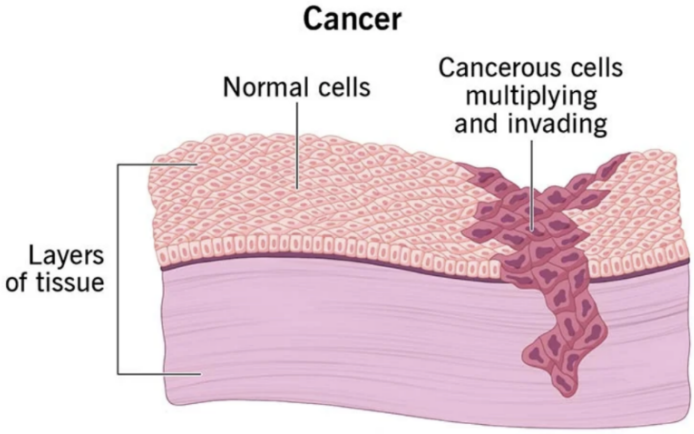
Veri madenciliği ve genetik algoritmalarla kanser geni araştırması

Kanser insan hayatının %25'inin ölümünden sorumludur. Kanserin erken ve doğru teşhisi hastanın hayatı için kritik öneme sahiptir. Gen analizi sayesinde kansere karşı ilaçlar ve tedaviler geliştirilir aynı zamanda kanserin tanımlanması ve sınıflandırmasında büyük yararı vardır. Bu araştırmada yumurtalık , prostat ve akciğer kanseri için gen izlenimlerinin analiz eder. Genetiği entegre edilmiş veri analizi için bir gen algoritması önerilmiştir. Bu entegre algoritma, veri ön işleme ve veri madenciliği için korelasyona dayalı tahminler elde edilir. Önerilen algoritma tarafından elde edilen bilgi, en anlamlı genleri belirlemesi sayesinde güçlü sınıflandırma sonuçları verir. Sınıflandırma doğruluğunu daha da arttırmak için Bagging ve stacking algoritmaları uygulandı. Sonuçlar rapor edilen verilerle karşılaştırıldı. Genetik verinin, haritanın tamamlanması kanser tespiti ve sınıflandırmasının maliyetini ve karmaşıklığını eninde sonunda azaltacaktır.

Kanser esas olarak; epitel hücrelerinde (karsinomlar), bağ/kas dokusunda (sarkomlar) ve beyaz kan hücrelerinde (lösemiler ve lenfomalar) gelişir. Normal bir hücrede DNA'ya zarar veren ve hücre replikasyon mekanizmasını bozan bir mutasyon, malign tümörlere (kanserler) sebep olur. Mutasyonları hızlandırabilecek tütün dumanı, radyasyon, kirli su ve hava gibi bir dizi kanserojen vardır. Bu nedenle, kanseröz bir duruma katkıda bulunan mutasyona uğramış genlerin tanımlanmasına ihtiyaç vardır.



Kanser tanımlama yöntemlerinden biri de genetik verilerin analizidir. İnsan genomu yaklaşık 10 milyon tek nükleotid polimorfizmi (SNP) içerir. Mikroarray teknolojisi, bir bireyin gen ekspresyonunu ve SNP'lerini elde etmek için kullanılır. Yüksek maliyet nedeniyle, genetik veriler

(hasta başına 15.000 kadar gen içerir) normalde sınırlı sayıda hasta (100-300 hasta) üzerinde toplanır. Bu kadar geniş veri setlerinden en bilgilendirici genlerin seçilmesi gerekmektedir. Bilgi vermeyen genlerin çıkarılması, karmaşıklığı azaltır ve en önemli genlerin tanımlanması, hastalıkların sınıflandırılması ve kanser türü gibi çeşitli sonuçların tahmin edilme şansını artırır. Edinilen bilgi, daha erken bir aşamada önleyici tedbirlerin alınmasına izin verecektir. Gen ekspresyonu sınıflandırma problemleri için çeşitli hesaplamalı zeka teknikleri uygulanmıştır. Bunlar; Fisher lineer diskriminant analizi, KNN, karar ağacı, çok katmanlı algılayıcı, destek vektör makineleri, artırma, ve kendi kendini organize eden haritalar, hiyerarşik kümeleme ve çizge kuramsal yaklaşımlardır.

Bu makale üç farklı kansere odaklanır: yumurtalık(Jinekolojik kansere bağlı ölümlerin en yaygın nedenidir. Kadınların yumurtalıklarında görülmektedir), prostat(Erkeklerin üreme sisteminin bir parçası olan prostat bezinde görülür .Erkeklerde en sık görülen ikinci kanserdir. Yaşla ilişkili bir şekilde artış gösterir.) ve akciğer kanser(Nefes borusunda, ana hava yolunda veya akciğer dokusunda meydana gelen kanser türüdür.). Her kanser için bir eğitim ve test verileri kümesi genlerin kalitesini analiz etmek için kullanılmıştır. Yukarıdaki veritabanlarına entegre bir gen arama algoritması uygulanmıştır. Alınan sonuçlar, önerilen entegre algoritmanın sağlamlığını incelemek için önceki bilgilerle karşılaştırıldı. Önerilen algoritmanın ana özellikleri de araştırıldı.

–Bu makale üç farklı kansere odaklanmaktadır: yumurtalık, prostat ve akciğer kanseri. Genlerin kalitesini analiz etmek için her kanser için bir eğitim ve test veri seti kullanılmıştır. Yukarıdaki veri setlerine entegre bir gen arama algoritması uygulanmıştır. Elde edilen sonuçlar, önerilen entegre algoritmanın sağlamlığını incelemek için önceki literatürle karşılaştırıldı. Önerilen algoritmanın temel özellikleri de tartışılmıştır.

Yumurtalık kanseri, yalnızca %29'luk uzun süreli sağkalım oranıyla özellikle öldürücüdür. Kanserin varlığını tespit etmek için kullanılan mevcut biyobelirteç, tümör hacmi ile ilişkilidir. Böylece kanser, çok sayıda hasta için iyileşme oranının yüksek olduğu erken evresinde tespit edilememektedir. Petricoin ve diğerleri, her biri 15.154 gen içeren 100 eşit dağıtılmış eğitim örneğini (yani 50 kanser ve 50 normal) analiz etmek için genetik algoritma ve kümeleme teknikleri uyguladı. Genetik algoritma için kodlama şeması, mantıksal kromozomlar iken uygunluk işlevi, mantıksal bir kromozomun bir lider küme haritası belirleme (yani, homojen kümeler oluşturma) yeteneğiydi. Analizleri, 116 ayrı test örneğine uygulandığında %97,4 tahmin doğruluğu ile sonuçlandı. Beş önemli gen (M/Z değerleri) yumurtalık kanseri göstergeleri olarak tanımlandı.

Table 1
Summary of the published cancer data sets and result analysis

Cancers	Decision		Training set		No. of genes	Testing set		Preprocessing method	Prediction tools	Training accuracy (%)	Testing accuracy (%)
	A	B	A	B		A	B				
Ovarian (old)	Cancer	Normal	50	50	15,154	50	66	Genetic algorithms +self-organizing maps	Iterative search algorithms	–	97.40
Ovarian (new)	Cancer	Normal	Random		15,154	Random				–	100.00
Prostate	Tumor	Normal	52	50	12,600	27	8	Signal-to-noise metric	*k*-nearest neighbor	90.00	86.00
Lung	MPM	ADCA	16	16	12,000	15	134	Correlation expression levels	15 diagnostics ratios	–	97.00

Prostat tümörleri tarihsel ve klinik olarak daha heterojen kanserler arasındadır. Prostat spesifik antijen (PSA) testi, prostat kanserinin erken teşhisinde faydalıdır. Singh, sınıf tahmini, gen ifadesi ölçümleri, gen sıralaması, permütasyon testi ve korelasyonu kullanarak yeni tespit yöntemlerini araştırdı. Analiz hem genotip hem de fenotip özelliklerini içeriyordu. Gleason skoru dışında genotip ve fenotip arasındaki korelasyon güçlü değildi. Deterministik modeller oluşturmak için her biri 12.600 gen içeren yüz iki örnek (50 normal ve 52 tümörlü) kullanıldı. Önemli genleri seçmek için bir sinyal-gürültü metriği kullanıldı. Tümörlü örneklerde, 317 gen daha yüksek ekspresyon seviyelerine sahipken, normal prostat örneklerinde 139 gen daha yüksek düzeyde eksprese edildi. Bir KNN kümeleme algoritması, yüksek doğrulukta tahminler sağladı. En önemli 29 genden oluşan son bir set tanımlandı. Test numuneleri (27 tümörlü ve 8 normal), eğitim veri seti ile karşılaştırıldığında genel mikrodizi yoğunluğunda yaklaşık 10 kat fark olan farklı kaynaklardan elde edildi.

Mezotelyoma (MPM) ve adenokarsinom (ADCA) gibi akciğer kanserlerini ayırt etmek zordur ve bu nedenle tanıları zordur. MPM, akciğeri kaplayan plevranın kanseridir. Benign veya malign olabilir. Sarkomatöz, epiyelyal ve karışık olmak üzere üç temel tipi vardır. ADCA, sıklıkla akciğer periferinde meydana gelen küçük hücreli dışı akciğer kanseri grubunun bir parçasıdır. Bu kanserlerin tedavisi büyük ölçüde erken ve kesin teşhislerine bağlıdır. Gordon, MPM ve ADCA arasında ayrım yapmak için genlerin ekspresyon seviyelerini kullanmışlardır. 15 teşhis oranı oluşturmak için eşit olarak dağıtılmış (yani 16 MPM ve 16 ADCA) 32 numunelik bir eğitim seti kullanıldı. ADCA ve MPM arasında ekspresyon düzeylerinde (yani ters korelasyon) yüksek bir farka sahip olanları belirlemek için tüm genler arandı. İstatistiksel olarak en önemli farklara ve 600'den büyük bir ortalama ifade düzeyine sahip sekiz gen (beş MPM ve üç ADCA geni) seçildi. Numune başına on beş ifade oranı, MPM'de nispeten daha yüksek seviyelerde ifade edilen beş genin her birinin ifade değerinin, ADCA'da nispeten daha yüksek bir seviyede ifade edilen her bir genin ifade değerine bölünmesiyle hesaplandı. Bu oranlar 149 numunede (15 MPM ve 134 ADCA) test edildi. Her oran %90'ın üzerinde doğruydu, iki veya üç oranın birleştirilmesi ise %95'in üzerinde bir tahmin doğruluğu sağladı.

Veri madenciliği algoritmaları, azaltılmış boyutluluk ile gen ekspresyonu veri setlerini analiz etmek için kullanıldı. Verilerdeki gizli kalıpları keşfetmek değerli olabilir ve bilgi verici genler, kontrol ayarı, tedavi seçimi vb. gibi keşiflere yol açabilir. Gelişmekte olan bir bilim olarak, veri madenciliği bir teoriler ve algoritmalar koleksiyonudur, örneğin istatistik, kaba küme teorisi, karar ağacı algoritmaları, destek vektör makineleri (SVM), sinir ağları vb.

Bu makalede, anlamlı bilgi, yüksek sınıflandırma doğruluğu üreten ve kişiselleştirmeye izin veren veri madenciliği algoritmalarına odaklandık. Analiz için DT ve SVM algoritmaları kullanıldı. Bilgi tabanını zenginleştirmek ve sınıflandırma doğruluğunu artırmak için “Bagging” ve “Stacking” kullanıldı.

DT algoritması, yorumlanabilirliği en üst düzeye çıkarmak için karar ağaçlarına veya if-then ifadelerine dayalı kurallar oluşturur. DT algoritması, kategorileri özetleyen ve tanımlayan, bunları sınıflandırıcılar halinde birleştiren ve tahminlerde bulunmak için kullanan kalıpları keşfeder. Bu araştırmada PART karar listesi kullanılmıştır. Her yinelemede kısmi bir C4.5 DT oluşturmak için ayır ve fethet stratejisini kullanırlar ve ardından kural olarak "en iyi" yaprağı kodlarlar. SVM'ler, bir dizi etiketli eğitim verisinden işlev oluşturma yaklaşımıdır. Fonksiyon bir sınıflandırma fonksiyonu olabilir veya fonksiyon genel bir regresyon fonksiyonu olabilir. DVM'ler, ikili sınıflandırma problemini doğru bir şekilde sınıflandırmak için girdi uzayı içinde optimal bir hiper düzlem bulmaya çalışır. Hiperdüzlem, hiperdüzlem ile ikili (pozitif ve negatif) örnekler arasında maksimum mesafe olacak şekilde seçilir. DVM'ler, sıfır olmayan Lagrange çarpanlarını belirlemek ve optimal hiperdüzlemi oluşturmak için ikili ikinci dereceden programlama problemini çözer. SVM'ler çeşitli şekillerde eğitilebilir. Özellikle basit ve hızlı bir yöntem sıralı minimum optimizasyon dur. SMO, büyük ikinci dereceden programlama problemlerini analitik olarak çözülen olası en küçük QP problemlerine böler. SMO, doğrusal SVM'ler ve seyrek veri kümeleri için en hızlıdır.

“Bagging”, çoklu oylama şeması yoluyla sonucu doğru bir şekilde tahmin etmek için verilerden çoklu modeller üretme yöntemidir. Çoklu modeller, önyükleme süreciyle oluşturulur (yani, öğrenme setini çoğaltır). Bu, modelleri eğitmek için birden çok yeni öğrenme seti sağlar. DT gibi temel sınıflandırıcı, test setlerinin yanı sıra çoklu öğrenme setlerini eğitmek, değerlendirmek ve tahmin etmek için kullanılır. Öğrenme setini bozmak, oluşturulan varsayımlarda önemli değişikliklere neden olabilir ve böylece doğruluğu artırır. Birkaç sınıflandırıcı, bir veri vektörü satırı için kendi tahminlerini üretir. Gerçek kararla birlikte her bir sınıflandırıcı tarafından yapılan tahminler, “Stacking” genelleme yöntemi için bir girdi veri seti olarak kullanılır. Nihai tahmini üretmek için bu türetilmiş veri setini değerlendirmek için DT gibi belirli bir temel sınıflandırıcı kullanılır. “Stacking”, genelleme hata oranını en aza indirmek için bir şemadır.

Sonuç olarak:
Entegre gen arama algoritması (veri madenciliği ile GA-CFS algoritması); yumurtalık, prostat ve akciğer kanserlerinin eğitim ve test genetik ekspresyon veri setlerine başarıyla uygulandı. Bu tekdüze uygulanabilir algoritma, yalnızca yüksek sınıflandırma doğruluğu sağlamakla kalmadı, aynı zamanda üç kanserin her biri için en önemli genlerin bir dizisini de belirledi. Bu gen setleri, tıbbi önemleri için daha fazla araştırma gerektirmektedir.

Genetik Algoritma:

Genetik Algoritmalar, evrimden ilham alan bir hesaplama modelleri ailesidir. Bu algoritmalar, basit bir kromozom benzeri veri yapısı üzerinde belirli bir soruna olası bir çözümü kodlar ve uygular. Kritik bilgileri korumak için bu yapılara rekombinasyon operatörleri. genetik algoritmalar Genetik algoritmaların sahip olduğu problemlerin aralığına rağmen, genellikle fonksiyon optimize edici olarak görülürler. Kanser için genetik algoritmanın görselleştirilmiş hali:

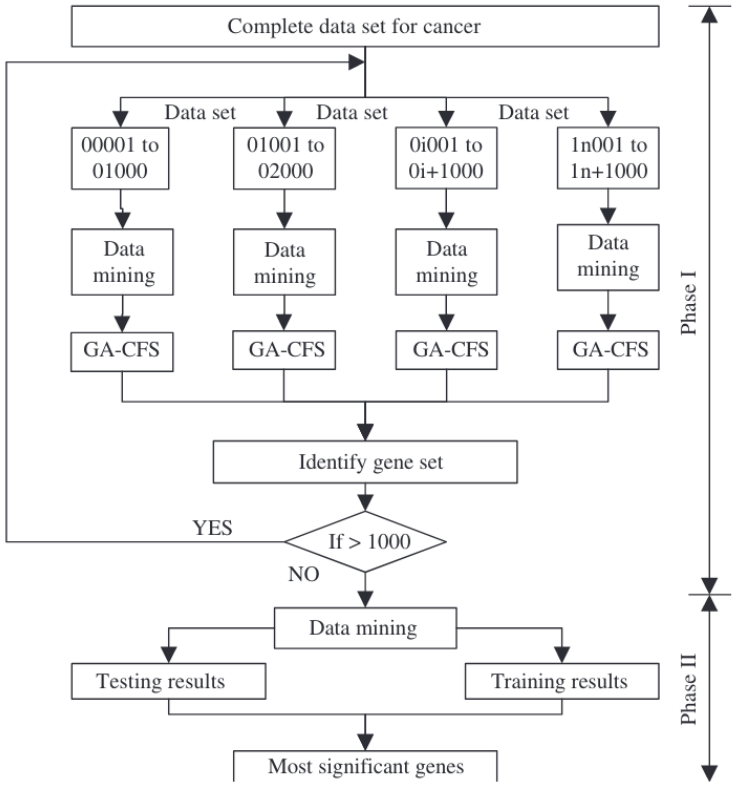


Fig. 1. Integrated gene-search algorithm.

Hazırlayanlar:

Ekin BEYKIN 121520052
Umay YENTÜR 121520041

Kaynakça:

Shah, Shital, ve Andrew Kusiak. “Cancer Gene Search with Data-Mining and Genetic Algorithms”. *Computers in Biology and Medicine*, c. 37, sy 2, Şubat 2007, ss. 251-61. *ScienceDirect*, https://doi.org/10.1016/j.compbiomed.2006.01.007.

Rawla, Prashanth. “Epidemiology of Prostate Cancer”. *World Journal of Oncology*, c. 10, sy 2, Nisan 2019, ss. 63-89. *PubMed Central*, https://doi.org/10.14740/wjon1191.

Jayson, Gordon C., vd. “Ovarian Cancer”. *The Lancet*, c. 384, sy 9951, Ekim 2014, ss. 1376-88. *ScienceDirect*, https://doi.org/10.1016/S0140-6736(13)62146-7.

Lung Cancer. https://www.cancerresearchuk.org/about-cancer/lung-cancer.

Mathew, Tom V. "Genetic algorithm." *Report submitted at IIT Bombay* (2012): 53.http://datajobstest.com/data-science-repo/Genetic-Algorithm-Guide-[Tom-Mathew].pdf