ile hesaplanan yardımcı matrisler yardımıyla **B** katsayı matrisi hesaplanır. **B** matrisinin bulunması regresyon problemleri için önemlidir ve bu çalışmada kullanılmamıştır. Öznitelik seçimi için yine yardımcı matrisler yardımıyla izdüşüm

matrisi hesaplanır. Boyutu iterasyon sayısı ile belirlenen izdüşüm matrisi sayesinde istenen sayıda öznitelik seçilir. Seçilen bu öznitelikler sınıf matrisini en iyi açıklayan özniteliklerdir.

Temel Bileşenler Analizi (PCA) ile oldukça benzerlik gösteren bu yöntemde veri kümesinin izdüşümü alınırken sınıf bilgileri de göz önüne alınır. Bu şekilde izdüşüm yapılırken doğruluk oranı artmış olur. KEKK yönteminin TBA'ya göre oldukça hızlı olması (TBA yöntemi ile temel bileşenler hesaplanırken,örneğin 40 öznitelik ve 216 örnek için geçen süre 21-22 saniye kadarken, bu süre KEKK için 0.22-0.23 saniye kadardır) bu yöntemi boyut azaltma ve öznitelik seçme işlemlerinde öne çıkarmaktadır.

3.2.1 Çekirdek kısmi en küçük kareler (ÇKEKK) yöntemi

Yüksek boyutlu örneklerin bir fonksiyon yardımı ile başka bir uzaya taşınması sonucu *öznitelik uzayı (çekirdek uzayı)* oluşturulur. ÇKEKK yönteminde doğrusal regresyon fonksiyonu oluşturulan bu öznitelik uzayında tanımlanır. Modelleme işlemi doğrusal olmayan bir uzayda doğrusal regresyon yöntemleri ile yapılır. Şekil 3.16'da görüldüğü gibi, verilerin bir φ fonksiyonu ile taşınması sonucu bu yeni uzayda daha kolay ayrılacağı varsayılır. Bu işlemi gerçekleştirecek fonksiyonun boyutunun belirsiz olması (sonsuz boyuta da sahip olabilir) problem oluşturmaktadır. Bu taşıma esnasında meydana gelen boyut belirsizliğini ortadan kaldırmak için *Çekirdek Hilesi* (Kernel Trick) kullanılır.



Şekil 3.16 : Verilerin çekirdek fonksiyonu ile başka bir uzaya taşınması.

Çekirdek hilesi (Kernel Trick)

Çekirdek hilesi Şekil 3.16'da görüldüğü gibi, örnek uzayındaki noktaların başka bir uzaya taşınması işleminin, φ fonksiyonu ile tek tek yapılması yerine tek bir işlem ile gerçekleştirilmesidir. φ fonksiyonunun önemini yitirdiği bu kısımda öznitelik uzayında noktalar arasındaki benzerlik iç çarpım yoluyla bulunmaktadır. Bu işlem $\mathbf{k(x_i,x_j) = < \varphi(x_i) , \varphi(x_j) >}$ şeklinde gösterilebilir. Burada $\mathbf{x_i}$ ve $\mathbf{x_j}$ örnek uzayındaki iki örneğe ait vektörlerdir (boyutları 1xN). $\mathbf{\varphi(x_i)}$ ve $\mathbf{\varphi(x_j)}$ ise $\mathbf{x_i}$ ve $\mathbf{x_j}$ 'nin öznitelik uzayına taşınmış durumlarını temsil etmektedir. **"k"** kulanılan çekirdek fonksiyonudur. Bir başka tabirle; verilerin önce öznitelik uzayına taşınması ve sonrasında gerçekleştirilen iç çarpım işlemi, verilerin örnek uzayında bir kernel fonksiyonundan geçirilmesi ile gerçekleştirilebilir (bu konuda ayrıntılı bir örnek EK-A'da verilmiştir).

Öznitelik uzayına taşınan veriler arasındaki benzerliklerin araştırılması ile *çekirdek matrisi* diğer adıyla *gram matris* oluşturulur. Gram matrisi **(K),** örnek kümesine ait her bir noktanın bilgisini içeren kare matristir. Bu matrisin her bir elemanı $\mathbf{K_{ij} = k(x_i,x_j)}$ işlemi ile hesaplanır**.** Burada **k** çekirdek fonksiyonunu temsil etmektedir. $\mathbf{x_i}$ 1xN (i=1,2,…,n) ve $\mathbf{x_j}$ 1xN (j=1,2,..,n) boyutlu olmak üzere örneklere ait satır vektörleridir. Bu işlem sonucunda elde edilen gram matris nxn boyutlu olmaktadır. Bu işlem matris formda eşitlik **(3.8)**'de verildiği gibi gösterilebilir.

$$K = \begin{bmatrix} \varphi(x_1)^T\varphi(x_1) & \varphi(x_1)^T\varphi(x_2) & \dots \\ \varphi(x_2)^T\varphi(x_1) & \ddots & \\ \vdots & & \end{bmatrix} = \begin{bmatrix} k(x_1,x_2) & k(x_1,x_2) & \dots \\ k(x_2,x_1) & \ddots & \\ \vdots & & \end{bmatrix} \quad \textbf{(3.8)}$$

Oluşturulan gram matris eğitim matrisi olarak kullanılır. Bunun yanında bir de test matrisine ihtiyaç vardır. Bu durumda bazı örneklerin gram matris oluşturulurken kullanılmaması gerekir. Özellikle sınıflandırma aşamasında sınıflayıcılara eğitim ve test kümeleri tanıtılırken yapılan bu ayrım göz önüne alınmalıdır. KEKK yöntemi denetimli bir öznitelik seçme yöntemi olduğu için bu noktaya dikkat edilmemesi hatalı sonuçların elde edilmesine sebep olabilir.

Test için kullanılacak çekirdek matrisi $\mathbf{K_t}$ şu şekilde oluşturulur; $\mathbf{K_{ij} = k(x_i,x_j)}$ **,** burada $\mathbf{x_i}$ 1xN (i=n+1,n+2,….,n+nt) ve $\mathbf{x_j}$ 1xN (j=1,2,..,n) boyutlu olmak üzere örneklere ait satır vektörleridir. Dolayısıyla $\mathbf{K_t}$ matrisi nt x n boyutlu bir matris olur.

Eğitim matrisi ve test matrislerinin ayrı ayrı oluşturulup merkezileştirilmesi (centralizing) gerekmektedir. Bu işlem **(3.9a)** ve **(3.9b)** de verilen formüller yardımıyla yapılır:

$$K = (I - \frac{1}{n} 1_n 1_n^T) K (I - \frac{1}{n} 1_n 1_n^T) \qquad \textbf{(3.9a)}$$

$$K_t = (K_t - \frac{1}{n} 1_{nt} 1_n^T K)(I - \frac{1}{n} 1_n 1_n^T) \qquad \textbf{(3.9b)}$$

Burada I n boyutlu birim matrisi göstermekteir. $1_{\mathbf{n}}$ ve $1_{\mathbf{nt}}$ sırasıyla boyutları n ve nt, elemanları 1 olan vektörleri temsil etmektedir.

ÇKEKK yönteminde kullanılacak algoritmaya geçmeden önce uygun bir çekirdek fonksiyonun tanımlanması gerekmektedir. Bunun için doğrusal, çok terimli ve Gauss çekirdek fonksiyonları[1] en çok kullanılanlardır. Bu çalışmada eşitlik **(3.10)** ile verilen çok terimli çekirdek fonksiyonu kullanılmıştır.

$$\mathbf{k(x_i,x_j) = (< x_i , x_j > + a)^b} \qquad \textbf{(3.10)}$$

Burada **b** polinomun derecesini ve **a** da sabit bir sayıyı göstermektedir. Optimum değerlerin belirlenmesi için deterministik bir yaklaşım olmadığından **a** ve **b** değerleri deneme-yanılma yoluyla belirlenmiş ve bu çalışma için **a=0** ve **b=2** alınmıştır.

ÇKEKK Algoritması

Buradaki ÇKEKK algoritması (Rosipal ve diğ., 2003) De Jong tarafından geliştirilen SIMPLS algoritmasının çekirdek fonksiyonu ile birleştirilmesi ile oluşturulmuştur (De Jong, 1993). Daha önce de bahsedildiği gibi **K** nxn gram matrisi ve $\mathbf{K_t}$ ise nt x n test matrisini temsil etmektedir. Gram ve test matrisleri **(3.9a)** ve **(3.9b)** 'deki eşitliklere göre merkezileştirilmiştir. **Y** normalize edilmiş sınıf etiketleri vektörünü (iki sınıf olduğu için bir boyutlu cevap kümesi yeterli olmaktadır) temsil etmektedir.

---

[1] Bu fonksiyonlar destek vektör makineleri kısmında verilmiştir. Diğer bazı çekirdek fonksiyonlar ve bu fonksiyonların nasıl oluşturuldukları hakkında daha fazla bilgi için Hofmann (2006) çalışmasına bakılabilir.

$\mathbf{K_{res}= K}$ 'ya eşitlenerek ve istenen sayıdaki gizli değişken değeri (p) atanarak iterasyona başlanır.

i=1 den p ye kadar

$$\mathbf{t_i} = \mathbf{K_{res}Y}$$

$$\mathbf{t_i = t_i \ / \| t_i \|}$$

$$\mathbf{u_i = Y\ (Y^T t_i)}$$

$$\mathbf{K_{res} = K_{res} - t_i\ (t_i^T K_{res})}$$

$$\mathbf{Y = Y - t_i\ (t_i^T Y)}$$

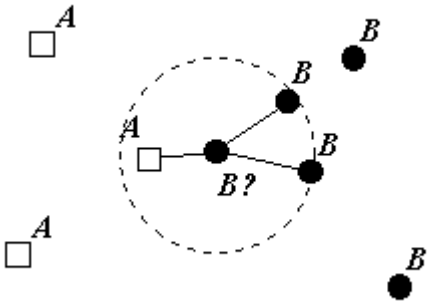
son

$\mathbf{T= [t_1\ t_2 \ldots\ t_p]}$, $\mathbf{U= [u_1\ u_2 \ldots\ u_p]}$ izdüşüm matrisleri olmak üzere, test örneklerinin çekirdek uzayındaki izdüşümleri $\mathbf{T_t= K_t U(T^T KU)^{-1}}$ şeklinde hesaplanır.

3.3 Sınıflama Yöntemleri

3.3.1 K- en yakın komşuluk

Denetimli bir öğrenme algoritması olan k-en yakın komşu (k-Nearest neighbor, kNN) algoritması da sınıflamada sıkça kullanılan bir algortimadır. kNN yeni bir örnek geldiğinde bu örneği en yakın k komşusunun etiketlerine bakarak sınıflandırmaktadır. Tam bir ön bilginin mevcut olmadığı durumlarda sıkça kullanılan bu gibi yöntemlerde, öznitelik vektörleri arasındaki mesafe ve benzerlik bilgisini kullanarak sınıflama yoluna gidilir. k-NN sınıflayıcısı bu anlamda basit bir yapıya sahiptir. Giriş vektörüne k adet en yakın komşu öznitelik vektörü bulunur ve daha sonra çoğunluğun kararı sınıf etiketi olarak atanır. Dikkat edilmesi gereken noktalardan biri belirsizliği ortadan kaldırmak için k sayının tek seçilmesi gerektiğidir. Bunun yanında k sayısının uygun belirlenememesi sınıflama kararının başarısını düşürecektir. Şekil 3.17'de görüldüğü gibi basit bir sınıflama problemi olduğu varsayılsın. k=1 seçilmesi durumunda sınıfı belirsiz olan nokta A sınıfına, k=3 seçilmesi durumunda ise B sınıfına atanacaktır. k sayısının büyük seçilmesi sınıflama başarısını arttıracağı gibi en yakın komşunun etkisini azaltacağından başarıyı azaltabilmektedir.



Şekil 3.17 : k-NN için "k" komşuluk değerinin belirlenmesi , k=3 (url12).

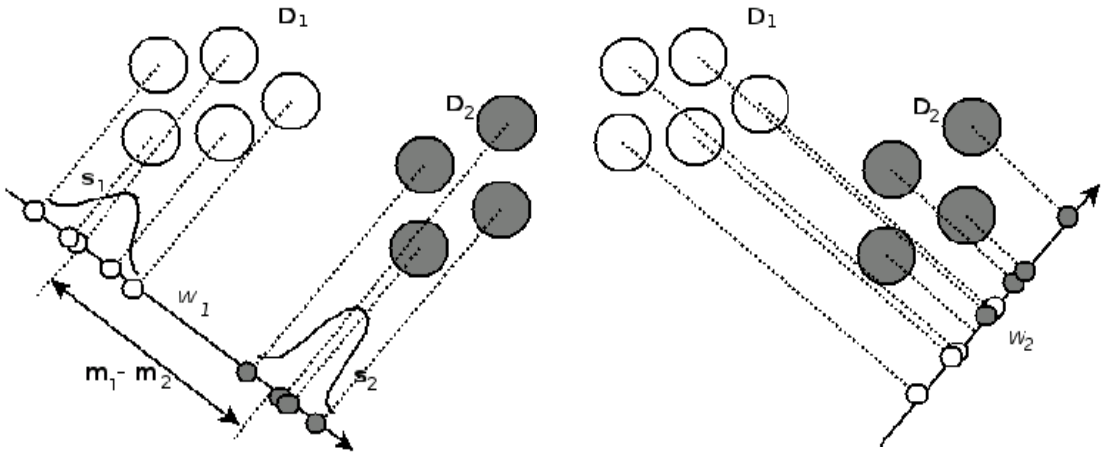
kNN sınıflayıcılarda, giriş vektörüne sınıf etiketleri atanırken herbir öznitelik vektörüne eşit önem verilmesi problem oluşturabilmektedir. Öznitelik vektörleri öbekleşme göstermeyip iç içe girdiği durumlarda sorun teşkil edecektir. Diğer bir problem ise bir giriş vektörü bir sınıfa atandığında , bu vektörün sınıfa üyeliğinin gücü hakkında herhangi bir bilgi bulunmamaktadır. Bu çalışmada k=1,3,5 için yapılan denemeler sonucu k değeri 3 olarak seçildi.

3.3.2 Lineer diskriminant analiz

Özniteliklerin boyutunu azaltma, *çok boyutluluk sorununundan* (curse of dimensionality) kurtulmak için kritik bir öneme sahiptir (Duda, 2001). Fisher LDA, farklı sınıflara ait örneklerin yüksek boyutlu bir uzaydan daha düşük boyutlu bir uzaya uygun bir izdüşüm alınarak taşınmasını sağlar. Uygun bir izdüşüm alınmasıyla iyi bir ayrım sağlanabilir. Fisher LDA daki ana fikir sınıflar arası saçılım matrisinin, sınıf içi saçılım matrisine oranını maksimum yapan doğruyu (w) bulmaktır. Şekil 3.18'de iki farklı w için temsili izdüşümler verilmiştir. Görüleceği üzere w_1 doğrusal ayrımı sağlarken w_2 bu ayrımı yeterince iyi sağlayamaz. Ayrımı en iyi yapacak olan doğruyu bulmak için eşitlik **(3.11)** ile verilen kriterin maksimum olması amaçlanmaktadır.

$$J(w) = \frac{\left|\tilde{m}_1 - \tilde{m}_2\right|}{\tilde{s}_1 + \tilde{s}_2} \qquad \textbf{(3.11)}$$

Burada $\left|\tilde{m}_1 - \tilde{m}_2\right|$ izdüşümü alınmış noktaların ortalamaları arasındaki mesafe olup eşitlik **(3.12)** de verildiği şekilde hesaplanır. Burada m_1 ve m_2 sınıflara ait ortalamaları göstermektedir.



Şekil 3.18 : İzdüşüm için seçilecek doğrunun (w) belirlenmesi (url13)

$$\left|\tilde{m}_1 - \tilde{m}_2\right| = \left|w^T (m_1 - m_2)\right| \qquad \textbf{(3.12)}$$

$\tilde{s}_1 + \tilde{s}_2$ ise izdüşümü alınmış noktaların sınıf içi saçılımını göstermektedir. Her bir sınıfa ait saçılım eşitlik **(3.13)** de verildiği gibi hesaplanır. Burada *y* izdüşümü alınan noktaların her birini göstermektedir. $\tilde{m}_i$ ise izdüşümü alınmış noktaların sınıf ortalamasını temsil etmketedir. (Duda ve diğ, 2001).

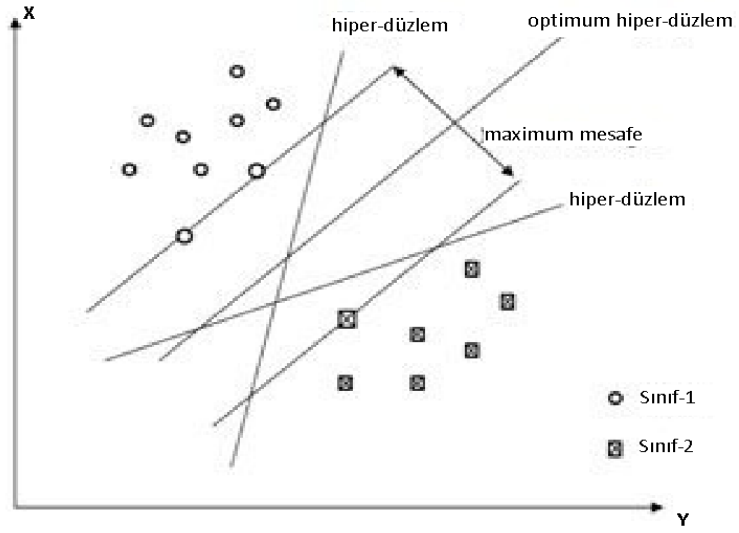
$$\tilde{s}_i = \sum_{y \in w} (y - \tilde{m}_i)^2 \qquad \textbf{(3.13)}$$

Sınıflama işlemi için test kümesindeki noktaların izdüşümleri eğitim kümesi yardımıyla belirlenen **w** doğrusu üzerine alınır. Eğitim kümesi göz önüne alınarak izdüşüm doğrusu üzerinde sınıfları ayıran optimum bir nokta belirlenir. Belirlenen bu noktaya göre test kümesindeki izdüşümü alınmış veriler iki sınıfa ayrılır. Böylece doğru ve yanlış sınıflandırılan noktalar ile başarım oranı hesaplanır.

3.3.3 Destek vektör makineleri

Destek Vektör Makinaları (Support Vector Machines,SVM) örüntü tanıma problemlerinde sıkça kullanılan bir yöntemdir. Destek vektör makinalarının en basit modeli doğrusal ayrılabilen örnekler için uygulanan modeldir. Bu durum Şekil 3.19 'daki gibi gösterilebilir. DVM ile sınıflandırmada genellikle{-1,+1} şeklinde sınıf etiketleri ile gösterilen sınıfa ait örneklerin, bir karar fonksiyonu kullanarak ayrılması hedeflenir. Şekil 3.19'da da görüleceği gibi bu ayrımı yapabilecek birden fazla hiper-düzlem (hyper-plane) tanımlanabilmektedir. Bu gibi durumlarda optimum

hiper-düzlemin belirlenmesi gerekmektedir. Kendisine en yakın noktalar arasındaki uzaklığı maksimum yapan hiper-düzlem optimum olandır.



Şekil 3.19 : İki sınıflı bir problem için hiper-düzlemler ve optimum hiper-düzlem

Şekil 3.20'de görüldüğü gibi noktalar arasındaki mesafeyi sınırlandıran noktalar destek vektörleri olarak adlandırılır. Optimum hiper-düzlemin bulunması için eşitlik **(3.14a)** ve **(3.14b)** kullanılır.

$$w.x_i + b \geq +1 \quad \text{her } y=+1 \text{ için} \qquad \textbf{(3.14a)}$$

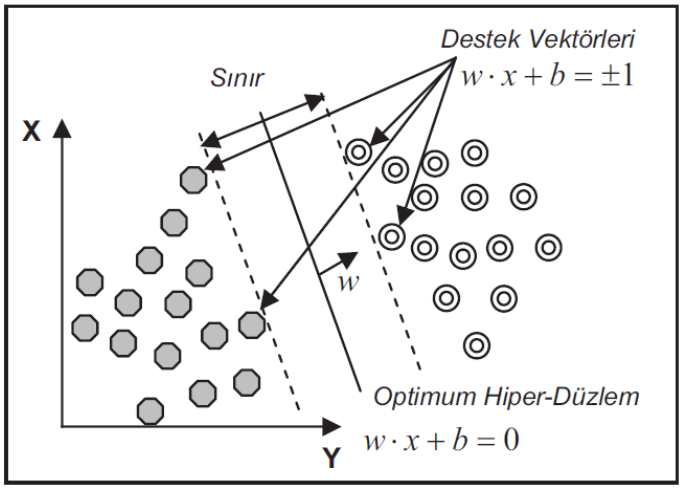
$$w.x_i + b \leq +1 \quad \text{her } y=-1 \text{ için} \qquad \textbf{(3.14b)}$$

Burada $x \in R^N$ olup N boyutlu bir uzaydaki girişleri, $y \in \{-1,+1\}$ ise sınıf etiketlerini, w hiper düzlemin normalini ve b eğilim değerini göstermektedir.

Optimum hiper-düzlemin belirlenebilmesi için bu düzleme parallel ve sınırlarını belirleyecek şekilde iki hiper-düzlem seçilir. Destek vektörleri yardımıyla seçilen bu düzlemler $w.x_i + b = \pm 1$ şeklinde ifade edilir. Optimum hiper-düzlemin sınırının maksimum olması için $||w||$ değerinin minimum hale gelmesi gerekmektedir. En iyi hiper-düzlemin belirlenmesi için, **(3.15a)** ve **(3.15b)** eşitlikleri ile verilen minimizasyon probleminin **w** ve **b** için çözülmesi gerekmektedir (Abe, 2010).

$$Q(w,b)=\left[\frac{1}{2}\|w\|^2\right] \qquad \textbf{(3.15a)}$$

$$y_i\,(w.x_i + b) - 1 \geq 0 \;,\; i=1,2,\ldots,M. \qquad \textbf{(3.15b)}$$

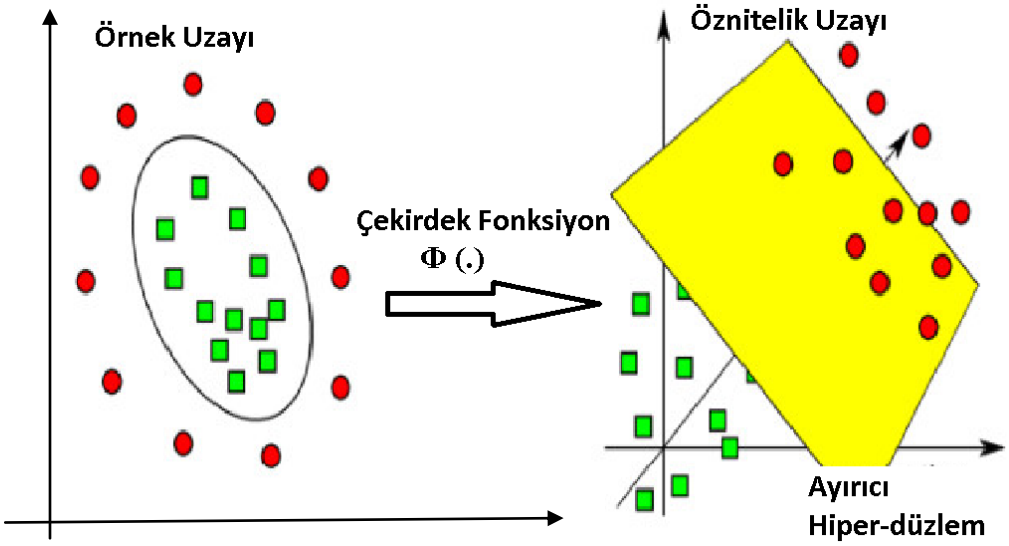


Şekil 3.20 : Doğrusal olarak ayrılabilen veri setleri için hiper-düzlemin belirlenmesi (Kavzaoğlu T.,Çölkesen İ., 2010)

İki sınıflı bu problemde sınıf-1 için y_i=1 ve sınıf-2 için y_i=-1 alınabilir. Bu optimizasyon problemi Lagrange çarpanları kullanarak çözülebilir (Abe, 2010).

Bu işlemlerin sonucunda karar fonksiyonu **(3.16)** da verildiği gibi ifade edilir (Osuna ve diğ,1997).

$$f(x)= sign(< w, x > +b)= sign\left(\sum_{i=1}^{k} \lambda_i y_i (x.x_i) + b\right) \qquad \textbf{(3.16)}$$



Şekil 3.21 : Verilerin çekirdek fonksiyonu ile başka bir uzaya taşınması ile doğrusal ayrılabilir duruma gelmesi.

Doğrusal olarak ayrılamayan veriler basitçe şekil 3.21'de de görüldüğü gibi doğrusal olmayan bir taşıma yöntemi ile daha yüksek boyuttaki bir uzaya taşınmaktadır. Bu sayede noktalar bir hiper-düzlem yardımıyla ayrılabilir hale gelmektedir (Vapnik, 1995). Aslında bu işlem daha önce de bahsedilen çekirdek fonksiyonu yardımı ile verilerin taşınması işlemidir. Bu sebeple DVM ve çeşitleri genel olarak çekirdek tabanlı metodlar (kernel-based methods) olarak da adlandırılır (İşcan Z., 2012).

Eşitlik **(3.15)**'de verilen karar fonksiyonunda yer alan iç çarpımın (burada doğrusal çekirdek fonksiyonunu temsil ettiği de düşünülebilir) yerine doğrusal olmayan bir çekirdek konularak yüksek boyutlu bir uzaya verilerin taşınması işlemi gerçekleştirilir. Yani $<x_i,x>$'in yerine $<\varphi(x_i),\varphi(x)>$ koyulur. Burada $\varphi(.)$ çekirdek fonksiyonudur. Literatürde en sık kullanılan çekirdek fonksiyonlarından bazıları **(3.17a), (3.17b)** ve **(3.17c)**'de verilmiştir. Burada x ve y veri matrisindeki vektörleri temsil etmektedir. İstatistiksel anlamda standart sapmaya karşılık gelen σ, Gaussian çekirdek için genişliğe karşılık gelen sabit bir sayıdır. Çok terimli çekirdek fonksiyonda kullanılan **a** sabit terimi ve **b** ise fonksiyonun derecesini göstermektedir.

Doğrusal Çekirdek: $$K(x,y) = x.y = x^T y \qquad \textbf{(3.17a)}$$

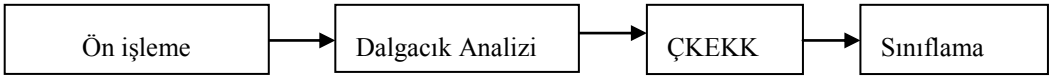
Gaussian Çekirdek: $$K(x,y) = e^{-(\frac{\|x-y\|^2}{\sigma})} \qquad \textbf{(3.17b)}$$

Çokterimli Çekirdek: $$K(x,y) = ((x.y)+a)^b \qquad \textbf{(3.17c)}$$

3.4 Proteomik Verilerin Analizinde Kullanılan Metodlar

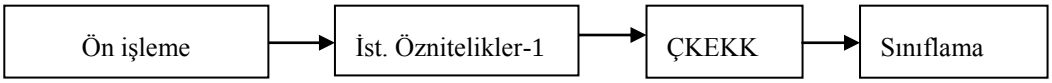
Bu çalışmada sırasıyla önişleme, öznitelik çıkarımı, öznitelik seçimi ve sınıflama işlemleri gerçekleştirilmiştir. Daha öncede bahsedildiği gibi ham olarak alınan data bir ön işlemeden geçirildi. Gürültülerden ve ilgisiz verilerden kurtulmak için, aynı zamanda boyut indirgemek için istatistiksel bir yöntem olan t-Test uyulandı. Öznitelik çıkarımı ve seçimi bu tür yüksek boyutlu verilerde büyük önem arz

etmektedir. Bu sebeple özellikle öznitelik çıkarımı ve seçimi üzerinde durulan bu çalışmada öznitelik çıkarımı için dört ayrı yöntem kullanıldı. İlk yöntem dalgacık analizinin öznitelik seçimi aşamasında kullanılmasıdır. Şekil 3.22'de önerilen metodun öznitelik seçimi için kullanılan dalgacık analizi literatürde daha öncede kullanılan etkili bir yöntemdir (Li ve diğ., 2007).



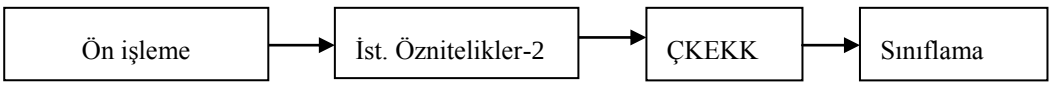
Şekil 3.22 : Proteomik verilerin analizinde kullanılan metodlar-1

Diğer üç yöntem ise istatistiksel hesaplamalara dayanan yöntemlerdir. İstatistiksel öznitelikler-1 yöntemi ile elde edilen öznitelikler Tang ve diğ. (2010) makalesine göre hesaplandı. Şekil 3.23'de önerilen metodun öznitelik seçimi adımında istatistiksel öznitelikler-1 kullanılmıştır.



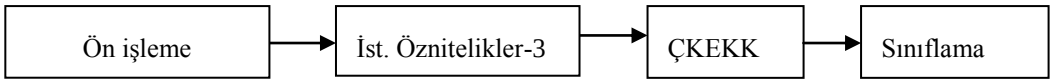
Şekil 3.23 : Proteomik verilerin analizinde kullanılan metodlar-2

İstatistiksel öznitelikler-2 yöntemi ise literatürde farklı bir problemde (EEG sinyallerini kullanarak epilepsi teşhisi için) yine öznitelik çıkarımı için Chandaka ve diğ., (2009) tarafından kullanılmıştır. İstatistiksel öznitelikler-2 yöntemi proteomik verilerinin sınıflandırılması amacıyla ilk defa bu çalışmada kullanılmış olup bu yöntem ile başarılı sonuçlar elde edilmiştir. Şekil 3.24'de önerilen metodda öznitelik seçimi için istatistiksel öznitelikler-2 kullanılmıştır.



Şekil 3.24 : Proteomik verilerin analizinde kullanılan metodlar-3

İstatistiksel öznitelikler-3 ise bir önceki yöntemde (istatistiksel öznitelikler-2) elde edilen özniteliklerden birinin farklı bir yolla elde edilmesiyle oluşturulmuş bir yöntemdir. Şekil 3.25'de önerilen metodun öznitelik seçimi aşamasında istatistiksel öznitelikler-3 kullanılmıştır.



Şekil 3.25 : Proteomik verilerin analizinde kullanılan metodlar-4

Öznitelik çıkarma işlemlerinden sonra öznitelik seçimi için biyoinformatikte ve kimyasal verilerin analizinde (chemometric) sık kullanılan ÇKEKK yöntemi kullanıldı. Son kısımda ise elde edilen verilerin sınıflandırılması için üç ayrı sınıflama metodu (LDA, DVM ve kNN) karşılaştırıldı.

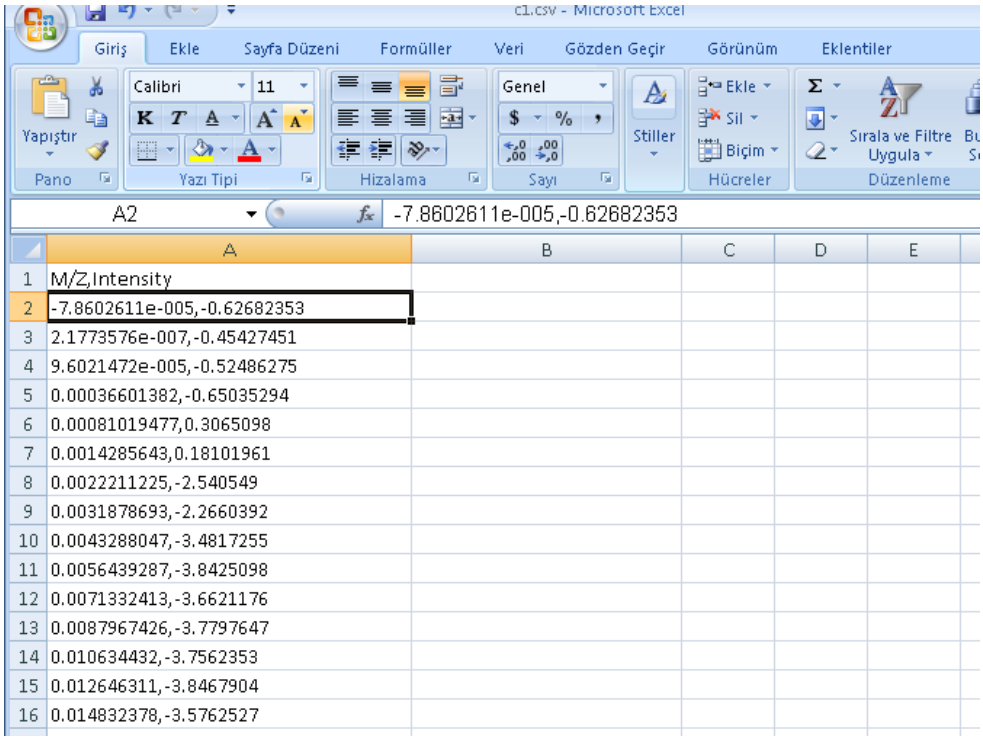
4. SONUÇLAR VE DEĞERLENDİRME

Bu çalışmada yumurtalık ve prostat kanserinin proteomik verilerinden teşhis edilmesi amaçlandı. Biyoinformatikte sıkça karşılaşılan bu tip yüksek boyut veriler için literaturde çok sayıda çalışmada denenen t-Test ve ÇKEKK bu çalışmada kullanıldı. Çalışmada odaklanılan öznitelik seçimi için de literatürde denenmiş olan iki yöntem, dalgacık analizi ve istatistiksel öznitelikler-1 kullanıldı. Bunların yanında daha önce başka bir alanda kullanılan istatistiksel öznitelik seçme yöntemi (istatistiksel öznitelikler-2) ilk defa bir takım değişiklikler yapılarak bu çalışmada denendi ve iyi başarımlar elde edildi. Son olarak da kullanılan bu istatistiksel verilerden birinin yerine (IP) farklı bir istatistiksel yöntem (istatistiksel öznitelikler-3) denenerek başarımın daha da artması sağlandı. Sonuçlardaki hata oranını azaltmak için 5-kat çapraz doğrulama gerçekleştirildi. Veriler her defasında %80 eğitim ve %20 test kümesi olacak şekilde rasgele ayrılıp sınıflama işlemi 50 defa tekrarlanarak elde edilen başarımların ortalamaları alındı.

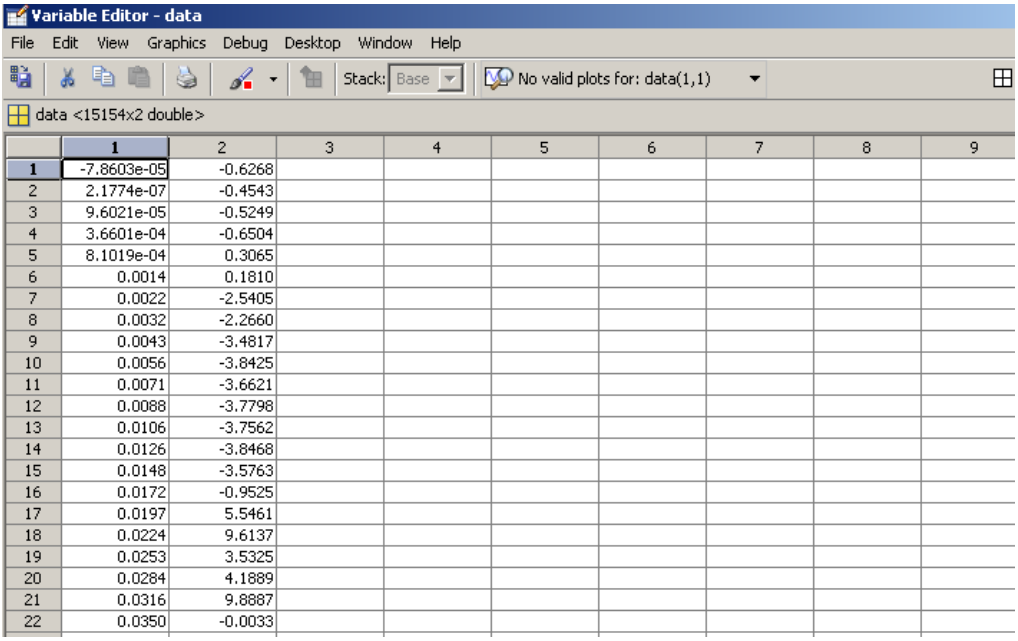
Kullanılan veriler FDA-NCI klinik proteomik programı (FDA-NCI Clinical Proteomics Program Databank) (**url7**) veritabanından alındı. Bu veritabınında tutulan KS verileri kütle/yük (m/z) oranı ve ilgili yoğunluk değeri olmak üzere iki sütundan oluşan dosyalar şeklinde kaydedilmektedir. Yumurtalık kanserine ait veriler yüksek çözünürlüklü SELDI-TOF-MS ile alınmıştır. Bu gruptaki her bir örneğe ait KS verisi ".txt" uzantılı olarak indirilmektedir ve boyutu çok fazladır (>370000). Prostat kanserine ait veriler ise düşük çözünürlüklü olup boyutu daha düşüktür (>15000). Prostat veri kümesindeki örnekler ise ".csv" uzantılıdır ve excel ile görüntülenebilmektedir.

Veri tabanından indilirilen JNCI_Data_7-3-02 isimli dosya prostat kanseri teşhisinde kullanılan örnekleri içermektedir. Şekil 3.26'da bu dosyadadan alınan "c1.csv" isimli örnek verinin excelde gösterimi verilmiştir. Görüldüğü gibi veriler kütle/yük oranına karşılık gelen yoğunluk değerleri şeklinde kaydedilmiştir. Aynı dosya MATLAB ortamında "uiopen('c1.csv')" komutuyla açılabilir. Bu işlem sonrasında veri *.mat* dosyası şeklinde kaydedilir. Bu dosyanın MATLAB ortamında gösterimi şekil

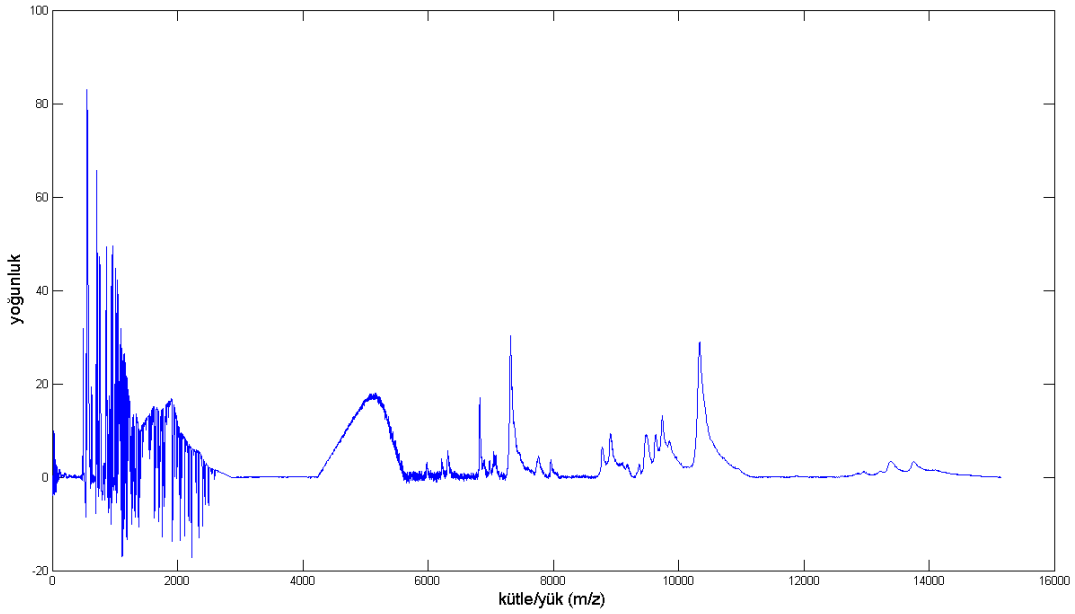
3.27'de verilmiştir. Boyutu 15154 olan bu verinin işlenmemiş hali ise Şekil 3.28'de grafik şeklinde gösterilmiştir.



Şekil 4.1 : İşlenmemiş örnek verinin excel'de gösterimi



Şekil 4.2 : İşlenmemiş örnek verinin MATLAB'da gösterimi



Şekil 4.3 : Örnek verinin MATLAB ortamında grafiksel gösterimi

Kullanılan iki veri kümesi içinde ön işleme kısmında ayrıntıları verilen işlemler, MATLAB (Bioinformatics Toolbox)'tan faydalanılarak gerçekleştirilmiştir. Bu işlemler sonucu yumurtalık kanserine ait verilerin boyutu 15000 ve prostat kanserine ait verilerin boyutu 12000 olarak elde edilmiştir. Böylece veriler önerilen metodların kullanılmasına hazır hale getirilmiştir.

Sınıflandırma sonuçlarının karşılaştırılabilmesi için doğruluk(accuracy, Ac), duyarlılık(sensitivity, Sn) ve belirlilik (specificity, Sp) gib performans terimlerinin hesaplanması gerekmektedir. Doğruluk, örneklerden doğru sınıflandırılanların tüm örneklere oranının yüzde ile verilmesi ile ele edilir. Duyarlılık ise gerçekten hasta olanların tüm hastalar içindeki oranı ile elde edilir. Son olarak belirlilik(seçicilik, özgüllük) ise gerçekte hasta olmayanların içinden kaçının sağlıklı olarak sınıflandırıldığının ölçütüdür. Bu terimler **(4.1a), (4.1b)** ve **(4.1c)** ile verilen matematiksel yollar ile elde edilir.

$$Doğ. = \frac{DP + DN}{DP + DN + YP + YN} \qquad \textbf{(4.1a)}$$

$$Duy.. = \frac{DP}{DP + YN} \qquad \textbf{(4.1b)}$$

$$Bel. = \frac{DN}{DN + YP} \qquad \textbf{(4.1c)}$$

Burada kullanılan DP (Doğru Pozitifler), DN (Doğru Negatifler), YP (Yanlış Pozitifler) ve YN (Yanlış Negatifler) şu şekilde tanımlanır:

DP (True Positives, TP): Doğru sınıflandırılan hasta sayısı

DN(True Negatives, TN): Doğru sınıflandırılan sağlıklı kişi sayısı

YP (False Positives, FP): Yanlış sınıflandırılan sağlıklı kişi sayısı

YN (False Negatives, FN): Yanlış sınıflandırılan hasta sayısı

Elde edilen sonuçlar parametrelerin (istatistiksel öznitelik çıkarımı için kullanılan pencere boyutu, dalgacık analizi için kullanılan dalgacık fonksiyonu ve ÇKEKK ile seçilen öznitelik sayısı) değişimine göre en iyi sonuçları göstermektedir. Kullanılan pencere boyutları 10,20,…100 arasında, öznitelik sayısı 2,4,….50 arasında ve dalgacık fonksiyonu db1,db2,…,db10 arasında değişecek şekilde üç sınıflayıcı için de test edilmiştir. Bu çalışmada işlemler 2.66GHz işlemci ve 4GB RAM'a sahip bilgisayar ile MATLAB ortamında gerçekleştirilmiştir. İlk olarak yumurtalık kanserine ait sonuçlar ve daha sonra da prostat kanserine ait sonuçlar verilmiştir.

4.1 Yumurtalık Kanseri Verileri ve Analizi

Yumurtalık kanseri veri kümesi 216 örneğe aittir. Bu veriler 95 sağlıklı kişi ve 121 yumurtalık kanseri teşhisi konmuş kişiye aittir. Bu veri kümesi kullanılarak kanserli ve sağlıklı bireyler sınıflandırıldı. Çizelge 4.1'de yumurtalık kanserinin sınıflandırılması ile elde edilen sınıflama başarımları verilmiştir.

Çizelge 4.1'de elde edilen sonuçların hangi parametreler ile elde edildiği sonuçların yanında verilmiştir. En iyi sonucun elde edildiği satırlar koyu olarak yazılmıştır. En iyi sonucun dalgacık analizi ile elde edilen özniteliklerin LDA ile sınıflandırılması işlemi ile elde edildiği görülmektedir. Bu yöntem ile yumurtalık kanserinin %95,3 doğruluk oranı ile teşhis edilebileceği görülmektedir. Burada dalgacık fonksiyonu için *db6* alınmış ve öznitelik sayısı 16 seçilmiştir. KNN sınıflayıcısı kullanıldığında en yüksek başarım, ilk defa bu çalışmada kullanılan istatistiksel öznitelikler-3 yöntemi ile elde edilmiştir.

Çizelge 4.1 : Yumurtalık kanseri için sınıflama sonuçları

	KNN				LDA				DVM			
	Doğ.	**Duy.**	**Bel.**	**Değ.**	**Doğ.**	**Duy.**	**Bel.**	**Değ.**	**Doğ.**	**Duy.**	**Bel.**	**Değ.**
Dalg.	0,928	0,939	0,921	db5 ÖS=12	**0,953**	**0,955**	**0,957**	**db6** **ÖS=16**	**0,947**	**0,948**	**0,952**	**db8** **ÖS=20**
İst-1	0,923	0,909	0,951	PB=50 ÖS=8	0,938	0,938	0,945	PB=50 ÖS=12	0,931	0,929	0,939	PB=70 ÖS=10
İst-2	**0,935**	**0,952**	**0,920**	**PB=80** **ÖS=4**	0,947	0,968	0,926	PB=60 ÖS=4	0,940	0,949	0,936	PB=60 ÖS=8
İst-3	**0,935**	**0,935**	**0,942**	**PB=80** **ÖS=4**	0,951	0,964	0,940	PB=60 ÖS=4	0,943	0,950	0,941	PB=80 ÖS=4

ÖS : Öznitelik Sayısı, PB: Pencere Boyutu, Değ: Değişkenler

Aynı zamanda literatürde daha önce kullanılan istatistiksel öznitelikler-1'e göre daha başarılı sonuçlar alınmıştır. Diğer iki sınıflayıcı ile en yüksek başarım dalgacık analizi ile elde edilmiştir. Diğer sonuçlarla karşılaştırıldığında bu çalışmada önerilen istatistiksel yöntemlerin oldukça başarılı olduğu görülmektedir.

Yumurtalık kanserinin sınıflandırılmasında en iyi sonuçlar göz önüne alınarak harcanan sürelerin karşılaştırılmıştır. Dalgacık analizi ile LDA 'nın birlikte kullanımında geçen süre 230 sn, istatistiksel öznitelikler-1 ve LDA 'nın birlikte kullanımı için bu süre 415 sn, istatistiksel öznitelikler-2 ile KNN 'in birlikte kullanımında geçen süre 152 sn ve istatistiksel öznitelikler-3 ile KNN 'in birlikte kulanımında geçen süre 145 sn'dir. Önerilen istatistiksel öznitelik çıkarma yönteminin kullanılması durumunda daha hızlı bir sınıflama yapılabileceği görülmüştür.

4.2 Prostat Kanseri Verileri ve Analizi

Prostat kanseri teşhisi için kullanılan veri kümesi toplam 322 örnekten alınmıştır. Bu veri kümesinde iki ayrı aşamada sınıflandırma yapıldı. İlk aşamada 259 tümörlü

birey ile 63 sağlıklı birey sınıflandırıldı. Daha sonra da tümöre sahip 259 kişi (iyi huylu-190 örnek- ve kötü huylu -kanser, 69 örnek- olarak) sınıflandırıldı. Çizelge 4.2 'de prostat kanserinin sınıflandırılması ile elde edilen sınıflama başarımları verilmiştir.

Çizelge 4.2 : Prostat kanseri için sınıflama sonuçları (1. Aşama: tümor tespiti, 2.Aşama: tümörün cinsinin belirlenmesi).

		KNN				**LDA**				**DVM**			
		Doğ.	**Duy.**	**Bel.**	**Değ.**	**Doğ.**	**Duy.**	**Bel.**	**Değ.**	**Doğ.**	**Duy.**	**Bel.**	**Değ.**
1. Aşama	Wave	0,958	0,971	0,913	db8 ÖS=4	0,956	0,972	0,894	db8 ÖS=12	0,958	0,968	0,921	db6 ÖS=10
	İst-1	0,959	0,969	0,919	PB=100 ÖS=32	0,925	0,954	0,814	PB=70 ÖS=20	0,928	0,946	0,856	PB=70 ÖS=14
	İst-2	0,969	0,982	0,921	PB=80 ÖS=10	0,964	0,980	0,911	PB=80 ÖS=10	0,968	0,982	0,917	PB=80 ÖS=10
	İst-3	**0,973**	**0,979**	**0,952**	**PB =80 ÖS=8**	**0,968**	**0,983**	**0,914**	**PB=80 ÖS=18**	**0,969**	**0,978**	**0,938**	**PB=80 ÖS=8**
2. Aşama	Wave	0,872	0,742	0,924	db7 ÖS=14	0,849	0,661	0,948	db10 ÖS=8	0,865	0,733	0,917	db6 ÖS=10
	İst-1	0,867	0,771	0,901	PB=40 ÖS=18	0,836	0,647	0,931	PB=30 ÖS=24	0,850	0,694	0,918	PB=40 ÖS=14
	İst-2	0,880	0,779	0,919	PB =50 ÖS=8	0,866	0,702	0,948	PB=50 ÖS=4	0,865	0,757	0,904	PB=50 ÖS =4
	İst-3	**0,889**	**0,797**	**0,923**	**PB=50 ÖS=16**	**0,887**	**0,772**	**0,933**	**PB =50 ÖS =28**	**0,882**	**0,782**	**0,921**	**PB=50 ÖS=30**

ÖS : Öznitelik Sayısı, PB: Pencere Boyutu, Değ: Değişkenler

Çizelge 4.2'de verilen sonuçlar iki aşamada incelenebilir. İlk aşamada tumorun (iyi veya kötü huylu) var olup olmadığı teşhis edilmeye çalışıldı. Bu kısımda en yüksek başarım %97.3 ile kNN sınıflayıcısı ve istatistiksel öznitelikler-3'ün birlikte kullanılmasıyla elde edildi. Bunun yanında kullanılan üç sınıflayıcı ile de en yüksek başarı istatistiksel öznitelikler-3 sonucu elde edildi. Benzer sonuç ikinci kısımda da elde edildi. Bu kısımda, var olan tümörün iyi (bening) veya kötü huylu (malign)

olduğu araştırıldı. Burada da en yüksek başarı %88.9 ile yine kNN sınıflayıcısı ve istatistiksel öznitelikler-3'ün birlikte kullanılmasıyla elde edildi. Birinci aşamada elde edilen sonuçlarla paralel şekilde, üç sınıflayıcı ile de en yüksek başarım istatistiksel öznitelikler-3 yönteminin kullanılması sonucu elde edilmiştir. Özetlenecek olursa; kullanılan veri kümesi üzerinde prostata ait tümörün var olup olmadığı %97.3, var olan tümörün çeşidi ise %88.9 doğruluk oranlarıyla tespit edilmiştir.

Sınıflama sürelerinin karşılaştırılması amacıyla ikinci aşama göz önüne alınmıştır. Bütün öznitelik çıkarma yöntemlerinde en iyi sonuç KNN ile elde edilmiştir. Dalgacık analizi kullanılarak yapılan sınıflama işleminde geçen süre 460 sn, istatistiksel öznitelikler-1 için bu süre 514 sn, istatistiksel öznitelikler-2 için 317 sn ve istatistiksel öznitelikler-3 için 283 sn'dir. Bu sonuçlar yumurtalık kanserinin sınıflandırılmasında elde edilen sonuçlar ile örtüşmektedir. En yüksek hız yine bu çalışmada önerilen istatistiksel öznitelik çıkarma yönteminin kullanılması ile elde edilmiştir.

4.3 Tartışma ve Gelecek Çalışma

Bu çalışmada kullanılan yöntem ile yumurtalık kanseri ve prostat kanseri yüksek başarım oranları ile sınıflandırıldı. Her iki problem göz önüne alındığında önerilen yöntemin prostat kanseri için daha iyi sonuç verdiği görüldü. Her ne kadar bu çalışmada karşılaştırmalar için performans kriterlerinden doğruluk göz önüne alınmış olsa da duyarlılık ve belirlilik de bu alanda yapılan çalışmalarda kullanılmıştır. Özellikle bu alanda yapılan çalışmaların klinik anlamda kullanılabilir hale gelmesi için belirlilik oranının % 99.6 olması gerektiği belirtilmiştir (Conrads ve diğ., 2003) . Günümüz itibariyle klinik anlamda kullanılmaya başlamasa da sadece kan örneğinden kanser teşhisi yapılabilmesi büyük bir kolaylık vaat etmektedir. Yapılan çalışmalar bunun çok da uzak bir ihtimal olmadığını göstermektedir. Bu nedenle sınıflandırma başarısını yükseltecek çalışmalar gün geçtikçe artmaktadır. Bundan sonraki çalışmalarda önerilen algortimanın farklı veri tabanları ve kanser türleri üzerinde test edilmesi düşünülmektedir. Aynı zamanda farklı yöntemler geliştirerek başarımların yükseltilmesine çalışılacaktır.

KAYNAKLAR

Abe S., 2010. Support Vector Machines for Pattern Classification (Advances in Pattern Recognition), Second Edition, Springer.

Baffi G., Martin E. B. ve Morris A. J., 1999. Non-Linear Projection to Latent Structures Revisited: the Quadratic PLS Algorithm. *Computers and Chemical Engineering*, 23,395-411.

Banez, L. L., Prasanna, P., Sun, L., Ali, A., Zou, Z., Adam, B. L., McLeod, D. J., Moul, J. W., and Srivastava, S., 2003. Diagnostic potential of serum proteomic patterns in prostate cancer. *J. Urol.* **170,** 442–446

Banks ve diğ., 1994. Ultrasonically assisted electrospray ionization for LC/MS determination of nucleosides from a transfer RNA digest. *Anal Chem*, 66, 406-414.

Belov ve diğ., 2000. Zeptomole-sensitivity electrospray ionizaiton-fourier transform ion cyclotron resonance mass spectrometry of proteins. *Anal Chem*, 72,2271-2279.

Ben-Dor ve diğ., 2000. Tissue Classification with Gene Expression Profiles. *Journal of Computational Biology*, 7, 559-583.

Berry ve diğ., 1995. Endogenous synthesis of galctose in normal men and patients with hereditary galactosaemia. *Lancet*,346,1073-1074.

Bhanot ve diğ., 2006. A robust meta classification strategy for cancer detection from MS data. Proteomics, 6, 592-604.

Biberoğlu G., 2003. Kütle Spektrometresi ve Tıp Alnında Kullanımı. T. Klin. Tıp Bilimleri, 23,491-498.

Blanco R. ve diğ., 2004. Gene selection for cancer classification using wrapper approaches. *Int. J. Recognit. Artif. Intell.*, 18, 1373-1390

Bramer S.E.V., 1998. An introduction to mass spectrometry. http://science.widener.edu/svb/massspec/massspec.pdf

Chandaka ve diğ., 2009. Cross-correlation aided support vector machine classifier for classification of EEG signals. *Expert Systems with Applications*, 36, 1329-1336.

Ceccarelli ve diğ., 2009. A scale space approach for unsupervised feature selection in mass spectra classification for ovarian cancer detection. *BMC Bioinformatics*,10 (Suppl 12),S9.

Conrads ve diğ., 2003. Cancer diagnosis using proteomic patterns. *Expert Rev. Mol. Diagn., 3(4),411-420*.

Datta S. ve DePadilla L.M., 2006. Feature selection and machine learning with mass spectrometry data for distinguishing cancer and non-cancer samples. *Statistical Methodology*, 3, 79-92.

Datta S. ve Datta S., 2002. Comparisons of validation of statistical clustering techniques for microarray gene expression data. Bioinformatics, 19(4), 459-466.

De Jong S., 1993. SIMPLS: An alternative approach to partial least square regression. *Chemometrics and Intelligent Laboratory Systems.*18 Issue:3 pages 251-263.

Duda R.O., Hart P. E. And Stork D.G., 2001. Pattern Classification, Second Ediyion, New York: John Wiley & Sons.

Geurts P. ve diğ., 2005. Proteomic mass spectra classification using decision tree based ensemble methods. *Bioinformatics*, 21 , 3138-3145.

Gögüş Ç., 2006. Prostat kanseri tanısında PSA 4ng/ml sınırı geçerli midir? *Üroonkoloji Bülteni*, 1 , 1-11.

Hilario M. ve Kalousis A., 2008. Approaches to dimensionality reduction in proteomic biomarker studies. *Briefings in Bioinformatics*, 9 (2), 102-108.

Hofmann M., 2006. Support Vector Machines-Kernels and the Kernel Trick. http://www.cogsys.wiai.unibamberg.de/teaching/ss06/hs_svm/slides/SVM_Seminarbericht_Hofmann.pdf

Inza I. ve diğ., 2004. Filter versus wrapper gene selection approaches in DNA microarray domains. *Artif. Intell. Med.,* 31, 91-103.

İşcan Z., 2012. *Development of Classification Methods for Electroencephalogram Based Brain Computer Interfaces*(doktora tezi). İTÜ, Fen Bilimleri Enstitüsü

Jong K. ve diğ, 2004. Feature selection in proteomic pattern data with support vector machines. *In proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, pp.41-48

Kavzaoğlu T., ve Çölkesen İ., 2010. UsDEstek Vektör Makineleri ile Uydu Görüntülerinin Sınıflandırılmasında Kernel Fonksiyonlarının Etkilerinin İncelenmesi. Harita Dergisi, 144, 73-82.

Kong ve diğ., 2006. Using proteomic approaches to identify new biomarkers for detection and monitoring ovarian cancer. *Gynecol Oncol*, 100(2), 247-253.

Krishnapuram ve diğ., 2004. Gene expression analysis: Joint feature selection and classifier design. *Kernel Methods in Computational Biology*, pp 299-318, MIT.

Li ve diğ., 2006. Gene feature extraction using t-test statistics and kernel partial least squares. *ICONIP*, pp. 11-20.

Li ve diğ., 2007. Awavelet-based data pre-processing analysis approach in mass spectrometry. *Computers in Biology and Medicine*, 37,509-516.

Liu ve diğ., 2003. A comparative study on feature selection and classification metdos using gene expression proles and proteomic patterns. *Genome Informatics*, 13, 51-60.

Liu QZ, Sung AH, Qiao MY, Chen ZX, Yang JY, Yang MQ, Huang XD, Deng YP., 2009. Comparison of feature selection and classification for MALDI-MS data. *BMC Genomics*, 2009,10.(Suppl 1):S3.

Molinaro A. ve diğ., 2005. Prediction error estimation: a comparision of resampling methods. *Bioinformatics*, 21, 3301-3307.

Morton M.S., 2006. *Early Detection of Ovarian Cancer Using Gabor Wavelet Phase Quantization and Binary Coding* (Master Thesis). Indiana University,Indianapolis

Neves ve diğ., 2010. Diagnosis of ovarian cancer with proteomic patterns in serum using independent component analysis and neural network. *International Journal of Biological and Life Sciences*, 6:4, 200-203

Osuna E.E., Freund R., Girosi F., 1997. Support Vector Machines: Training and Applications, A.I. Memo No. 1602, C.B.C.L Paper No. 144, Massachusetts Institute of Technology and Artificial Intelligence Laboratory, Massachusetts.

Pengyi Yang, Zili Zhang, Bing B. Zhou, Albert Y. Zomaya, 2010. A clustering based hybrid system for biomarker selection and sample classification of mass spectrometry data. *Neurocomputing*, 73, 2317-2331.

Petricoin ve diğ., 2005. Serum proteomic patterns for detection of prostate cancer cancer. *J Natl Cancer Inst*, 94(20), 1576-1578

Poon TCW, Yip T-T, Chan ATC, Yip C, Yip V, Mok TSK, Lee CCY, Leung TWT, Ho SKW, Johnson PJ., 2003. Comprehensive proteomic profilling identifies serum proteomic signatures for detection of hepatocellular carcinoma and its subtypes. *Clinical Chemistry*,49:5, 752-760.

Qu,Y. Adam, B.-L., Yasui, Y., Ward, M. D., Cazares, L. H., Schellhammer, P. F., Feng, Z., Semmes, O. J., and Wright G. L., Jr., 2002. Zept Boosteddecision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminated prostate cancer from nonprostate patients. *Clin Chem*, 48,1835-1843.

Qu Y. ve diğ., 2003. Data reduction using a discrete wavelet transform in discriminant analysis of very high dimensional data. Biometrics, 59, 143-151.

Rashed ve diğ., 1995. Diagnosis of inborn errors of metabolism from blood by acylcarnitines and amino acids profiling using automated electrosprey tandem mass spectrometry. *Pediatric Research*. 38, 324-331.

Rosipal R., 2003. Kernel Partial Least Squares for Nonlinear Regression and Discrimination. Neural Network World, 13(3), 291-300.

Rosipal R. ve Trejoe L.J., 2001. Kernel Partial Least Squares Regression in Reproducing Kernel Hilbert Space, *Journal of Machine Learning Research,* 2: 97-123.

Rosipal R., Trejo L.J., ve Matthews B., 2003. Kernel PLS. SVC for Linear and Nonlinear Classification. *In Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003),* Washington DC, 2003

Tang ve diğ., 2010. Ovarian cancer classification identification based on dimensionality reduction for SELDI-TOF data. *BMC Bioinformatics*,11:109.

Taşkın V., Doğan B. ve Ölmez T., 2012. Yumurtalık kanserinin kütle spektrometresi verilerinden kısmi en küçük kareler yöntemi ile teşhisi. *Biyomut,* syf: 151-154.

Vapnik, V.N., 1995, The Nature of Statistical Learning Theory, Springer-Verlag, New York.)

Vreken ve diğ., 2000. Analysis of plasmenylethanolamines using electrospray tandem mass spectrometry and its applicaiton in screening for peroxisomal disorders. *J Inher Metab Dis*. 23,429-433.

Wold S., 1992. Non-Linear Partial Least Squares Modelling II: Spline Inner Function, *Chemometrics and Intelligent Laboratory System.,* 14, 71–84.

Wold S., Kettanch-Wold N. ve Skegerberg B., 1989. Non-Linear PLS Modelling. *Chemometrics and Intelligent Laboratory Systems,* 7,53–65

Wu ve diğ., 2003. Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics*,19, 1636-1643.

Yu ve diğ., 2005. Ovarian cancer identification based on dimensionality reduction for high-throughput mass spectrometry data. *Bioinformatics*, 21, 2200-2209

Yurdakul ve diğ., 2012. Akciğer kanseri ile serum protein profilleri arasındaki ilişkinin seldi-tof/ms yöntemi ile değerlendirilmes**.** *Turkiye Klinikleri Journal of Medical Science*. 32 (4).

Zhang ve diğ., 2006. Biomarker discovery for ovarian cancer using SELDI-TOF-MS. *Gynecol Oncol,* 102(1), 61-66.

Zhu ve diğ., 2003. Detection of cancer-specific markers amid massive mass spectral data. *PNAS,* 100(25), 14666-7149(5),752-760.

Zhukov TA, Johanson RA, Cantor AB, Clark RA, Tockman MS., 2003. Discovery of distinct protein profiles specific for lung tumors and pre-malignant lung lesions by SELDI mass spectrometry. *Lung Cancer*. 2003;40267-279

Url-1<http*://www.kanser.org/toplum/?action=kanser-turleri&kat_id=1*>,alındığı tarih: 30.12.2012

Url-2 < *http://ovariancancer.jhmi.edu/earlydx.cfm*>, alındığı tarih: 30.12.2012.

Url-3 < *http://www.cancer.gov/statistics/finding* >, alındığı tarih: 25.10.2012

Url-4 <www.bayar.edu.tr/besergil/kutle_1.pdf>, alındığı tarih: 25.9.2012

Url-5 < *http://www.bayar.edu.tr/besergil/* >, alındığı tarih: 12.09.2012

Url-6 < http://www.bayar.edu.tr/besergil/2_iyon_dedektorleri.pdf >, alındığı tarih: 25.9.2012

Url-7 <*http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp*>, alındığı tarih: 30.06.2012.

Url-8 < *http://www.yildiz.edu.tr/~nicoskun/*>, alındığı tarih: 30.12.2012.

Url-9 < *http://yunus.hacettepe.edu.tr/~roner/denbi.htm* >, alındığı tarih: 30.12.2012.

Url-10<*http://www.mathworks.com/help/wavelet/gs/introduction-to-the-wavelet-families.html#f3-1009153*>, alındığı tarih: 12.11.2012

Url-11<*http://astronomy.swin.edu.au/cosmos/E/Equivalent+Width*>, alındığı tarih: 12.11.2012

Url-12<http://web.cs.hacettepe.edu.tr/~oner/sunum/index.php?page=teori >, alındığı tarih: 25.9.2012

Url-13<*http://vellum.cz/~mikc/oss-projects/CarRecognition/doc/dp/node32.html* >, alındığı tarih: 12.11.2012

EKLER

EK-A: Çekirdek uzayında iç çarpım

EK-A

İki boyutlu bir örnek uzayında (**X**) tanımlı bir noktanın altı boyutlu bir çekirdek uzayına (**F**) taşınması işlemi, φ: **X➔F**, x ➔ φ(x) , tanımlanan bir φ fonksiyonu ile gerçekleştirilsin. İki boyutlu uzaydaki bir noktanın x = (x_1 , x_2), φ fonksiyonu ile taşınması işlemi eşitlik **(A 1.1)** de verildiği gibi tanımlansın.

$$\varphi_1 = 1, \varphi_2 = \sqrt{2}x_1, \varphi_3 = \sqrt{2}x_2, \varphi_4 = x_1^2, \varphi_5 = \sqrt{2}x_1x_2, \varphi_6 = x_2^2 \qquad \textbf{(A 1.1)}$$

Çekirdek uzayına diğer adıyla öznitelik uzayına taşınan iki noktanın bu uzaydaki öklid iç çarpımı şu şekilde hesaplanır:

$$\varphi(x)^T \varphi(\hat{x}) = (1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, \sqrt{2}x_1x_2, x_2^2)(1, \sqrt{2}\hat{x}_1, \sqrt{2}\hat{x}_2, \hat{x}_1^2, \sqrt{2}\hat{x}_1\hat{x}_2, \hat{x}_2^2)^T \qquad \textbf{(A 1.2)}$$

$$\varphi(x)^T \varphi(\hat{x}) = 1^2 + 2x_1\hat{x}_1 + 2x_2\hat{x}_2 + x_1^2\hat{x}_1^2 + 2x_1x_2\hat{x}_1\hat{x}_2, x_2^2\hat{x}_2^2) \qquad \textbf{(A 1.3)}$$

$$\varphi(x)^T \varphi(\hat{x}) = (1 + x_1\hat{x}_1 + x_2\hat{x}_2)^2 \qquad \textbf{(A 1.4)}$$

$$\varphi(x)^T \varphi(\hat{x}) = (1 + x^T\hat{x})^2 \qquad \textbf{(A 1.5)}$$

Yukarıda gerçekleştirilen işlem aslında örnek uzayında tanımlanan bir çok terimli fonksiyonun gerçekleştirilmesidir. Bu işlem en genel hali ile **(A 1.6)** da verildiği şekilde tanımlanabilir:

$$\varphi(x)^T \varphi(\hat{x}) = (1 + x^T \hat{x})^d \qquad \textbf{(A 1.6)}$$

Çekirdek hilesi olarak tanımlanan işlem, örnek uzayında tanımlanan bir *çekirdek fonksiyonu* yardımıyla yukarıdaki işlemlerin gerçekleştirilmesidir. Böylece öznitelik uzayındaki iç çarpımın daha basit bir yolu önerilmektedir. Bu durumda φ taşıma fonksiyonlarının ne olduğunun ve taşıma işleminin nasıl yapıldığının önemi kalmamaktadır. Öznitelik uzayında noktaların birbirleriyle olan ilişkileri belirlenen bir çekirdek fonksiyonu (k) ile tanımlanır. Bu işlem en genel hali ile eşitlik **(A 1.7)** de verildiği gibi ifade edilebilir.

$$k(x, \hat{x}) = \varphi(x)^T \varphi(\hat{x}) \qquad \textbf{(A 1.7)}$$

ÖZGEÇMİŞ

Ad Soyad: Vedat TAŞKIN

Doğum Yeri ve Tarihi: Karşıyaka - 1986

Adres: İTÜ. Elek. Elek. Fak. Oda No:3005

E-Posta: taskinv@itu.edu.tr

Lisans: TOBB Ekonomi ve Teknoloji Üniversitesi, Elek. Elek. Müh. (2008)

TEZDEN TÜRETİLEN YAYINLAR/SUNUMLAR

- V. Taşkın, B. Doğan, and T. Ölmez. " Yumurtalık Kanserinin Kütle Spektrometresi Verilerinden Kısmi en Küçük Kareler Yöntemi ile Teşhisi" , *Biyomedikal Mühendisliği Ulusal Toplantısı (BİYOMUT)*, 151-154, İstanbul, Ekim 2012 .

- Vedat Taşkın, Berat Doğan, Tamer Ölmez. "Prostate Cancer Classification from Mass Spectrometry Data by Using Wavelet Analysis and Kernel Partial Least Squares Algorithm", *International Journal of Bioscience, Biochemistry and Bioinformatics*, vol. 3, no. 2, pp. 98-102, 2013.