Genetik Araştırmalarda Machine Learning ve Veri Madenciliği

Erdal COŞGUN[a], Ergun KARAAĞAOĞLU[a]

[a] Hacettepe Üniversitesi, Tıp Fakültesi, Biyoistatistik Anabilim Dalı, 06100, Ankara, Türkiye

Özet: *Genetik araştırmalardaki yeni keşifler bilgi teknolojilerindeki ilerleme ile paralel gerçekleşmiştir. Özellikle "Human Genome Project" bu iki bilim dalı arasındaki ilişkinin en net biçimde ortaya konulduğu ortak projelerden biridir. Genetik araştırmalardan elde edilen bilgiler tıp bilimi dışında birçok bilim dalında kullanılmaktadır. Ancak sadece bu verilerin elde edilmesi, depolanması veya tanımlayıcı analizlerin yapılması yeterli olmamaktadır. Çünkü bu veriler konunun uzmanlarının bile farkında olmadığı gizli yapıları içermektedir. Bu yapıları ortaya çıkarmada kullanılan iki önemli yaklaşım vardır: Machine Learning ve Veri Madenciliği.*

Söz konusu yöntemler istatistik bilimi ile bilgi teknolojilerinin ortak noktalarıdır. Çalışmamızda konu ile yakından ilgili olmayan kişilerin hep karıştırdığı bu iki kavramın, genetik araştırmalardaki önemi ve etkileşimi üzerinde durulacaktır. Her iki yöntem de "yapay zeka" terimine ait alt birimlerdir. Machine Learning, bilgisayarların "öğrenmesi" kolay algoritma ve teknikleri dizayn etmeye , geliştirmeye odaklanır. Belirli bir veri seti üzerinde çalışmayı hedeflemez. Geliştirdiği algoritmalar her sorunu çözmeyi hedefler. Veri madenciliği ise Machine Learning yöntemlerini kullanarak "GERÇEK" veriler üzerinde tanımlama, sınıflama, tahmin ya da kümeleme amaçlı çalışır. Bu doğrultuda danışmanlı ve danışmansız öğrenme algoritmaları kullanır. Machine Learning, istatistiksel analizinde herhangi bir modelleme yapılmamış genetik araştırmalarda kullanılabilir. Bu sayede sadece bir veri setine özgü sonuçlar yerine belirli hastalıklara özgü modeller ortaya konabilir. Veri madenciliği ise kendine ait hazır ve denetlenen veri tabanlarına sahip araştırmalarda kullanılabilir. Bu ayrımın farkında olarak bilgi teknolojilerinin getirdiği yenilikleri istatistiksel analizlerde kullanmak gereklidir.

Bu motivasyon ile çalışmamızda bu iki yaklaşımın genetik araştırmalardaki önemini ve olası katkılarını göstermek amacıyla iki çeşit veri seti üzerinde uygulama örnekleri yapılmıştır. Bunlardan birincisi "gen ekpresyon" (mikroarray) verileri, ikincisi ise "tek nükleotit polimorfizmi (SNP)" verileridir. İlk veri seti genel kullanıma açık olan mikroarray "Kolon Kanser" veri setidir. Bu veri setinde 22'si normal, 42'si hasta olan toplam 62 bireye ait 2000 genin ekspresyon değeri mevcuttur. İkinci veri seti ise Amerika Birleşik Devletleri, Alabama Üniversitesi Hastanesi Nöroloji ve Epidemiyoloji Bölümü'nden sağlanan "Warfarin" adlı ilaca ait yapılan çalışmanın verileridir. Bu veri setinde 290 bireye ait 980 bin SNP bilgisi mevcuttur. Her iki veri seti için önemli girdi değişkenler (gen veya SNP) belirlenmiş ve en iyi tahmin (prediction) modeli belirlenmiştir. Warfarin çalışması için "Boosted Regresyon Ağacı" (BRT) yöntemi 0.69'luk belirtme katsayısı değeri ile en açıklayıcı model olmuştur. Kolon Kanser çalışmasında ise 0.79'luk doğru sınıflama değeri ile Random Forest yöntemi en iyi sonucu vermiştir. Literatürde şimdiye kadar yapılanın aksine sadece uygulama sonuçları değil henüz deneme aşamasında olan "Gen3e" adını verdiğimiz, [R] programlama dilindeki yazımı büyük oranda biten, genetik analiz paketinin temel akış şeması ortaya konulmuştur. Böylece "genetik araştırma-bilgi teknolojileri-istatistik" tek bir çatı altında kullanıldığında ne kadar faydalı oldukları somutlaştırılmıştır.

Çalışmadaki analizlerde [R] programlama dilinin analiz paketleri kullanılmıştır. Bununla birlikte açık kodlu genom ilişki analiz programı "Plink", SNP`lerde kayıp veri analiz programı "FastPhase", çok boyutlu genetik veri analiz programı "HDBStat!" programı kullanmıştır. Tüm analizler Alabama Üniversitesi, Biyoistatistik Bölümü, İstatistiksel Genetik Ünitesi"nin "High-performance computing (HPC) "leri kullanılarak yapılmıştır.