DNA Dizilimlerinin Sınıflandırılmasında Karar Ağacı Algoritmalarının Karşılaştırılması

Bihter Daş[1], İbrahim Türkoğlu[2]

[1]Fırat Üniversitesi, Teknoloji Fakültesi, Yazılım Mühendisliği
23119, Elazığ/Turkiye
bihtedas@gmail.com

[2]Bingöl Üniversitesi, Mühendislik ve Mimarlık Fakültesi, Bilgisayar Mühendisliği Bölümü
12000, Bingöl/Türkiye
iturkoglu@bingol.edu.tr

Özet

DNA dört tip nükleotidden oluşan bir zincir moleküldür. Son yıllarda DNA üzerinde yapılan çalışmalarda, DNA 'daki nükleotid dizilişlerinin birbiri ardı sıra tekrar ettiği görülmüştür. STR olarak adlandırılan ve çeşitli alanlarda kullanılan bu tekrarlar genetik hastalıklarda, adli amaçlı kimlik tespitinde, babalık araştırmaları ve tümör biyokimyasal araştırılmaları gibi birçok farklı amaçlar için kullanılmaktadır.

Bu makale çalışmasında, 4 bakteri türünün farklı uzunluklardaki DNA dizilimleri alınarak, bu dizilimlerde tekrar eden nükleotid çiftlerin frekansı bulunmuş ve elde edilen bu frekans değerlerine Karar Ağacı algoritmalarından J48, LMT ve RandomForest uygulanarak bir sınıflandırılma yapılmıştır. Sınıflandırma sonucunda RandomForest algoritmasının, J48 ve LMT algoritmalarından sınıflandırma başarımının çok daha yüksek olduğu görülmüştür.

Anahtar Kelimeler: *DNA dizilimi, özellik çıkarımı, sınıflandırma, STR, karar ağaçları.*

Abstract

DNA is a chain molecule consists of four kind of nucleotides. In recent year's studies on DNA, it is showed that sequences of nucleotide in DNA are repeated one after another. These repeats called STR are used in various fields such as genetic diseases, forensic identification, fatherhood research, biochemical tumor research.

In this paper, it was found the frequency of repetitive nucleotide pairs taking DNA sequences of different lengths of the four bacterial species and a classification was made to obtained frequency values by using J48, LMT and Random Forest algorithm from Decision Trees. Classification results show performance of RandomForest algorithm is much higher than performance of J48 and LMT methods.

Keywords: *DNA sequence, feature extraction, classification, STR, decision trees.*

1. Giriş

Biyoinformatik, özellikle moleküler biyoloji ile bilgisayar teknolojisini ve bununla ilişkili veri işleme aygıtlarını bünyesinde barındıran bilimsel bir disiplindir. Bir diğer tanımla, karmaşık biyolojik verilerin derlenmesi ve analiz edilmesinde bilişim teknolojilerinin kullanılması esasına dayanan ve biyolojik olayların moleküler düzeyde açıklanmasına yardımcı olan bir bilimdir.

DNA dizilim hizalama biyoinformatiğin en temel problemlerinden biridir. DNA diziliminin temel yapıtaşı olan nükleotidin yapısında fosfat, seker ve organik baz bulunur. Organik bazlar adenin (A), timin (T), sitozin (C) ve guanin (G)'dir. Nükleotidler hangi organik bazı içeriyorlarsa o bazın ismiyle adlandırılırlar.

Son yıllarda DNA üzerinde yapılan araştırmalarda; DNA'daki bazların ve baz dizilimlerin birbiri ardı sıra tekrar ettiği belirlenmiştir. Tekrarlayan bu nükleotid dizileri; kromozomal sentromeri çevreledikleri için "satellit (uydu)" ismini almışlardır. Mikrosatellitler, DNA lokusları; 2-6 nükleotit uzunlukta kısa, tekrarlanan DNA dizilerini ifade etmektedir [1,2]. Mikrosatellitler; basit dizi tekrarları (Simple Sequence Repeats, SSR) ya da kısa ard arda tekrarlar (Short Tandem Repeats, STR) olarak da adlandırılırlar.[1,3]. Bu tekrarlar genetik kökenli haritalamada, tümör biyokimyasal araştırmalarda, adli bireysel kimlik tespitinde, babalık ve nüfus genetik analizlerde yaygın bir şekilde kullanılmaktadır [4] Çeşitli veri madenciliği tekniklerinin kullanılması DNA dizilimlerinin sınıflandırılması açısından büyük öneme sahiptir.

Bu makalede DNA dizilimlerini sınıflandırmada Karar Ağacı algoritmalarından J48, LMT(Lojistik Model Ağacı) ve RandomForest (Rasgele Ağaç) kullanılmıştır.

2. Verilerin elde edilmesi

Bu çalışmada deneysel veriler için National Center for Biotechnology(NCBI) sitesi Gen bankasından [5] gerçek

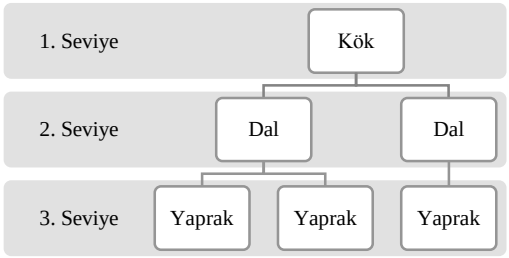
veriler alınmıştır. Alınan Escherichia coli, Bacillus cereus, Buchnera aphidicola ve Enterobacter cloacae gibi 4 bakteri türü sınıflandırılma için kullanılmıştır. Tablo 1'de kullanılan bakteri türleri, erişim numaraları, sayıları ve ortalama uzunlukları verilmiştir

Tablo 1: Kullanılan veri türlerinin açıklaması

Sıra	Bakteri Türü	Erişim Numarası	Bakteri Örnek Sayısı	Dizilimdeki Nükleotid Sayısı
1	Escherichia coli	AE14075	31	102
2	Bacillus cereus	NC_016771	40	94
3	Buchnera aphidicola	AF012886	35	96
4	Enterobacter cloacae	EU606203	48	104

3. Uygulamada Kullanılan Karar Ağacı Algoritmaları

Veri Madenciliği ile sınıflandırmada en çok kullanılanlardan yöntemlerden birisi de karar ağaçlarıdır. Karar ağacının yapısında her bir düğüm bir niteliği temsil eder. Dallar ve yapraklar ağaç yapısının elemanlarıdır. En üstteki eleman kök, en alttaki eleman yaprak ve bunların arasında kalan elemanlar ise dal olarak isimlendirilir [10]. Karar ağacında kurallar kökten yaprağa doğru inilerek (IF-THEN rules) yazılır. Karar ağacı yönteminde bir olayın sonuçlandırılmasında sorunun cevabına göre hareket edilir. Şekil 1' de bir karar ağacının yapısı gösterilmektedir.



Şekil 1: Bir Karar Ağacının yapısı

a. J48 Algoritması

Enformasyona dayalı, verilerden ilgili özellikleri seçmek için otomatik işlem yeteneğine sahip, Naive Bayes, ID3, Lojistik Regresyon gibi algoritmalarına göre sınıflandırması en yüksek algoritmadır. Enformasyon kazancının en iyi olduğu noktadan örnekleri bölen yinelemeli algoritmadır. IF-THEN kurallarına dayalı bir karar ağacı ve üyelik fonksiyon kümeleri çıktısı verir. Ağaç yapısı, denekleri bölme ve ağacın en iyi kök değişkeninin seçilmesi süreci ile başlayıp yukarıdan aşağıya doğru inşası gerçekleştirilmektedir. J48, anlamlı olmayan diğer bir deyişle zayıf dalları kesmek için etkin bir budama işlemi yapabilmektedir. Bunun nedenlerinden biri, karar ağaçlarının amacının veri keşfetmek değil, veriler üzerinde basit bir sınıflandırma modeli oluşturmak olmasındandır [11].

b. LMT Algoritması

Lojistik Model Ağacı Algoritması, Lojistik Regresyon ve Karar Ağacı Öğrenmesi algoritmalarını birleştiren denetimli bir sınıflandırma modelidir. Bu ağaçlar yapraklarında lojistik azalma fonksiyonları olan sınıflandırma ağaçlarıdır. [12]. LMT, LogitBoost tekrar sayısını bulmak için çapraz doğrulama kullanır. Algoritma ikili ve çoklu sınıf hedef değişkenleriyle ilgilenebilir.

c. RandomForest Algoritması

2001 yılında Breiman tarafından geliştirilen bu algoritmada sınıflandırıcı için amaç, tek bir karar ağacı üretmek yerine her biri farklı eğitim kümelerinde eğitilmiş olan çok sayıda ağacın kararlarını birleştirmektir. Farklı eğitim kümeleri için aynı dağılımlı rasgele özellik seçimi kullanılır. Karar ağaçlarını oluştururken, her seviyedeki özniteliği belirlerken önce bütün ağaçlarda birtakım hesaplamalar yapılarak nitelik belirlenir, daha sonra diğer ağaçlardaki nitelikler birleştirilerek en fazla kullanılan öznitelik seçilir. Seçilen öznitelik ağaca dahil edilerek diğer seviyelerde aynı işlemler tekrarlanır. Algoritmayı başlatmak için her bir düğümde kullanılan değişkenlerin sayısı ve geliştirilecek ağaçların sayısı kullanıcı tarafından belirlenmelidir. RandomForest (RO), ağaç üretmek için CART (Classification and Regression Tree) algoritmasını kullanır. Düğüm ve dallar bu algoritmanın özelliklerine uygun olarak oluşturulur.

4. Uygulama Sonuçları

Bu çalışmada yer alan uygulama, Gen bankasından alınan Escherichia coli, Bacillus cereus, Buchnera aphidicola ve Enterobacter cloacae gibi 4 bakteri türden oluşan toplam 154 örnek üzerinde gerçekleştirilmiştir. Makalede kullanılan J48, LMT ve RandomForest karar ağacı algoritmaları için 154 satır 17 sütun içeren bir giriş matrisi oluşturulmuştur. Bu matrisin ilk 16 sütununu DNA dizilimde bulunan nükleotid çitlerinin frekans dağılım değerleri oluşturmaktadır. Bu türlerin DNA dizilimlerinin karakteristik matrisini oluşturmak için DNA dizilimlerinden özellik çıkarımı yapılmıştır. Bu özellik çıkarımı için Tablo 1'de verilen her bir türün ortalama uzunluğuna göre dizilimde yer alan nükleotid çiftlerin frekans dağılımı bulunmuştur. Bir DNA diziliminde yer alabilecek nükleotid çiftleri AG, AC, AT, AA, GA, GC, GT, GG, CA, CG, CT, CC, TA, TG, TC, TT şeklindedir. Uygulamada her bir örnek için bu nükleotid çiftlerinin dağılım frekansı denklem (1) 'e göre hesaplanmıştır [4].

$$f_{S_iS_j}=\frac{S_iS_j}{|S|-1} \quad (1)$$

Denkleme göre S_iS_j nükleotid çiftlerinin DNA diziliminde bulunma sayısı, $|S|$ ise DNA diziliminin uzunluğudur.

Örneğin 'GATCAACATATTGTGACGCACGT' şeklinde verilen DNA diziliminde nükleotid çitlerinin frekanslarının denklem (1)'e göre hesaplanması tablo 2'de verilmiştir.

Nükleotid çiftlerinin frekans dağılım değerleri ile oluşturulan giriş matrisine Karar Ağacı algoritmalarından J48, LMT ve RandomForest algoritmaları uygulanmış ve uygulama sonuçları tablo 3'de gösterilmiştir. Sonuçlara bakıldığında RandomForest algoritmasının %100 başarılı bir sınıflandırma yaptığı görülmektedir. Buna rağmen LMT algoritmasının performansı ise diğer algoritmalara göre daha düşüktür.

Tablo 2: Örnek dizilimin matris sütunları

Nükleotid Çifti	Frekans Değeri	Nükleotid Çifti	Frekans Değeri
AA	0.0526	TG	0.0434
AT	0.157	CA	0.1304
TA	0.0434	AC	0.0526
TT	0.434	TC	0.0434
GG	0	CT	0
GA	0.0869	CC	0
AG	0	GC	0.0434
GT	0.0869	CG	0.0869

Tablo 3: Karar Ağaçları Algoritmalarının Sınıflandırma Sonuçlarının Karşılaştırılması

Sınıflandırma Durumları	**J48 Algoritması**	**LMT Algoritması**	**RandomForest Algoritması**
Doğru Sınıflandırma	% 93	%86	100
Yanlış Sınıflandırma	%7	%14	0
Kappa İstatistiği	0.8406	0.6835	1
Mutlak Ortalama Hata	0.1211	0.2459	0.074
Kök Karesel Hata	0.2461	0.3118	0.1265
Bağıl Mutlak Hata	%26.93	%54.65	%16.45
Kök Bağıl Karesel Hata	%51.95	%72.14	%26.70

5. Sonuç

STR lokusları belli bir nükleotid dizilimine sahip ardarda tekrarlanan ünitelerden oluşmaktadırlar. Farklı alanlarda değişik amaçlarla kullanılan STR lokuslarından genetik hastalıklarda, adli amaçlı kimlik tespitinde, babalık araştırmaları ve tümör biyokimyasal araştırılmalarında yararlanılmaktadır.

Bu makale çalışmasında, bakteri türlerinin farklı uzunluklardaki DNA dizilimleri alınarak, bu dizilimlerde tekrar eden nükleotid çiftlerin frekansı bulunmuş ve elde edilen bu frekans değerlerine Karar Ağacı algoritmalarından J48, LMT ve RandomForest algoritmaları uygulanarak bir sınıflandırma yapılmıştır. Sınıflandırma sonucunda RandomForest algoritmasının J48 ve LMT algoritmalarından başarımının daha yüksek olduğu ve hatta %100 doğruluk oranıyla sınıflandırma yaptığı görülmüştür.

Kaynaklar

[1] Dönbak L. *"*Kısa ardarda tekrar eden dna dizilerinin adli amaçlı dna çalışmalarındaki yeri **"** *Türkiye Klinikleri Tıp Bilimleri* 22(2):233-8,2002.

[2] Allendorf F. W., Luikart G. H. ve Aitken S. N.,Conservation and the genetics of populations.(2nd Edition) , WileyBlackwell, USA, 2013.

[3] Butler J.M., Forensic dna typing: biology, technology, and genetics of str markers. Elsevier academic press, New york, 2005.

[4] Zhou Q., Jiang Q. Ve Wei D. ,"A new method for classification in dna sequence", the 6th international conference on computer science&education (ıccse 2011) august 3-5, 2011.

[5] http://www.ncbi.nlm.nih.gov/Genbank/genomes/bacteria.

[6] Ayhan S., Erdoğmuş Ş., "Destek vektör makineleriyle sınıflandırma problemlerinin çözümü için çekirdek fonksiyonu seçimi ",iibf dergisi, 9(1), 175- 198, 2014.

[7] Karagülle F., "Destek vektör makinelerin kullanarak yüz bulma" ,yüksek lisans tezi, 2008.

[8] A.şengür, i.türkoğlu ve m.c.ince, wavelet packet neural networks for texture classification, expert systems with applications, 32(2), mart 2007.

[9] Sengur, A., "multiclass least-squares support vector machines for analog modulation classification", expert systems with applications, 36(3), 6681-6685 (2009).

[10] Alpaydın, E, ―Zeki Veri Madenciliği: Ham Veriden Altın Bilgiye Ulaşma Yöntemleri, Bilişim 2000 Veri Madenciliği Eğitim Semineri Notları, 2000.

[11] Altıkardeş Z. A. Erdal, H., Baba, F., Fak, A.S., ―ABPM Ölçümü Olmaksızın Karar Ağaçları Algoritması ile Non-Dipper / Dipper Öngörüsü, IX. Ulusal Tıp Bilişimi Kongresi. Antalya, Türkiye, 2012.

[12] Ardıl E., "Esnek Hesaplama Yaklaşımı İle Yazılım Hata Kestirimi", yüksek lisans tezi, 2009.