YALOVA ÜNİVERSİTESİ BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

Genetik Korelasyona Göre Sağlıklı Bir Kişi ile Hastalıklı Bir Kişinin Kanser Olup Olmadığını Ayırt Etme

ÖZET

Kanser, genetik varyasyon ve mutasyon tarafından yönlendirilen bir hastalıktır. Exome dizilemesi, yüzlerce tümörde bu varyantları ve mutasyonları keşfetmek için kullanılabilir. Burada, somatik mutasyonların saptanması için bir analiz aracı olan VarScan 2'yi ve tümör-normal çiftlerden gelen ekson verilerindeki kopya sayısı değişikliklerini (CNA'lar) sunmaktayız.

En güncel yaklaşımlardan farklı olarak, algoritmamız göğüs kanseriden gelen verileri aynı anda okur; Sezgisel ve istatistiksel bir algoritma, dizi varyantlarını tespit eder ve bunları mcc,knn naviebayes ile sınıflandırır. Bu yöntemleri, Kanser Genom Atlası'nın (TCGA) bir parçası olarak karakterize edilen 151 yüksek dereceli göğüs tümöründen alınan ekzom sekansı verilerinin analizine uyguluyoruz.Bu çalışmada, kanserin teşhisi ve tahmin edilmesi problemi Rastgele Orman, k En Yakın Komşu Algoritması, Bayes Ağları ve Yapay Sinir Ağları gibi çeşitli algoritmalar kullanılarak ele alınmıştır.

GİRİŞ (INTRODUCTION)

Göğüs kanseri ülkemizde, dünyada olduğu gibi en sık görülen kanserdir (1). Göğüs kanseri, hücre çoğalması ve genetik kararsızlığa yol açan çok sayıda moleküler değişiklikle daha invaziv (yayılan) ve rezistan (dirençli) özellik kazanabilen karmaşık bir hastalıktır. Bu heterojenite moleküler düzeyde farklı alt gruplar yaratmakta ve farklı klinik sonuçlara, sağaltım yanıtlarına neden olmaktadır. Daha iyi klinik sonuçlara ulaşmak üzere çalışmalar, hastalığın nedeni olan moleküler yapıları belirlemeye, hastalık evrelerini, hastalığa ve kişiye özgü farklılıkları ortaya koyarak, hastalığın teşhisi ve tanısını, izleyebilecek kanser belirteçlerine göre değişiklik gösterir (2).

Dinamik hücresel değişimleri eş zamanlı izlemeyi gerekli kılan bu yaklaşım, ancak insan genom analizinin tamamlanması sürecinde gelişen teknoloji ile klinik çalışmalara yansımıştır. Genom çalışmaları; hastalıkların moleküler temellerini anlamamızı sağlamıştır. Ancak, hızla değişen hücresel fonksiyonlar proteom çalışmaları ile açıklık kazanmaktadır. Çünkü aynı genom farklı proteom çıktılarına neden olabilir

(3). Kanser araştırmalarında proteomik teknolojiler, işlevsel ve düzenleyici yolakları ayrımlayan ve tanımlayan (4), doku ve biyolojik İnsan yapısındaki dokuların hepsi aynı genetik materyali içermektedir. Fakat her hücrede aynı genler aktif değildir. Hangi genlerin aktif olup hangilerinin aktif olmadığı bilgisi, hücrelerin nasıl bir fonksiyona sahip olduğu ve bazı genler normal şekilde çalışmadığında hücrenin bu olaydan nasıl etkileneceği bilgisini vermektedir. Hücrenin bu aktiflik bilgisi, gen ifadesine eşittir [3].

DNA dizi teknolojisi sayesinde yüzlerce genin ifade düzeyleri eş zamanlı olarak incelenebilmektedir.[7].Bu teknoloji sayesinde hasta ve sağlıklı hücrelerin gen ifadelerinin karşılaştırılması neticesinde hastalığa neden olan genlerin belirlenmesi mümkün olabilmektedir. DNA dizi gen ifade verilerini sınıflandırmak için literatürde knn, navibayes, mcc pca kullanıldı.

Bu çalışmada, kanser DNA dizi veri setlerinin sınıflandırmasında, teşhis doğruluğu arttırmak adına veri setlerine öznitelik seçimi uygulanması önerilmiştir. Önerilen yöntem Genetik Algoritma ile gerçekleştirilmiş olup, elde edilen sonuçlardan, öznitelik seçimi uygulanmasının bazı algoritmaların doğruluğunu önemli ölçüde arttırdığı görülmüştür.

1. METODLAR (METHODS)

1.1 Veri Seti

Tüm yöntemler göğüs kanseri teşhisi için kanser veri seti üzerinde değerlendirilir: Göğüs veri setinde 569 veri ve 212 tümor,357 normal olmak üzere 569 örnek bulunmaktadır

1.2 Sınıflandırma Algoritmaları

1.2.1 kNN (k En Yakın Komşuluk) Algoritması

2. K en yakın komşuluk algoritması sorgu vektörünün en yakın k komşuluktaki vektör ile sınıflandırılmasının bir sonucu olan denetlemeli öğrenme algoritmasıdır. Bu algoritma ile yeni bir vektörü sınıflandırabilmek için doküman vektörü ve eğitim dokümanları vektörleri kullanılır. Bir sorgu örneği verilir, bu sorgu noktasına en yakın k tane eğitim noktası bulunur. Sınıflandırma ise bu k tane nesnenin en fazla olanı ile yapılır. K en yakın komşuluk uygulaması yeni sorgu örneğinin sınıflandırmak için kullanılan bir komşuluk sınıflandırma algoritmasıdır.

K en yakın komşuluk algoritması çok kolaydır. K en yakın komşulukları bulmak için sorgu örneği ile eğitim dokümanları arasındaki en küçük uzaklıklar dikkate alınır. En yakın komşuları bulduktan sonra bu komşulardan kategorisi en çok olanın kategorisi dokümanın kategorisini tahmin etmekte kullanılır.

2.1.1 Naive Bayes Sınıflandırma Algoritması

Naive Bayes sınıflandırma algoritması, adını Matematikçi Thomas Bayes ‘den alan bir sınıflandırma algoritmasıdır. Naive Bayes sınıflandırması olasılık ilkelerine göre tanımlanmış bir dizi hesaplama ile, sisteme sunulan verilerin sınıfını yani kategorisini tespit etmeyi amaçlar.

Naive Bayes sınıflandırmasında sisteme belirli bir oranda öğretilmiş veri sunulur. Öğretim için sunulan verilerin mutlaka bir sınıfı/kategorisi bulunmalıdır. Öğretilmiş veriler üzerinde yapılan olasılık işlemleri ile, sisteme sunulan yeni test verileri, daha önce elde edilmiş olasılık değerlerine göre işletilir ve verilen test verisinin hangi kategoride olduğu tespit edilmeye çalışılır. Öğretilmiş veri sayısı ne kadar çok ise, test verisinin gerçek kategorisini tespit etmek o kadar kesin olabilmektedir.

Naive Bayes sınıflandırma yönteminin birçok kullanım alanı bulunabilir fakat, burada neyin sınıflandırıldığından çok nasıl sınıflandırıldığı önemli olmaktadır.

Bir Bayes yaklaşımı olarak, n boyutlu uzayda tanımlı olan X vektörü (x_1,.., x_2), m adet sınıf bulunan C_k (C_1,.,C_n) veri kümesinde son olasılığı maksimize eden bir sınıf etiketi C arar.

$$P(C_i \mid \mathbf{X}) \propto P(\mathbf{X}|C_i)P(C_i)$$

(BayesianTeoremi)

Bu teorem bir rassal değişken için koşullu olasılıklar ile önsel olasılıklar arasındaki ilişkiyi gösterir

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

P(A|B) : B olayı gerçekleştiği durumda A olayının meydana gelme olasılığı
P(B|A) ; A olayı gerçekleştiği durumda B olayının meydana gelme olasılığı
P(A) ve P(B) : A ve B olaylarının önsel olasılıklarıdır.

Bayes ağları, ilgili değişkenler arasındaki olasılıksal ilişkileri kodlayan bir grafik modelidir. İstatistiksel tekniklerle bağlantılı olarak kullanıldığında, grafik model veri modelleme için çeşitli avantajlara sahiptir. Birincisi, model tüm değişkenler arasındaki bağımlılıkları şifrelediği için, veri girişlerinin eksik olduğu durumları kolayca ele alır.

İki, Bayes ağı, nedensel ilişkileri öğrenmek için kullanılabilir ve bu nedenle, bir sorun alanı hakkında bilgi edinmek ve müdahalenin sonuçlarını tahmin etmek için kullanılabilir. Üçüncüsü, modelin hem nedensel hem de olasılıksal bir semantiğe sahip olmasından dolayı, (çoğunlukla nedensel biçimde ortaya çıkan) ön bilgi ile veriyi birleştirmek için ideal bir temsildir.

Genetik algoritmalar, daha iyi yaklaşımlar üretmek için bireylerin popülasyonu üzerinde çalışır. Popülasyon her nesilde, bireylerim problem alanlardaki uyum düzeylerine göre seçme ve doğal genetik özelliklerinden elde edilen operatörlere göre bir araya getirilme süreci ile oluşturulur. Bireyler gibi yavrular da mutasyona uğrayabilir. Bu süreçle, doğal adaptasyonda olduğu gibi, bireylerin çevreye uyumunun yüksek olduğu popülasyonların oluşmasını sağlar. Bu çalışmada, nüfus içindeki her birey bir öngörü modelini temsil etmektedir. Gen sayısı, veri kümesindeki toplam öznitelik sayısıdır. [19]

Karar Ağaçları (Descision Trees)

Veri madenciliğinde en çok kullanılan tekniklerden biridir. Bir data, karar ağacına göre sınıflandırılmak istendiğinde, sınıf etiketleri bilinen bir data setine ihtiyaç duyulur. Data seti üzerinde karar verme basamakları uygulanarak, fazla sayıdaki kayıtlı veriler, az sayıda gruplarına bölünür.

Her bölme işlemi yapıldığında, özellikleri bakımından birbirine benzer datalar grup haline gelir.
Karar Ağaçlan, (Descision Trees) adından da anlaşılacağı üzere kararın verilebilmesi için ağaç biçiminde bir yapı oluşturmaktadır. Karar ağaçlarında dal, yaprak gibi gerçek bir ağaçtakine benzer, bir takım özel terimler kullanılmaktadır.

Bir karar ağacında, karar düğümleri (decision nodes) ve yaprak düğümleri (leaf nodes) bulunur. Karar düğümleri veri, setinde karar vermek, sınıflandırma yapmak ya da tahmin etmek için kullanılan nitelikler olup, bunlar iki ya da daha fazla dala (branch) ayrılabilmektedir.
Yaprak düğümleri ise kararları tutmaktadır. Ağacın en tepesinde bulunan düğüm, kök düğüm (root node) olarak adlandırılır. Bir karara ulaşabilmek için ağaç kökünden yaprak düğümlere kadar belirli bit yol (path) izlenmektedir.

SVM (Support Vector Machine, Destekçi Vektör Makinesi)

Sınıflandırma konusunda kullanılan oldukça etkili ve basit yöntemlerden birisidir. Sınıflandırma için bir düzlemde bulunan iki grup arasında bir sınır çizilerek iki grubu ayırmak mümkündür. Bu sınırın çizileceği yer ise iki grubun da üyelerine en uzak olan yer olmalıdır. İşte SVM bu sınırın nasıl çizileceğini belirler.
Bu işlemin yapılması için iki gruba da yakın ve birbirine paralel iki sınır çizgisi çizilir ve bu sınır çizgileri birbirine yaklaştırılarak ortak sınır çizgisi üretilir

Logistic Regression

Lojistik regresyon, yalnızca iki değere sahip olabilen bir sonucun olasılığını öngörür (yani, ikiye bölünebilir). Tahmin, bir veya birkaç öngörücünün (sayısal ve kategorik) kullanımına dayanır. Doğrusal regresyon evet/hayır, var/yok gibi binary(ikili) sistemde ifade edilebilecek değerler için uygun değildir. Çünkü,0 ve 1 aralığının dışında değer tahmin edebilir.
Lojistik regresyon, 0 ile 1 arasındaki değerlerle sınırlı lojistik eğrisi üretir. Lojistik regresyon lineer bir regresyona benzer, ancak eğri olasılık yerine hedef değişkenin olasılıkları' nın doğal logaritması kullanılarak oluşturulur.
Logistik regresyonda odds ve odds ratio kavramları vardır. Örneğin; bir torbada 2 mavi, 3 kırmızı, 5tane sarı top olsun.Mavi gelme olasılığı 2/10 iken gelmeme olaslığı 8/10 dur. (2/10) / (8/10) olasılık oranıdır.

1.1.1 Random Forest (Rastgele Orman) Sınıflandırma Algoritması

Random Forest, öğretim ve test setleri olarak ayrılan veri seti üzerinde, Karar Ağacı Algoritmasını **N** defa kullanarak daha iyi tahminler yapmamızı sağlayan bir modeldir. **N** defa çalıştırılan algoritmalar sonucunda elde edilen tahminlerin ortalaması alınarak daha doğru bir tahmin üretir. Fakat Karar Ağacı Algoritmasını kullandığımız için üretilen tahmin, her tahmin edilmek istenen sayı için farklı olmayabilir. Yani belirli aralıktaki sayılar için bu değer aynıdır. Örneğin Lineer Regresyonda 7. ve 8. noktaya denk gelen rakam farklı iken, Random Forest Regresyonda da bu değer aynı olabilir.

Breiman tek bir karar ağacı üretmek yerine çok sayıda ve çok değişkenli ağaçların her birinin farklı eğitim kümeleriyle eğitilmesi sonucu ortaya çıkan kararların birleştirilmesini önerir. Bir sınıflandırıcı yerine birden çok sınıflandırıcı üreten ve sonrasında onların tahminlerinden alınan sonuçlar ile yeni veriyi sınıflandıran öğrenme algoritmasıdır. Büyük veri tabanlarında performansı iyidir. Dengesiz veri seti sınıfında hata dengeleme yöntemlerine sahiptir. Kaybolan verilerin büyük olasılıkla doğruluğu korunur ve kaybolan verilerin tahmin edilmesinde etkili bir metottur [17] [18].

3. DENEYSEL SONUÇLAR (EXPERIMENTAL STUDY)

3.1 Sınıflandırma Algoritmalarının Karşılaştırılmasında Kullanılan Kriterler

3.1.1 Doğruluk Oranı (Accuracy)

Model başarımının ölçülmesinde kullanılan en popüler ve basit yöntem, modele ait doğruluk değeridir. Doğru sınıflandırılmış örnek sayısının (TP +TN), toplam örnek sayısına (TP+TN+FP+FN) oranıdır.

$$Doğruluk = \frac{(TP + TN)}{(TP + FP + FN + TN)}$$

3.1.2 Duyarlılık (Recall)

Doğru sınıflandırılmış pozitif örnek (TP) sayısının, toplam pozitif örnek sayısına (TP+FN) oranıdır.

3.1.3 MCC (Matthews Correlation Coefficient)

MCC, sınıf dağılımlarının dengesiz olduğu durumlarda bile en iyi sonucu vermektedir. MCC'de çıktı değeri -1 ile +1 arasında değişmektedir. 0 değeri rastgele sınıflandırma durumunu, -1 değeri, var olan gerçek değerler ile sınıflandırıcının verdiği kararların tamamen birbirinden zıt olduğunu göstermektedir. +1 ise sınıflandırma başarısının tam doğru olduğunu göstermektedir. MCC değeri şu şekilde hesaplanır [20]:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

3.2 Deneysel Çalışma

3.2.1 Göğüs Kanseri Veri Seti

		ACCUARY	RECALL	MCC	AUROC
Random Forest		100.0	1.000	1.000	1.000
KNN		100.0	1.000	1.000	0.886
Navie Bayes		88.04	0.880	0.764	0.898
SVM		95.07	0.951	0.895	0.737
Decision Tree		91.91	0.919	0.830	0.919
Logistic		97.89	0.979	0.955	0.578
BayesNet		96.13	0.961	0.917	0.813

Tablo 1: *İşlenmemiş göğüs kanseri veri setine uygulanan algoritmalar sonucu elde edilen sonuçlar*

Accuary (Doğruluk Oranı) ölçütüne göre en iyi sonucu KNN (K nearest neighborhood) ve Random Forest algoritmaları göstermiş olup, diğer algoritmalar bu ölçüte göre sırasıyla Logistic, Bayesian Network, SVM (Support Vector Machine), Decision Tree ve Navie Bayes şeklinde sıralanabilir.

Recall (Duyarlılık) ölçütüne göre en iyi KNN (K nearest neighborhood) ve Random Forest algoritmaları göstermiş olup diğer algoritmaların bu ölçüte göre sırasıyla Logistic, Bayesian Network, SVM (Support Vector Machine), Decision Tree ve Navie Bayes şeklinde sıralanabilir.

MCC (Matthews Correlation Coefficient) ölçütüne göre en iyi sonucu KNN (K nearest neighborhood) ve random forest algoritmaları göstermiş olup diğer algoritmaların bu ölçüte göre sırasıyla Logistic, Bayesian Network, SVM (Support Vector Machine), Decision Tree ve Navie Bayes şeklinde sıralanabilir.
AU-ROC ölçütüne göre en iyi sonucu Random Forest algoritması göstermiş olup, diğer algoritmalar bu ölçüte göre sırasıyla Decision Tree, Navie Bayes, KNN (K nearest neighborhood), Bayesian Network, SVM (Support Vector Machine) ve Logistic şeklinde sıralanabilir.

Random Forest		94.71	0.968	0.783	0.987
KNN		92.21	0.895	0.598	0.934
Navie Bayes		85.43	0.853	0.732	0.824
SVM		93.04	0.955	0.531	0.647
Decision Tree		90.86	0.901	0.589	0.939
Logistic		97.89	0.979	0.948	0.587
BayesNet		92.83	0.929	0.751	0.772

Tablo 2: *PCA (Principal Component Analysis) Algoritma ile öznitelik seçimi uygulanmış göğüs kanseri veri setine uygulanan sınıflandırma algoritmaları sonucu elde edilen sonuçlar*

Accuary (Doğruluk Oranı) ölçütüne göre en iyi sonucu Logistic algoritması göstermiş olup, diğer algoritmalar bu ölçüte göre sırasıyla Random Forest, SVM, Bayesian Network, KNN (K nearest neighborhood), Decision Tree ve Navie Bayes şeklinde sıralanabilir.

Recall (Duyarlılık) ölçütüne göre en iyi Logistic ve Random Forest algoritmaları göstermiş olup diğer algoritmaların bu ölçüte göre sırasıyla SVM (Support Vector Machine), Bayesian Network, Decision Tree, KNN (K nearest neighborhood) ve Navie Bayes şeklinde sıralanabilir.

MCC (Matthews Correlation Coefficient) ölçütüne göre en iyi sonucu Logistic algoritması göstermiş olup diğer algoritmaların bu ölçüte göre sırasıyla Random Forest, Bayesian Network, Navie Bayes, KNN (K nearest neighborhood), Decision Tree ve SVM (Support Vector Machine) şeklinde sıralanabilir.
AU-ROC ölçütüne göre en iyi sonucu Random Forest algoritması göstermiş olup, diğer algoritmalar bu ölçüte göre sırasıyla Decision Tree, KNN (K nearest neighborhood), Navie Bayes, Bayesian Network, SVM (Support Vector Machine) ve Logistic şeklinde sıralanabilir.

SONUÇ (CONCLUSION)

Bu makalede, kanser teşhisi ve tahmini problemi, makine öğrenmesi teknikleri ile ele alınmıştır. İşlenmemiş veri setleri ve öznitelik seçimi yapılmış kanser genome dizi veri setleri üzerinde Rastgele Orman, Bayes Ağları, Naive Bayes, Svm,Logistic,Bayes Network ,k En Yakın Komşu ve yöntemleri uygulanarak, bu yöntemlerin karşılaştırmalı bir değerlendirmesi yapılmıştır. Elde edilen sonuçlardan göğüs kanseri veri setinde Naive Bayes algoritmasının, göğüs kanseri veri setlerinden yapılmış veri setlerinde Rastgele Orman algoritmasının doğruluk oranı yüksek sonuçlar verdiği gözlemlenmiştir.

KAYNAKLAR (REFERENCES)

[1] Jemal A, Siegel R, Ward E, et al. Cancer statistics. CA Cancer J Clin, 2007; 57 (1): 43-669.

[2.] Petterson SD, Aebersold RH. Proteomics: First decade and beyond. Nature Genetics Supplement 2003; 33: 311-32.

[3]. Baskın Y. Tıpta teknolojik gelişimin neden olduğu kavram değişimleri: kişiselleştirilmiş tıp. Türk Hijyen ve Deneysel Tıp Dergisi, 2007; 64(2): 54-59.

[4]. Baskın Y. Yigitbası T. Clinical Proteomics of Breast Cancer. Current Genomics, 2010; 11 (7): 528-536.

J.A. Cruz, D.S. Wishart, "Applications of Machine Learning in Cancer Prediction and Prognosis", *Cancer Informat,* 2006.

G. S. Özcan, "Bütünleştirici Modül Ağlarıyla Gen Düzenleme Analizi" *Başkent Üniversitesi,*

2014.

H. Ü. Lüleyap, “Moleküler Genetiğin Esasları” *İzmir: Nobel Kitabevi*, 2008.

[1] H. S. BAL, F. Budak, “Mikroarray Teknolojisi,” *Uludağ Üniversitesi Tıp Fakültesi Dergisi,* 2012.

[2] Ö. Şimşek, “Mikroarray Teknolojisi ve Diş Hekimliğinde Kullanımı,” *Atatürk Üniversitesi Diş Hekim Fakültesi Derg*isi, 2013.

[3] K. Shakya, H. J. Ruskin, G. Kerr, M. Crane, J. Becker, “Comparison of Microarray Preprocessing Methods,” *Springer New York,* 2010.

[4] K. Ipekdal, “Microarray Technology,” 2011.

[5] H. Liu, I. Bebu, X. Li, “Microarray probes and probe sets.,” *Front Biosci (Elite Ed),* 2010.

[6] R. Díaz-Uriarte, S. Alvarez de Andrés, "Gene selection and classification of microarray data using random forest.," *BMC Bioinformatics*, 2006.

[7] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics*, 2000.

[8] H. Liu, J. Li, and L. Wong, "Classification and Study on Feature Gene Patterns Selection Expression and Profiles Methods Using Proteomic" *Genome Informatics*, 2002.

[9] A. Statnikov, C. F. Aliferis, I. Tsamardinos, D.Hardin, and S. Levy, "A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis" *Bioinformatics*, 2005.

[10] J. Khan, J. S. Wei, M. Ringnér, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, and P. S. Meltzer, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks.", 2001.

[11] M. Ringnér and C. Peterson, "Microarray-based cancer diagnosis with artificial neural networks," *Biotechniques*, 2003.

[12] Enrico Glaab, Jaume Bacardit, Jonathan M. Garibaldi, Natalio Krasnogor, Using Rule- Based Machine Learning for Candidate Disease Gene Prioritization and Sample Classification of Cancer Gene Expression Data, 2012.

[13] Leo Breiman, Adele Cutler, Random Forests, 2005.

[14] Leo Breiman, Machine Learning, 2001.

[15] Fernando Gómez, Alberto Quesada, Genetic algorithms for feature selection in Data Analytics

[16] Liu Y., Cheng J., Yan,C., Wu X., Chen F., Research on the Matthews Correlation Coefficients Metrics of Personalized Recommendation Algorithm Evaluation, *International Journal of Hybrid Information Technology*, 2015.