Veri madenciliği, büyük veri ve sağlıkta kullanımı

Doç. Dr. Gökhan Silahtaroğlu

1987 yılında Kara Harp Okulu Elektrik Elektronik Bölümü'nden mezun oldu. Yüksek lisans ve doktora çalışmalarını İstanbul Üniversitesi Sayısal Yöntemler Anabilim Dalı'nda tamamladı. 2012 yılında Yönetim Bilişim Sistemleri ABD'de doçent oldu. Halen İstanbul Medipol Üniversitesi Yönetim Bilişim Sistemleri Bölüm Başkanlığı görevini yürüten Dr. Silahtaroğlu'nun başlıca araştırma alanları; veri madenciliği, büyük veri analizi, makine öğrenmesi ve sistem analizi/tasarımıdır.

Büyük veri kavramı, son yıllarda oldukça çeşitli alanlarda sıkça kullanılmaya başlanmıştır. Öyle ki bu alana uzak insanlar bile bu kavramdan söz etmekte ve yararlanılması gereken bir kaynak olduğunu vurgulamaktadır. Büyük veri kavramı, medya ve iletişimden kara yolları dahil her türlü ulaşıma, sağlık kuruluşlarındaki muayene ve tedaviden günlük tekstil harcamalarımıza, yediğimiz yemeklerden gezdiğimiz yerlere kadar her şeyin dijital ortamda kayıt altına alınmasıyla ortaya çıkmıştır ve gelişimini halen sürdürmektedir. Dijital kayıtları daha önce yalnızca bilişim profesyonelleri yapar ve arşivlerken günümüzde artık yediden yetmişe, okuryazardan üniversite mezununa kadar herkes dijital ortama veri kaydedebilmektedir. Kaydedilen verilerin yoğunluğu oldukça fazladır; ortalama bir vatandaş dahi günde birkaç megabayt veriyi çok rahatlıkla dijital ortama aktarmakta hatta binlerce makine ve insanın kullanımına sunmaktadır. Bu kadar çok veri olunca da büyük veri kavramının ortaya çıkması kaçınılmaz olmuştur. Ancak büyük veri sadece çok veri değildir. Çokluk veya miktar büyük verinin sadece bir özelliğidir (1).

Aslında bu çokluk kavramını da sayısallaştırmak gerekmektedir. Büyük verinin çokluğu, terabayt seviyesine çıkmasıyla başlamıştır. Bilgisayara kaydettiğiniz her bir karakterin (harf, noktalama işareti veya sayı) dijital ortamda bir baytlık yer işgal ettiğini düşünebilirsiniz. 1 terabayt ise 1.099.511.627.776 bayta karşılık gelmektedir. Daha rahat okumamız için kabaca 1012 bayt diyebiliriz. Bazı alanlarda birikmiş verilerin boyutu ise 1015 bayta karşılık gelen petabayt ile ifade edilmektedir. Ancak bu kadar çok veriniz olsa bile büyük veriniz olmayabilir. Çünkü büyük verinin gerçekten büyük olması için başka özelliklere de sahip olması gerekmektedir. Bu özelliklerden biri verinin çeşitli formatlarda olması gerekliliğidir. Yani veriler, sayı, düz metin, resim, video, ses, mekânsal gibi farklı formatlarda olmalı ve bu formattaki veriler birbirleriyle entegre olmalıdır. Dahası bu entegrasyonun veri madenciliği algoritmalarıyla işlenebilir hale dönüştürülebilmesi gerekmektedir. Bu özelliği daha iyi kafamızda canlandırabilmemiz için şöyle bir örnek verebiliriz: Özel bir hastaneden, internet üzerinden randevu almaya çalıştığınızı düşünelim. Öncelikle sisteme kayıt olmanız gerekir. Ancak o kadar bilgiyi kim girecek? Bir de bakıyorsunuz ki "Facebook ile kaydol" seçeneği var. Bir tıkla, kaydınız tamam, artık randevu işlemine geçebilirsiniz. Bu arada ne oldu? Muhtemelen "Facebook"ta paylaştığınız dün akşamki yemeğinizin resmi ve alt yazısıyla randevu için yazdığınız şikâyet sözcükleri entegre oldu. Sizin gibi kim bilir kaç kişi daha bu şekilde randevu alıp, kayıt yaptırıyor. Hiç farkında olmadan bir bakıma veri hizmeti sunmuş oluyorsunuz. Bu entegrasyonun, diğer sosyal medya, e-devlet, internet aramaları ile olan boyutu da işin içine girince büyük veri ortaya çıkmaya başlıyor.

Tam da bu noktada artık büyük verinin tanımını daha kolay yapabiliriz: "Büyük veri, çeşitli kaynaklardan toplanmış, resim, metin ses gibi farklı formatlarda olan, sürekli çoğalan ve değişen terabayt ve daha fazla miktardaki verilerin bütünüdür."

Sağlıkta Büyük Veri

Bu tanımdan yola çıkarak web üzerinde yer alan tüm yazı, video ve seslerin hep birlikte büyük veriyi oluşturduğunu söyleyebilmekteyiz. Sağlık alanında büyük verinin kullanımı bazen şaşırtıcı bir şekilde çok basit bir uygulamayla yapılabilirken bazen de çok komplike düzenekler, algoritmalar ve veri depolama sistemlerinin kurulmasını gerektirebilmektedir. Örneğin bir arama motorunun verilerini bölgesel bazda inceleyerek herhangi bir hastalığın varlığını ve yayılma hızını evde tek başımıza bile öğrenebiliriz. Mesela geniz akıntısı sözcüğünün, geçtiğimiz son üç günde ülkenin hangi bölgesinde, ne kadar arandığını ve bu arama sıklığının son bir aydaki her bir üç günlük dönemlerle kıyaslandığında artıp artmadığını ve bu arama sıklığının bölgesel olarak ilerleme veya kayma gösterip göstermediğinin raporlanması günümüzde çok basit bir hale gelmiştir.

Büyük veriyi "tüm internet kaynakları" diyecek kadar geniş tanımladıktan sonra, biraz daha özele inerek, "Sağlıkta büyük veri nerededir ve nasıl elde edilmelidir?" sorusuna yanıt arayalım.

Sağlık alanında büyük veri kaynakları Devlet, Sağlık Bakanlığı, sağlık sigortacılığı şirketleri, her seviyeden sağlık kuruluşları, sağlık çalışanları, ilaç firmaları, bireysel sağlık hizmeti sunan hekimler, araştırmacılar, TÜBİTAK, hastalar ve yeni nesil oyuncular olarak adlandırabileceğimiz perakendeciler ve telekomünikasyon firmalarıdır (2). Her bir kaynağın sağlayacağı büyük verinin niteliği ve önemi farklıdır. Sözü edilen kaynakların veri üretim şekilleri veya veriyi ortaya çıkarış biçimleri de çeşitlidir. Tablo 1 üzerinde ana veri kaynakları ve bunların sağladığı hizmetler görülmektedir. İşte bu hizmetlerden oluşacak tüm veriler

büyük verinin bir parçası olacaktır. Örneğin hastaların sağlayacağı beslenme ve aktivite raporları, klinik çalışmalardan elde edilen PubMed, Bio Bank verileri, sosyal medyada konuşulan sağlık ile ilgili konular, sorular ve cevaplar, hasta şikâyetleri ve bunlara daha anlaşılır hale getirmek için sağlık çalışanlarının hastalara sorduğu sorular, laboratuvar sonuçları gibi verilerin hepsi bir bütün halinde büyük veriyi oluşturmaktadır. Ancak, bu verilerden anlamlı sonuçlar çıkarabilmek için tüm bu verilerin "belirli" bir düzen içinde birbirine entegre edilmesi gerekmektedir.

Büyük Veri Temelinde Son Yıllarda Sağlık Alanında Yapılan Çalışmalar

Amerika Birleşik Devletleri'nde Asthmapolis markası altında astım hastalarının kullandığı inhalerle entegre çalışan bir sensör geliştirilmiştir. 2013 yılında başlayan bu çalışmada inhalerüzerinden elde edilen çevreyle ilgili bilgiler, başka verilerle birleştirilerek hasta ve hastalık için çeşitli çözümlerin elde dilmesinde kullanılmaktadır (3). Özellikle kırsal veya ulaşımın zor olduğu bölgeler için hava durumu tahminleri, havadaki polen, kül gibi maddelerin miktarının ölçümü ve tahminiyle ilgili verilerle, bölgedeki astım hastaları, ilaç depolarındaki ilaç miktarı ve kullanım istatistiklerini kullanarak GPS temelli bir model ile hastaları uyarmak ve gerekli ilaç tedarikini sağlamak artık çok kolay bir şekilde yapılabilmektedir.

Twitter üzerinden hastanın kendi sağlık parametrelerini bir merkeze göndermesiyle başlayan süreç içinde makine öğrenmesi algoritmalarının hastanın gönderdiği metindeki parametre değerlerini ayıklayıp, hastanın sağlık durumuyla ilgili yorum yaparak, yine hastayı alması gereken önlemler açısından uyardığı bir sistem başarıyla çalışmaktadır (4). Buradan hareketle, hasta ve hekim görüşmelerinin kayıt edilmesi, yazılı hale dönüştürülmesi, işlenmesi ve yapısallaştırılmasının ardından bu verilerin hasta yakınmalarının, daha iyi anlaşılması veya hastaların günlük dille anlattıklarının tıbbi literatüre çevrilmesi mümkündür. Academic Hospital of Rennes (Fransa) bünyesindeki veri ambarı üzerinde haftalık sorgular yaparak influenza yaygınlığının ölçülmesi ve buna bağlı olarak gerekli önlemlerin alındığı bir çalışma yürütülmektedir (5). Bu sayede hekimlerin bireysel olarak karşılaştığı vakaları, zaman ve bölge bazında kümeleyerek, makro düzeyde toplum sağlığı izlenebilmekte ve buradan yola çıkarak önleyici sağlık hizmetleri hayata geçirilebilmektedir.

Benim de içinde olduğum, cihazlardan alınan anlık verilerle, yoğun bakım

Devlet/Bakanlık	Sağlık Eğitimi ve Bilgi Üretimine Katkı	Finansal Kaynak Optimizasyonu	Halk Sağlığı İzleme ve Denetimi	İletişim	Koruma	Yasalar, Anlaşmalar, Yönergeler	Medikal Performans Optimizasyonu
Sağlık Sigortacılığı					Hastalık Yönetimi		
Sağlık Kuruluşları / Hekimler						Kötüye Kullanımın Engellenmesi	
İlaç Firmaları			Ürün Geliştirme		Klinik Çalışmalar		Erken ve Etkin Müdahale
Yeni Nesil Oyuncular (Perakende, Telekomikasyon)					Müşteri Davranışı Tespiti		
Bireysel Hekimler, Diğer Sağlık Çalışanları		Koçluk, Mentörlük		Klinik Karar Destek Sistemleri			
Araştırmacılar / TUBITAK							
Hastalar			Prognosis		Bilgi takası P2P		

Tablo 1: Ana veri kaynakları ve bunların sağladığı hizmetler

hastalarında ortaya çıkacak olan komplikasyonların, henüz çıkmadan tahmin edilmesi çalışmaları da ülkemizde yürütülmeye başlamıştır.

Sonuç

İstatistiksel analiz teknikleri, elle toplanan hasta verileri üzerinde yapılan çalışmalarda yıllardır başarıyla uygulanmaktadır. Bu çalışmalarda öne çıkan unsur, kullanılan istatiksel teknik ya da yazılım olarak görünse de aslında verinin elde edilmesi, veri miktarının çokluğu ve güvenilirliği daha fazla önem taşımaktadır.

Yaşadığımız dijital çağda her gün gözümüzün önünden birçok veri akıp gitmektedir. Bunlar; forumlarda hasta ve hasta yakınlarının konuşmaları, sosyal medyadan hekimlere yöneltilen sorular ve yanıtları, arama motorlarından yapılan ilaç ve/veya hastalık aramaları gibi yapılandırılmamış verilerdir. Sağlıkta büyük veri ise tetkik kayıtları, halk sağlığı verileri, sağlık sigortası verileri, bilimsel araştırma sonuçları ve sosyal ağlarda kullanıcılar tarafından üretilen verilerin bütününü temsil etmektedir. Günümüzde sadece son iki hafta içinde arama motorlarından yapılan sağlıkla ilgili aramaları, bölgesel olarak kümelediğimizde, herhangi bir yerdeki salgın veya yaygın hastalığın varlığını ortaya çıkarabilmekteyiz. Gelinen bu noktada sağlık kuruluşlarımızda tutulan hasta kayıtları, tanı ve tetkik sonuçları istatistiksel yöntemlerin dışında, yapay sinir ağları, karar ağaçları gibi veri madenciliği yöntemleriyle analiz edilebilir durumdadır. Hatta bu verilerin internet üzerinden elde edilecek büyük verilerle birleştirilmesiyle analize bambaşka bir boyut kazandırılabilir. Söz konusu büyük verinin bu şekilde analizi, koruyucu ve önleyici sağlık hizmetlerine katkı sağlayacaktır. Hizmet alacak hastanın, sağlık kurumundan içeri girişinden itibaren elde edilecek resim ve videoların makine öğrenmesi algoritmalarıyla analizi, dış bulguların tespiti ve değerlendirmesi açısından hekime bir karar destek hizmeti sunacaktır. Tetkiki yapılan hastanın, sonuçları saniyeler içinde yüzbinlerce hasta ile karşılaştırılıp olası tanı ve tedavi yöntemleri, belirli olasılık ve güven seviyeleriyle hekime sunulabilecektir. Elde edilecek bu yeni bilgiler bilimsel araştırmalarda kullanılabileceği gibi sağlık eğitimlerinde de kullanılabilecektir. Tüm bu çalışmalar yapılırken hasta mahremiyeti göz ardı edilmemelidir. Ayrıca bu çalışmalar, hekimi dışarıda tutmak üzere değil aksine hekime tanı ve tedavisinde yardımcı olmak üzere tesis edilmelidir. Örneğin hekimin bir görüntüleme tetkik sonucunu, kuşkulandığı bir hastalık üzerine bilgisayar ortamına sürüklemesiyle, bu tetkikin, o hastalık tanısı konmuş binlerce hastayla eşleştirilmesi ve hekime kuşkusu hakkında bilgi vermesi hekimin tanılarını güçlendirecektir.

Kaynaklar

1) M. D. Assuno, R. N. Calheiros, S. Bianchi, M. A. S. Netto, and R. Buyya, "Big Data Computing and Clouds: Trends and Future Directions," J. Parallel Distrib. Comput., vol. 79–80, pp. 3–15, 2015.

2) S. D. Young, "Behavioral Insights on Big Data: Using Social Media for Predicting Biomedical Outcomes," Trends Microbiol., vol. 22, no. 11, pp. 601–602, 2014.

3) J. Sarasohn-Kahn, "Making Sense of Sensors: How New Technologies Can Change Patient Care California HealthCare foundation," Pew Internet Am. Life Proj. Manhattan Res. IMS Res., no. February, 2013.

4) L. R. Nair, S. D. Shetty, and S. D. Shetty, "Applying Spark Based Machine Learning Model on Streaming Big Data for Health Status Prediction," Comput. Electr. Eng., 2017.

5) G. Bouzillé et al., "Leveraging Hospital Big Data to Monitor Flu Epidemics," Comput. Methods Programs Biomed., vol. 154, pp. 153–160, 2018.