Gırtlak Kanseri Ameliyat Verilerinin K-means Yöntemiyle Analizi

E.Dinçer
esradincer@maltepe.edu.tr
Maltepe Üniversitesi
Bilişim Bölümü

N.Duru
nduru@kou.edu.tr
Kocaeli Üniversitesi
Bilgisayar Mühendisliği Bölümü

Özetçe

Bu çalışmada, veri madenciliğinde bir kümeleme tekniği olan k-means algoritmasını kullanan bir yazılım geliştirilerek, gırtlak kanseri ameliyat verilerinin analizinin yapılması amaçlanmıştır. K-means algoritması ile veri kümesi içindeki yoğunlaşmalar belirlenmiş ve grafik aracılığı ile görüntülenmiştir. Gerçek tıbbi veriler kullanılarak geliştirilen yazılım, karar vermede yardımcı bir araçtır, araştırma, denetim ve eğitim amacıyla kullanılabilir. Literatür taramasında benzer başka bir çalışmaya rastlanmamıştır.

Abstract

Analyzing the laryngeal cancer operations data by the k-means algorithm. In this study, a software tool was developed to analyze the laryngeal cancer operations data by using the k-means algorithm, which is a clustering algorithms in data mining. The algorithm was used to point out the intensities in the data set and to displayed in charts. The software which was run on real-world medical data, is a tool for taking desicions and it can be used to support research, to enhance supervision, and to aid in teaching activities. According to our search, there is not any application, which is similar to this.

1. Giriş

İnsanların deneyimlerden sonuç çıkartma yeteneği, geçmişten uygun örneklerin tanınması yeteneğine bağlıdır. Hastalıklara teşhis koyan bir doktor, öncelikle deneyimlerinden benzer vakaları tanımlar ve ardından bu vakaların bilgilerini eldeki probleme uygular. Bilinen vakaların sınıflandırılmış kayıtlarının tutulduğu veri tabanı, yeni vakaya benzeyen kayıtları bulmak için taranır. Mevcut hasta için en etkili tedavi, muhtemelen benzer hastaların sonuçlarından elde edilen bilgilerle yapılan tedavidir [1].

Bu çalışmada, gırtlak kanseri ameliyat verilerinin geçmiş kayıtlarını analiz ederek, tıp doktorlarının tıbbi bilgi elde etmesini sağlayan bir yazılım uygulaması geliştirilmiştir. Uygulamada, verilerin kümelenmesi ve verilerin içindeki yoğunlaşmaların görüntülenmesini sağlamak amacıyla veri madenciliğinde bir kümeleme algoritması olan k-means algoritması kullanılmıştır. Çalışmada kullanılan gırtlak kanseri ameliyat verileri Kocaeli Üniversitesi Tıp Fakültesi Hastanesi, Kulak, Burun ve Boğaz Bölümünden alınmıştır. Veri tabanı 1995 yılından bu yana toplanan 400 kaydı içermektedir. Algoritmanın seçilme nedenleri ve kümelenmiş görüntünün sağladığı yararlar 4.2'de açıklanmıştır.

Veri madenciliği, veri keşfi sürecinde bir adımdır. Büyük miktarda veriyi analiz ederek, içindeki kümelenmeleri, örüntüleri, birliktelikleri ve istisnaları keşfetmeye çalışma işlemidir. Endüstri, medya ve veri tabanı araştırma alanlarında veri madenciliği gittikçe daha fazla yer almaya başlamıştır [2]. Farklı bilgilerin madenlenmesi için birliktelik kuralı, sınıflandırma ve kümeleme gibi bir çok yöntem bulunmaktadır.

Tıbbi veri tabanlarında değişik amaçlarla yapılmış çok sayıda veri madenciliği çalışması bulunmaktadır. Bunlardan biri, göğüs kanseri ve cilt lezyonları verilerini sınıflandırmak için k-nearest, bayesian, karar ağacı ve Dempster-Shafer teorisi yöntemlerini birbirleriyle karşılaştıran çalışmadır [3]. Veri madenciliği algoritmalarını sınamak için bir standart haline gelen PIDD (Pima Indian Diabet Database) diabet veri tabanını kullanan Breault, diabet hastalığı konusunda yapılan tahminlerinin doğruluğunu göstermek için kaba kümeler (rough sets) yöntemini kullanmıştır [4]. Tıbbi veri tabanı üzerinde yazılım uygulaması geliştiren çalışmalara literatürde sık rastlanmamaktadır. Yapılmış çalışmalara bir örnek olarak; geliştirilen DMAP (Data Minig with Apriori) isimli bir yazılım aracılığı ile Apriori algoritması kullanılarak diabet hastalarının sosyal durumları ortaya çıkarılmıştır [5].

Tıp alanında k-means algoritmasını kullanan pek çok çalışma gerçekleştirilmiştir. Bunlardan bazıları şu şekilde özetlenebilir: İlaçların olumsuz etkileri konusunda risk faktörlerini araştıran bir çalışmada, her bir ilacı sınıflandırırken k-means algoritmasından yararlanılmıştır [6]. Diş hekimliğinde ameliyat sonrası akut ağrıların analizi için yapılan bir çalışmada, akut ağrı çeken hasta gruplarının özellikleri k-means algoritması ile ortaya çıkarılmıştır [7]. Psikiyatri alanında gerçekleştirilen bir çalışmada, antisosyal kişilik bozukluğu gösteren adli suçluların bilgileri k-means algoritması kullanılarak ve küme sayısı verilmeden analiz edilmiştir [8]. Gırtlak kanseri ameliyat verileri üzerinde, k-means kümeleme algoritması kullanılarak yapılan başka bir veri madenciliği çalışmasına literatür taramasında rastlanmamıştır.

2. Veri Madenciliği İle Veri Keşfi

Veri tabanlarındaki hızlı gelişme, işlenen verinin kullanışlı bilgiye otomatik dönüşümünü yapacak akıllı yeni araç ve teknolojilerin ortaya çıkmasına ihtiyaç oluşturmuştur. Bu nedenle bir araştırma alanı olarak veri madenciliğinin önemi gittikçe artmaktadır [9]. Veri madenciliğinde eldeki veri türüne ve elde edilen sonuçların kullanım amacına göre değişen farklı bir çok teknik bulunmaktadır. Bu çalışmada kümeleme tekniği kullanılmıştır. Kümeleme tekniğinde, veriler dağılımlarına göre irdelenerek, doğal sınıflandırmalar oluşturulur. Kümeleme işleminde temel prensip, sınıf içi benzerliği maksimum, sınıflar arası benzerliği minimum yapmaktır [10].

Çalışmada kullanılan k-means algoritması en iyi bilinen ve yaygın kullanılan bir kümeleme algoritması ve bölümleme tekniğidir [11]. İlk olarak J. MacQueen tarafından 1967 yılında tanıtılmıştır. Bu yöntem yıllardır bilimsel ve endüstriyel uygulamalarda en yoğun kullanılan kümeleme algoritması haline gelmiştir. Algoritmaya k-means adı verilmesinin nedeni, algoritmanın çalışmasından önce sabit bir küme sayısına ihtiyaç duyulmasıdır. Küme sayısı k ile gösterilir ve elemanlarının birbirlerine olan yakınlıklarına göre oluşacak grup sayısını ifade eder. Buna göre k önceden bilinen ve kümeleme işlemi bitene kadar değeri değişmeyen sabit bir pozitif tamsayıdır [12]. Kümeleme işlemi, verilerin en yakın veya benzer oldukları küme merkezleri (centroid) etrafına yerleştirilmesi ile gerçekleştirilir. Çalışma yönteminde genellikle, öklit bağıntısı temel alınarak kümeleme yapılmaktadır. Algoritmanın başında k sayısı giriş parametresi olarak verilir. Eğer küme sayısı belirli değil ise deneme yoluyla en uygun sayı bulunur veya bu değer algoritmaya dışardan verilir. K adet rastgele küme merkezi belirlenir veya ilk k eleman merkez olabilir. Elemanların merkezlere yakınlıkları hesaplanarak, yakın oldukları merkezlere göre kümeleme yapılır. Oluşan kümelerin ortalamaları hesaplanarak yeni küme merkezleri belirlenir. Bu işlem kümelenecek eleman kalmayıncaya kadar sürer [13].

Çok yaygın kullanımı olan bu algoritmanın zayıf yanları da bulunmaktadır: öncelikle algoritmanın başında giriş parametresi olarak bir k sayısının verilmesine gerek vardır. Elde edilecek sonuçlar k

sayısına göre değişkenlik gösterebilir. Eğer küme sayısı belirli değil ise deneme yoluyla en uygun sayı bulunur. Aşırı gürültü ve istisna veriler algoritmayla hesaplanan ortalamayı değiştirdiği için k-means algoritması gürültü ve istisnaya karşı çok duyarlıdır. Algoritma çakışan kümelerde iyi sonuç vermemektedir ve sadece sayısal veriler ile kullanılabilmektedir [14].

3. Gırtlak Kanseri Hakkında Genel Bilgiler

Kanser, anormal hücrelerin kontrolsüz çoğalması ve yayılması olarak bilinen bir grup hastalığa verilen isimdir. Gırtlak kanseri ise, Kulak Burun Boğaz Hekimliğinde en sık görülen kanser türlerinden biridir. Gırtlak (larenks), boğazın hemen altında ses tellerinin bulunduğu bir organdır ve gıda alımı sırasında besinlerin nefes borusuna kaçmasını engeller. Gırtlak kanseri, gırtlağın herhangi bir kısmında gelişebilir ve çoğu zaman ses kısıklığı ile erken bulgu verir. Genellikle 50-60 yaş grubundaki erkeklerde sık görülür. Sigara bu kanser türü için en önemli risk etkenidir. Yoğun alkol kullanımı da riski arttırır. Tedavi tümörün türü, yeri ve evreye göre belirlenir. En önemli tedavi şekli cerrahidir ve bunun yanında ışın tedavisi kullanılabilir.

TNM sınıflaması, malignant (kötü huylu) tümörler için standart bir kanser evrelendirme sistemidir. Bir çok tümörün kendi TNM sınıflaması bulunur. Sınıflandırmanın genel şekli şöyledir: Zorunlu parametreler T, N ve M'dir. T: Tümör büyüklüğünü temsil eder ve 0 ile 4 arası değer alır. N: Node değeri, tümörün lenflere yayılımını ifade eder ve 0 ile 3 arası değer alır. M: Metastas, 0 veya 1 değeri alır ve diğer organlara yayılımını ifade eder. Gırtlak kanseri vakalarında dikkate alınmadığından metastas parametresi veri tabanında tutulmamaktadır.

Tedavide patolojik ve diğer görüntü verilerine dayanarak tümör hakkında tahmini bilgi edinilir ve kanserin evresi öngörülür. Bu bilgiler ışığında yapılacak ameliyat şekline karar verilir. Hastanın yaşı da, göz önünde tutulan bir etkendir. Ameliyat sırasında nadiren de olsa öngörülen bilgilerden farklı bir görüntüyle karşılaşılabilir. Tümör büyüklüğü öngörülenden farklı olabilir ve kanser evresi tahmin edilenden daha yüksek olabilir. Bu durum ameliyat sırasında uygulanan tekniğin değiştirilmesini gerektirir. Ameliyatla alınan tümör daha sonradan nüks edebilir. Hastaların durumu daha sonraki yıllarda da izlenir ve tümör 5 yıl içinde nüks etmez ise hastanın sorundan kurtulduğu kabul edilir [13].

4. K-Means Yöntemiyle Verilerin Analizi

4.1 Veri Tabanının Hazırlanması

Diğer bir çok veri tabanında olduğu gibi tıp veri tabanlarının da program aracılığı ile kullanılması için bozuk verilerden temizlenmesi ve düzenlenmesi gerekmektedir. Hatalı girilmiş bilgilerin olması sık rastlanır bir durumdur [14]. Çalışmada kullanılan veri tabanında rastlanan hatalar şunlardır: Boş bırakılmış alanlar, birkaç bilginin birleştirilerek aynı alana yazılması, bir durumun değişik kayıtlarda birkaç farklı isimle yer alması, benzer şekilde aynı anlama gelen bilgilerin farklı formatlarda yazılması, yanlış yazılmış tıbbi terimler.

Bu çalışmada kullanılan gırtlak kanseri ameliyat bilgilerinin tutulduğu veri tabanı öncelikle bozuk verilerden temizlenmiş ve program aracılığı ile kullanılabilir şekilde düzenlenmiştir. Veriler Excel tablosu şeklinde alınmış ve Access veri tabanına aktarılmıştır. Düzenlemeler için Access sorgu nesnesi kullanılmıştır. Çalışmada kullanılmak amacıyla düzenlenen veri tabanı Tablo 1'de yer almaktadır. Veri tabanında aşağıdaki işlemler yapılmıştır:

Hasta ismi ve adresi gibi analiz için gerekli olmayan alanlar çıkartılmıştır. Boş (null) değer içeren sayısal alanlara 0 değeri atanmıştır. K-means algoritmasında karakter alanlar kullanılamadığından, patoloji, survive ve operasyon alanları sayısal karşılıklarına çevrilmiştir. Patoloji sonucu yassı epitel hücre (tabloda "Yeh Ca" olarak kısaltılmıştır) ise bu alanın değeri 1, diğer sonuçlar için 0 yapılarak veri tabanına yeni eklenen patoloji_kodu alanına yerleştirilmiştir. Nüks bilgisi, ölüm nedeni ve tarihini içeren survive alanından nüks ve hayatta bilgileri alınarak veri tabanına yeni eklenen aynı isimli alanlara yerleştirilmişlerdir. Nüks varsa nüks isimli alanın değeri 1, yoksa 0 yapılmıştır.

Benzer şekilde, hasta hayatta ise hayatta isimli alanın değeri 1, aksi durumda 0 yapılmıştır.

Tablo 1: Çalışmada kullanılan gırtlak kanseri ameliyat bilgileri veri tabanı

Yas	Patoloji	preop_T	preop_N	pre_evre	postop_T	postop_N	post_evre	Opr	Opr_tarihi	nuks	hayatta	ptj_kodu
45	YEH Ca	T2	N0	2	T2	N1	3	FLL	14.02.1996	0	1	1
33	epidermoi	T4	N2	4				TL	02.09.1997	0	0	0
62	Ağır displ	T1	N0	1				KRD	11.03.1997	0	1	0
44	YEH Ca	T3	N1	3	T3	N2	4	NTL	23.02.1998	0	0	1
60	YEH Ca	T3	N0	3				TL	18.12.1997	0	1	1
61	Ağır displ	T1	N0	1				KRD	10.05.1996	0	0	0
50	YEH Ca	T3	N1	3	T3	N0	3	TL	24.07.1996	0	0	1
48	YEH Ca	T1	N0	1				KRD	18.07.1998	0	1	1
59	YEH Ca	T3	N2	4	T4	N2	4	TL	07.07.1998	0	0	1
51	YEH Ca	T3	N0	3	T4	N2	4	TL	21.11.2000	1	0	1
45	YEH Ca	T2	N0	2	T2	N1	3	FLL	14.02.1996	0	1	1
56	Epidermoi	T1	N0	1				SGL	03.04.2001	0	0	0

Veri tabanındaki alanların açıklamaları:
Yaş: Hastanın yaşı.
Patoloji: Ameliyat öncesinde tespit edilen patolojik tetkik sonucu. Yeh yassı epitel hücre anlamına gelmektedir.
Preop_T: Teşhis sırasında tespit edilen tümör büyüklüğü.
Preop_N: Teşhis sırasında tespit edilen node (lenf) büyüklüğü.
Preop_evre: Preop_T ve Preop_N değerlerinden hesaplanan kanser evresi.
Postop_T: Ameliyat sırasında görülen tümör büyüklüğü.
Postop_N: Ameliyat sırasında görülen node (lenf) büyüklüğü.
Postop_evre: Postop_T ve Postop_N değerlerinden hesaplanan kanser evresi.
Opr: Yapılan ameliyatın ismi. Veri tabanında yedi çeşit ameliyat bulunmaktadır.
Opr_tarihi: Ameliyatın yapıldığı tarih.
Nüks: Ameliyat sonrası tümör nüks ederse bu alanın değeri 1 olur. Aksi durumda 0'dır.
Hayatta: Ameliyat sonrası hasta hayatta ise bu alanın değeri 1 olur. Aksi durumda 0'dır.
Ptj_kodu: Ameliyat öncesinde tespit edilen patolojik tetkik sonucu "Yeh Ca" ise bu alanın değeri 1 olur. Diğer bütün sonuçlar için 0'dır.

4.2 Analiz Aracının Geliştirilmesi

Bu çalışmada k-means algoritması ile veriler kümelenerek, içindeki yoğunlaşmalar belirlenmiş ve grafik aracılığı ile görüntülenmiştir. En sık kullanılan yöntem olması nedeniyle küme elemanları arasındaki uzaklığın hesaplanmasında öklit bağıntısı tercih edilmiştir. K-means algoritması aşağıdaki özellikleri nedeniyle tercih edilmiştir:

Küme sayısı olan k değeri parametre olarak algoritmaya dışardan verilmektedir. Verilerin kaç kümeye ayrılacağının net olarak bilinmediği durumlarda farklı değerler vererek sonuçları izlemek mümkün olmaktadır. Bu olanak uygulamanın analizini esnek hale getirmektedir.
Algoritmanın uygulanması kolaydır ve hızlı çalışmaktadır [11]. Veriler matristen okunarak algoritmaya verilmektedir. Veri sayısına oranla iterasyon sayısı düşüktür.

Değişik dağılımlarda başarılı sonuçlar alınabilmektedir. Veriler birbirine çok yakın veya çok uzak değilse, kümeleme işlemleri başarıyla gerçekleştirilmektedir.

Kategorik verilerle çalışacak şekilde adapte edilebilmektedir. Algoritma rakamsal verilerle çalışmaktadır. Ancak kategorik veriler uygun rakamsal karşılıklarına çevrilerek algoritma tarafından işlenebilir.

Kümeleme sonuçları hem grafik olarak hem de yazı ve rakamlarla kolayca ifade edilebilmektedir. Algoritma sonucunda kümelenen veriler matrise yerleştirilmektedir. Buradan grafiğe kolayca aktarılabilmekte, ve her küme farklı renkle gösterilebilmektedir. Kümelerin eleman sayıları ve diğer verilere olan oranları kolayca hesaplanıp, yazılı olarak ifade edilebilmektedir.

Uygulama, Borland Delphi 6.0 kullanılarak Windows XP işletim sistemi üzerinde geliştirilmiştir. Veriler Access veri tabanında tutulmuştur.

Uygulama aracılığı ile veriler dört farklı açıdan analiz edilmiştir. Her bir analiz ayrı bir arayüz ile görüntülenmiştir. Arayüz ekranlarına şekil 1'de görülen bir anamenü üzerinden erişilmektedir. Oluşturulan arayüzler şöyledir:

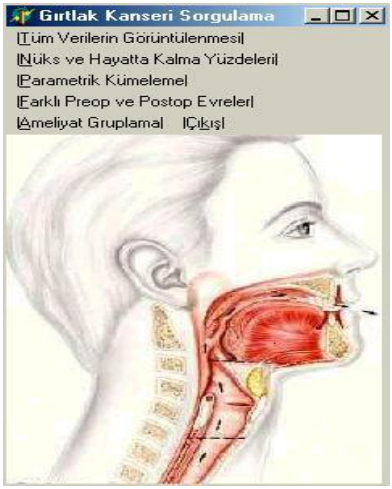
1. Nüks ve Hayatta KalmaYüzdeleri: Gelecek vakaların tahminlerinde kullanılmak amacıyla, geçmiş vakaların tümör nüks etme yüzdeleri ve

hayatta kalma yüzdeleri incelenir. Sadece bu arayüzde k-means algoritması kullanılmamıştır.

2. Parametrik Kümeleme: Veri tablosunun seçilen iki alanı arasındaki etkileşim izlenir.

3. Farklı Preop ve Postop Evreler: Doğru öngörülen ve öngörülemeyen ameliyat öncesi evreler incelenerek ameliyat öncesi tahmin başarısı değerlendirilir.

4. Ameliyat Gruplama: Başarılı ameliyatlar izlenerek, gelecek ameliyat tercihlerine karar verilir.



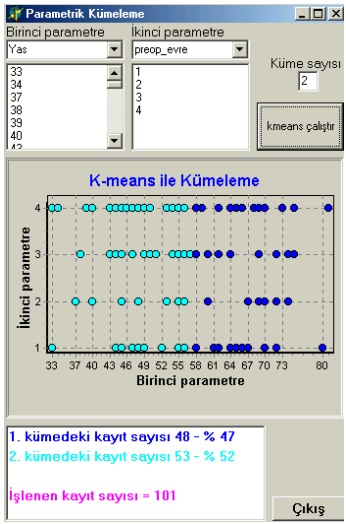
Şekil 1: Uygulamanın ana menüsü

Kümelenmiş veriler ve içlerindeki yoğunlaşmalar grafik üzerinde gösterilmiştir. Arayüzlerde k-means algoritması ile oluşturulacak küme sayısı, kullanıcı tarafından 2 ile 9 arasında bir değer ile belirlenebilmektedir. Sınama sonuçlarına göre, 9 kümeden daha fazlası verimli olmamaktadır. Kullanıcı "K-means çalıştır" tuşunu tıklayarak kümeleme sonucunu grafik üzerinde görüntüleyebilmektedir. Her bir küme ayrı bir renk ile gösterilerek birbirinden ayrıştırılır.

Kümelenmiş sonuçların kısa açıklaması arayüz ekranlarının altındaki bir pencereden görüntülenmektedir. Açıklama, her bir kümedeki kayıt sayısını ve bu sayının işlenen tüm kayıtlara oranını içermektedir. Her bir açıklama satırı, grafikteki kümesi ile aynı renktedir. Bu kısımda ayrıca kümeleme sonucunun yazılı ifadeleri ile arayüze ilişkin mesajlar ve açıklamalar bulunmaktadır.

4.2.1 Parametrik Kümeleme

Bu arayüzde geriye dönük inceleme kolaylığı ve ileriye dönük karar verme desteği sağlanması amaçlanmıştır. Kullanıcı, değişken parametreler kullanarak değerlendirme yapabilmektedir. Geçmiş veriler analiz edilerek, mevcut ve gelecek vakalar için tahminde bulunulabilmektedir. Veri tabanından iki alan seçilerek, bunların değerleri arasındaki etkileşim izlenmektedir. Örneğin yaş ile tümör boyutu veya yaş ile uygulanmış ameliyatlar gibi. Seçilen veriler k-means algoritması ile kümelenerek, veri içindeki yoğunlaşmalar kullanıcıya görüntülenmektedir. Bu arayüz geliştirilen uygulama içinde, kullanıcının alanlar seçerek işlem yapabildiği tek arayüzdür. Eğer kullanıcı birbiriyle ilişkisiz alanlar seçerse, analiz sonucu anlamsız olabilir.



Şekil 2: Parametrik kümeleme arayüzü

Arayüz görüntülendiğinde şekil 2'de görüldüğü gibi, veri tabanındaki bütün alanların isimleri üstte yer alan, iki açılan kutu içinde listelenir. Kullanıcı, alan isimlerinin üstünü tıklayarak herhangi iki

tanesini seçer. Seçim sonrası alanların değerleri, her bir değerden bir tane olacak şekilde liste kutusunda listelenir.

Seçilen alanların grafik üzerinde kümelenmiş görüntüsünün oluşturulması için, kullanıcı küme sayısını girerek "K-means çalıştır" tuşunu tıklar. Grafik üzerindeki her bir seri seçilen bir alanı temsil eder.

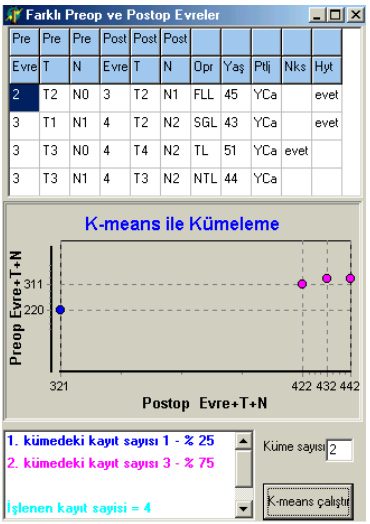
Şekil 2'deki örnek görüntüde, yaş ve ameliyat öncesi evre alanları seçilmiş, ilgili veriler k-means algoritması ile iki kümeye ayrılmıştır. Bu seçimle, ameliyat öncesi öngörülen evrelerin yaşlarla ilişkisi incelenmektedir. Kayıt sayıları birbirine yakın iki kümeden birincisi 33 ile 57 yaş arası, ikincisi ise 57 ile 83 yaş arasındaki vakaları içermektedir. Birinci kümede, 43 yaş sonrası dördüncü evredeki vakaların çokluğu dikkat çekmektedir. Dördüncü evreyi, yoğunluk açısından üçüncü ve birinci evreler izlemektedir. İkinci kümede ise 57 yaş ile 70 yaş arası dördünce evrede, 61 ve 67 yaş arasında birinci evrede yoğunluk izlenmektedir.

Program kodunda, ameliyat ismi ve patoloji kodu gibi bütün karakter alanlar k-means algoritmasıyla işlenebilmek için sayısal karşılıklarına dönüştürülmüştür.

4.2.2 Farklı Preop ve Postop Evreler

Bu arayüzde doğru öngörülen ve öngörülemeyen ameliyat öncesi evrelerin görüntülenerek incelenmesi ve bu şekilde ameliyat öncesi tahmin başarısının değerlendirilmesi amaçlanmıştır. Kanser ameliyatlarında, nadiren de olsa ameliyat öncesinde tahmin edilen evre, ameliyat sırasında görülen evreden farklı olmaktadır. Bu da ameliyat öncesi evreye göre yapılan hazırlığın, ameliyat sırasında değiştirilmesi ve yeni kararlar alınması anlamına gelmektedir. Gerçek hayatta az rastlanır bir durum olduğundan, veri tabanında bu inceleme için listelenen kayıt sayısı azdır. Şekil 3'deki örnekte sadece 4 kayıt görüntülenmektedir.

Burada evre değerleri, ameliyat öncesinde ve sonrasında birbirinden farklı kayıtlar listelenmektedir. Koşulu sağlayan her bir vaka için listelenen bilgiler şunlardır: Ameliyat öncesi ve sonrasına ait evre değeri, tümör ve node kodları, uygulanan ameliyatın kısa adı, hastanın yaşı, patoloji kodu, nüks ve hayatta bilgileri. Grafikte ameliyat öncesi ve sonrasına ait evre, tümör ve node değerleri karşılaştırılmaktadır.



Şekil 3: Farklı preop ve postop evreler arayüzü

Program kodunda, evre, tümör ve node değerlerinin k-means algoritmasında kullanılabilmesi için alanlar birleştirilerek tek bir sayı haline getirilmiştir. Birleştirmede izlenen yol şöyledir: alanlar analizdeki önemlerine göre sıralanmış ve sayısal değerlerini koruyarak diğerleriyle karşılaştırabilmek için 1, 10 ve 100 katsayı değerleriyle çarpılarak birbiriyle toplanmıştır; (evre kodu)*100+(tümör değeri)*10+(node değeri). Ameliyat öncesi ve sonrası için bu şekilde oluşturulan sayılar k-means algoritmasının her bir boyutunu oluşturur. Şekil 3'de görülen örnekte, ilk kaydın preop evresi 2, tümör değeri 2 ve node değeri 0'dır. Bu değerler grafikte 220 sayısına karşılık gelmektedir. Aynı şekilde postop evresi 3, tümör değeri 2 ve node değeri 1'dir. Bu da grafikte 321 sayısına karşılık gelmektedir. Şekil 3'de listelenen 4 kayıt iki kümeye ayrılmıştır. Postop evresi 4 olan kayıtlarla preop evresi 3 olan kayıtlar üç elemanlı büyük kümeyi oluşturmuşlardır. Grafikten görüldüğü gibi, ameliyat öncesinde üçüncü evrede olduğu öngörülen vakalar, ameliyat sırasında dördüncü evrede bulunmuştur. Öngörü farkı, en çok üçüncü

evrede, patolojisi “yehca” olan 40-50 yaş arası vakalarda ortaya çıkmıştır.

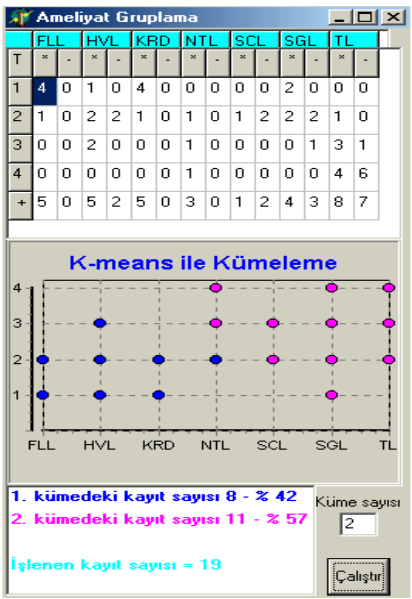
4.2.3 Ameliyat Gruplama

Bu arayüzde başarılı ameliyatların izlenerek, gelecek ameliyat tercihlerine destek sağlanması amaçlanmıştır. Burada gırtlak kanseri ameliyatları, vakalara göre kümelenmiştir. Vakalar ameliyat öncesi tümör değerleri ile temsil edilmektedir. Kümeleme sırasında aynı tümör değerine sahip vakaların sayıları belirlenmiştir.

Aynı tümör değerine sahip her vaka için başarılı ve başarısız ameliyatların sayıları hesaplanmıştır. Hesaplama yöntemi şöyledir; eğer ameliyat sonrasında tümör nüks etmiş ise ameliyat başarısız kabul edilir, eğer tümör nüks etmemiş ve hasta hayatta ise ameliyat başarılı kabul edilir. Böylece arayüzde, gerçekleştirilen ameliyatlar görüntülenerek, hangi tümör değerlerine hangi ameliyatların ne çoklukta yapıldığı ve başarılı/ başarısız sayıları izlenir. Başarılı ve başarısız ameliyatların sayıları grafiğe yansıtılmamıştır.

Arayüz vasıtasıyla, kümelenmiş ameliyatlar listelenir. Birçok ameliyatın ismi uzun olduğu için isimler yerine kısaltmalar kullanılmıştır. Şekil 4’de görülen tabloda ameliyat öncesi tümör değerlerine karşılık olarak, başarılı ve başarısızlık hesaplaması her bir ameliyat için ayrı ayrı listelenmektedir. Başarılı sütunu ‘*’ ile, başarısız sütunu ise ‘-‘ ile gösterilmiştir. Şekildeki örnekte FLL kodlu ameliyat, T1 için 4 kez, T2 için 1 kez başarıyla yapılmıştır. Bu ameliyatta hiç başarısız vaka kaydedilmemiştir. Son satırda bütün ameliyatların başarılı ve başarısız sayıları toplanmıştır. Örnekte, FLL kodlu ameliyatın başarı sayısı 5, başarısız sayısı 0’dır. Bu tabloda, veri tabanında nüks ve hayatta alanları dolu olan vakalar hesaplamaya dahil edilmiştir. Her kayıt için bu alanlar dolu olmadığından, burada listelenen kayıt sayısı, veri tabanındaki toplam kayıt sayısından daha azdır.

Şekil 4’deki grafikte ameliyatlar ile ameliyat öncesi tümör değerleri karşılaştırılmaktadır. Burada tümör tiplerine uygulanan ameliyat çeşitleri kümelenmiştir. Kümeleme için veri tabanından okunan vakaların nüks ve hayatta alanları dikkate alınmamıştır. Grafikte gösterilmediğinden tümör tiplerine uygulanan ameliyat sayıları hesaplanmamıştır. Bu nedenle tabloda listelenen ile grafikte ele alınan kayıt sayıları farklıdır. Tabloda 45 kayıt için listeleme yapılırken, grafikte 19 kayıt kümelenmiştir. Şekildeki örnekte, ameliyatlar iki kümeye ayrılmıştır. Sayıca az olan ilk kümede, daha çok 1 ve 2. derece tümörlere uygulanan üç ameliyat yer almıştır. Büyük kümede ise, 2. derecenin üstündeki tümörlere uygulanan ameliyatlar bulunmaktadır. 2. derece tümörlere bütün ameliyat çeşitlerinin uygulandığı, SGL kodlu ameliyatın bütün tümör türlerine uygulanabildiği görülmektedir. Diğer taraftan 2., 3. ve 4. derecelere uygulanan NTL ve TL kodlu ameliyatlar, başarı rakamlarına göre incelendiğinde, TL’nin daha riskli bir teknik olduğu göze çarpmaktadır. Buna rağmen TL, sayıca NTL’den fazla gerçekleştirilmiştir.



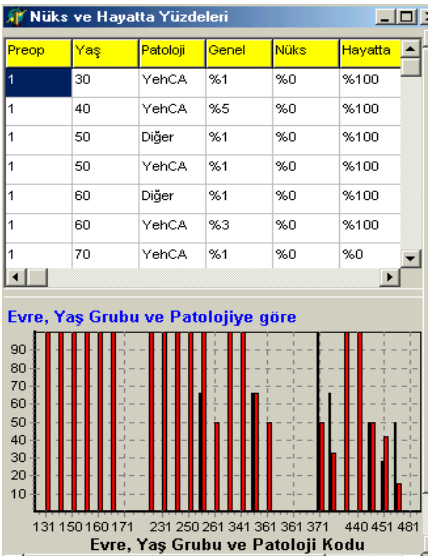
Şekil 4: Ameliyat gruplama arayüzü

Program kodunda ameliyatlar k-means algoritmasıyla işlenebilmek için sayısal karşılıklarına dönüştürülmüştür. Bunun için ameliyat isimlerine alfabetik sırada birden yediye kadar sıra numarası verilmiştir.

4.2.4 Nüks ve Hayatta Kalma Yüzdeleri

Bu arayüzde diğer üç arayüz ekranından farklı olarak veri kümeleme yerine sınıflama yapılmıştır. Analizi yapılacak kayıtların algoritmaya uygulanamaması nedeniyle, burada k-means algoritması kullanılmamıştır.

Bu arayüzün amacı, mevcut ve gelecek vakalar için, ameliyat sonrasında tümörün nüks etme olasılığının ve hastanın hayatta kalma olasılığının tahmin edilmesine destek sağlamaktır. Bunun için geçmiş vaka verileri sınıflandırılarak, tümör nüks etme yüzdeleri ve hayatta kalma yüzdeleri görüntülenir. Veriler, hasta yaş grubu, ameliyat öncesi evresi ve patolojik bilgiye göre sınıflanmıştır. Yüzdeler her sınıf için ayrı ayrı hesaplanmıştır.



Şekil 5: Nüks ve hayatta kalma yüzdeleri arayüzü

Arayüz görüntülendiğinde hesaplanan sonuçlar listelenir. Her bir sınıftaki kayıt sayısının bütün kayıtlara oranı hesaplanarak "Genele Oranı" adı altında ayrı bir sütunda listelenmiştir. Grafikte çubuk (bar) seri tipi kullanılmıştır. Şekil 5'de görüldüğü gibi her bir çubuk bir sınıfı temsil etmektedir. Grafikte görüntülenen yeşil renkli seri, her bir sınıfın tümör nüks yüzdesini, kırmızı renkli seri ise hayatta kalma yüzdesini göstermektedir.

Sınıflar hasta yaş grubu, ameliyat öncesi evresi ve patolojik bilgilerin birleştirilmesinden oluşturulmuştur. Yaş grubunda onlar basamağı aynı olan yaşlar gruplanmıştır. Örneğin 50-59 arası 50 grubu, 60-69 arası 60 grubu gibi. Birleştirilen alanların bilgileri sayısal karşılıklarına dönüştürülerek, her bir sınıf üç basamaklı bir sayı ile temsil edilmiştir. Sayının yüzler basamağını evre kodu, onlar basamağını yaş grubu bilgisinin onlar basamağı, birler basamağını değeri 0 veya 1 olan patoloji kodu oluşturmaktadır. Program kodunda sayının hesaplanmasında izlenen yol şöyledir: alanlar analizdeki önemlerine göre sıralanmış ve sayısal değerlerini koruyarak diğerleriyle karşılaştırabilmek için 1, 10 ve 100 katsayı değerleriyle çarpılarak birbiriyle toplanmıştır; (evre kodu)*100+(yaş grubunun onlar basamağı)*10+ (patoloji kodu). Örneğin evre kodu 1, yaş grubu 70 ve patolojik tetkik sonucu "Yeh Ca" ise bu sınıfın kodu 171'dir. Evre kodu 2, yaş grubu 60 ve patolojik tetkik sonucu "Diğer" ise bu sınıfın kodu 260'dır.

Örneğin Şekil 5'de görülen ilk satırda, birinci evrede olan ve patolojik tetkik sonucu "YehCa" (yassı epitel hücre) olan 30 yaş grubundaki vakalar yer almaktadır. Bu gruptaki vakalar sayıca bütün kayıtların yüzde birini oluşturmaktadır. Vakaların hiç birinde tümör nüksü kaydedilmemiş ve vakaların tamamı hayattadır. Bu grup 131 kodu ile grafikte yer almıştır. Nüks oranı 0 olduğundan yeşil seri görüntülenmemektedir. Hayatta oranı 100, kırmızı seri ile gösterilmiştir. Yedinci satırda, birinci evrede olan ve patolojik tetkik sonucu "YehCa" olan 70 yaş grubundaki vakaların nüks ve hayatta bilgileri veri tabanında yer almadığı için oranları 0'dır.

5. Sonuç

Mevcut hasta için en etkili tedavi, muhtemelen benzer hastaların sonuçlarından elde edilen bilgilerle yapılan tedavidir. Bu çalışmada, tıp alanında geçmiş kayıtları kullanmanın önemi dikkate alınarak, gırtlak kanseri ameliyat verileri için bir analiz aracı geliştirilmiştir. Çalışmanın amacı, veri madenciliğinde bir kümeleme tekniği olan k-means algoritmasını kullanarak geliştirilen bir yazılım aracılığıyla, gırtlak kanseri ameliyat verilerinin analizini yapmaktır. Geliştirilen

yazılımın gerçek veri tabanı üzerinde kullanılması, uygulamanın etkinliğini görmeyi sağlamıştır. Yazılım, tıp doktorlarının geçmiş kayıtları hasta dosyalarını tek tek taramaya gerek kalmadan analiz ederek, ileriye dönük tahminde bulunabilmelerini kolaylaştırmaktadır. Karar almada yardımcı olabilecek bir araçtır. Literatür taramasında, gırtlak kanseri ameliyat verileri üzerinde, k-means algoritması kullanarak yapılan başka bir veri madenciliği çalışmasına rastlanmamıştır.

Bu çalışmada k-means algoritması ile veriler kümelenerek, içindeki yoğunlaşmalar belirlenmiş ve grafik aracılığı ile görüntülenmiştir. Küme sayısının, kullanıcı tarafından 2 ile 9 arasında bir değer olarak belirlenmesine olanak sağlanmıştır. Sınama sonuçlarına göre, 9 kümeden daha fazlası verimli olmamıştır. Çalışmada kullanılmak üzere k-means algoritmasının seçilmesinde şu nedenler etkili olmuştur; Küme sayısının okunan bir parametre olması analizi esnek hale getirmektedir. Algoritmanın uygulanması kolaydır ve hızlı çalışmaktadır. Değişik dağılımlarda başarılı sonuçlar alınabilmektedir. Kategorik verilerle çalışacak şekilde adapte edilebilmektedir. Kümeleme sonuçları hem grafik olarak hem de yazı ve rakamlarla kolayca ifade edilebilmektedir. Gırtlak kanseri ameliyatlarının verileri çok boyutlu olmadığından ve noktalar düzgün dağıldığından kümeleme yüksek iterasyon gerektirmemiştir.

Geliştirilen yazılım aracılığı ile veriler analiz edilerek aşağıdaki yararlar elde edilir;

1. Değişken parametreler kullanılarak geçmiş verileri analiz edilir ve bilgiler mevcut vakaların değerlendirilmesinde dikkate alınır,
2. Mevcut ve gelecek vakalar için ameliyat sonrasında tümörün nüks etme olasılığı ve hastanın hayatta kalma olasılığı değerlendirilir,
3. Doğru öngörülen ve öngörülemeyen ameliyat öncesi evreler görüntülenerek incelenir ve bu şekilde ameliyat öncesi tahmin başarısı değerlendirilir,
4. Başarılı ameliyat bilgileri izlenerek, gelecek ameliyat tercihlerinde fikir alınır,
5. Yazılım araştırma, denetim ve eğitim etkinliklerinde de kullanılabilir.

Veri madenciliği çalışmalarında genellikle veriler çeşitli kısıtlar içeren paket programlar aracılığı ile analiz edilmektedir. Kullanıcı açısından bu çalışmada geliştirilen yazılımın öğrenme süresi çok kısadır ve kullanılması kolaydır. Uygulamanın tıp doktorlarının kullanımına uygun şekilde verileri çeşitli açılardan analiz etmesi hedeflenmiştir. Çalışmada kullanılan gırtlak kanseri ameliyat verileri, Kocaeli Üniversitesi Tıp Fakültesi Hastanesi, Kulak, Burun ve Boğaz Bölümünden alınmıştır.

5.1 Gelecekteki Çalışmalar İçin Öneriler

Veri madenciliğinde yeni gelişen teknolojilerin çoğu henüz tıp alanında kullanılan yazılımlara dahil edilmemiştir. Tıbbi veri tabanları üzerinde yazılım uygulaması geliştiren çalışmaların azlığı dikkat çekmektedir. Bu çalışma, tıp alanında geliştirilecek ve veri madenciliği algoritmalarını içerecek yazılımlar için bilgilendirici bir kaynak olabilir ve bu çalışma ile, çeşitli tıbbi verilerin değerlendirilmesi ve analizine kümeleme algoritmalarının katkısı gözlenebilir.

Tıp alanında veri madenciliği uygulamalarının ilerlemesinin önünde bazı engeller bulunmaktadır. Öncelikle çalışmalarda kullanılacak verilerin bulunmasında zorluk yaşanmaktadır. Tıp alanındaki mevcut verilerde belli bir standardın olmayışı verilerin işlenmesini zorlaştırmaktadır. Çalışmanın yönlendirilmesi için konuyla ilgili ve bilgisayar uygulamaları konusunda bilgi sahibi bir tıp doktorunun bulunması ve destek alınması gerekebilir. Bu zorlukların en aza indirilmesi, tıp alanında veri madenciliği uygulamalarının artmasını sağlayacaktır.

Kaynakça

[1] Berry, M.; Linoff, D.: “Data Mining Techniques”, Wiley Publishing Inc.,2004

[2] G. Piatetsky-Shapiro and W. J. Frawley. Knowledge Discovery in Databases. AAAI/MIT Press, 1991.

[3] Aslandogan, A., Mahajani, G., Taylor, S.: "Evidence Combination in Medical Data Mining," itcc, p. 465, International Conference on Information Technology: Coding and Computing (ITCC'04) Volume 2, 2004.

[4] Breault, J. L.: Data Mining Diabetic Databases: Are Rough Sets a Useful Addition?,
Computing Science and Statistics, Vol:34, 2001.
[5] Duru, N; An Application of Apriori Algorithm on a Diabetic Database, R. Khosla et al. (Eds.): KES 2005, LNAI 3681, pp. 398.404, 2005. Springer-Verlag Berlin Heidelberg 2005
[6] Evans, S., Lloyd, J., Stoddard, G., Nekeber, J., Samone, M.: Risk Factors For Adverse Drug Events. The Annals of Pharmacotherapy: Vol. 39, No. 7, pp. 1161-1168, 2005.
[7] Vickers E., Boocock H., Harris R., Bradshow J., "Analysis of the acute postoperative pain experience following oral surgery", Australian Dental Journal, 51(1):69-77, (2006 Mar).
[8] Morana H., Camara F., Arboleda-Florez J., "Cluster analysis of a forensic population with antisocial personality disorder", Forensic Science International, (2006, jan 23).
[9].Han, J.; Kamber, M.: "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers Inc., 2001
[10] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R.: Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996.
[11] Roiger, R.J., Geatz M. W.: Data Mining, A Tutorial-based Primer, Addison Wesley,2003.
[12] McQueen, J.: "Some Methods for classification and Analysis of Multivariate Observations", Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1:281-297, 1967
[13] http://en.wikipedia.org
[14] Kantardzic M. : Data Mining, Concepts, Models, Methods, and Algorithms IEEE Press,2001.