VERİ MADENCİLİĞİ VE YER BİLİMLERİ

Mehmet Seval KAYGULU*

Günümüzde veri birikimi büyük bir hıza ulaşmıştır. Elde edilen gözlemsel sonuçlara göre var olan veriler her 18 ayda hemen hemen iki katına çıkmaktadır. Bu yüzden, veri depolarının büyüklükleri de “terabyte” lar düzeyine erişmiştir. Bu boyuttaki bilgilerin elle ve klasik yöntemlerle incelenmesi, önemli, anlaşılabilir ve kullanılabilir bilgilere ulaşılması hemen hemen imkânsızdır.

Çünkü, geleneksel olarak, araştırmacılar işlemleri el ile yönetirler. İstatistiksel metotlarda ise veriler çoğunlukla rasgele seçilirler ve sınırlı sayıda olurlar. Aynı zamanda, araştırmacı konuyu ve verileri iyi tanıyor olmalıdır. Bunlarla beraber, veri miktarındaki ve verileri tanımlayan özelliklerin yani sahaların (attribute) sayısındaki artışlar nedeni ile geleneksel yöntemler yavaş kalmakta, pahalıya mal olmakta ve gerekli tüm bilgiye ulaşılamamaktadır.

Bu nedenle, 80'li yıllardan bu yana bu konuda bilgisayar kullanımı yaygınlaşmıştır. Özellikle tıp ve ticaret alanlarında pek çok örnek bulunmaktadır. Buna karşın, yerbilimlerinde bilgisayar kullanımının oldukça sınırlı olduğu gözlenmiştir.

VERİ MADENCİLİĞİ

Bilgisayar kullanımı ile bilgiye ulaşma çalışmaları sınıflandırma (classification) ve salkımlama (clustering) teknikleri ile 80'li yıllarda başlanmıştır. 90'lı yılların başında birden çok inceleme tekniğinin bir arada kullanılmasını içeren “Veri Madenciliği-VM” (Data Mining-DM) kavramı geliştirildi. Verileri depolama olanaklarının artması ve ucuzlaması nedeni ile daha önce gereksiz kabul edilen veriler de saklanmaya başlandı. Veri çeşitleri de artış gösterdi. Metin ve tablo biçimindeki verilere ses, resim ve hareketli görüntüler de eklendi. Büyük boyutta ve değişik tipte veri barındıran veri depoları (Data Warehouse) oluşturuldu.

Bu gelişmeler bilgiye ulaşmayı daha karmaşık hale getirdi. 90'lı yılların ortalarında bu işlem “Veri Tabanlarından Bilgi Keşfetme-VTBK” (Knowledge Discovery From Databases-KDD) adı altında daha sistematik bir hale getirildi. VM ise VTBK işlemler zincirinin bir halkası olarak tanımlanmıştır. Ancak VM ismi daha yaygın olarak bazen, VTBK yerine de kullanılmaktadır.

VTBK birbirlerini etkileyen işlem basamaklarından oluşur. Çalışmanın herhangi bir aşamasında, çoğunlukla, daha önceki işlemlere yeniden dönmek gerekir (iterative). Bilim insanları VTBK'nın işlem basamaklarını çeşitli şekillerde tanımlamışlardır. Ancak temel söylemler aynıdır. Aşağıda kısaca bu basamakları anlatacağız.

Konunun tanınması

Üzerinde çalışılacak verilerin tanınması, konu hakkında bilgi edinilmesi gerekecektir. Bunun için ilgili kitap veya uzmanlardan yararlanılabilir ve, eğer varsa, bu konuda daha önce yapılmış araştırmalar incelenebilir.

*173 sk. No:27 D:13 Basın Sitesi, 35440 İZMİR

Üzerinde VM uygulamasının yapılacağı örnek veri tabanının seçilmesi

VM uygulamalarında kullanılan verilerin en önemli farkı gerçek dünyadan alınmış olmalarıdır. Bu yüzden, gerekli sahalardan bir veya bir kaçı bulunmayabilir. Bazı elemanların da özelliklerinde eksiklik veya yanlışlıklar bulunabilir. Veri kirliliği veya parazit diyebileceğimiz bu durumların giderilmesine çalışılmalıdır.

Veriler hakkında bir ön bilgiye sahip olduktan sonra verilerin sunum yapısı (ilişkisel veri tabanı, veri küpü gibi) ve mantıksal yapısı yani veri tabanı yönetim sistemi seçilir. Bilgiye ulaşma aşamasında elde edilecek sınıfların ve kalıpların (pattern) sayısını incelenebilir düzeyde tutmak ve anlamsız oluşumları baştan engellemek için veri tabanındaki değişkenleri, eğer mümkünse, azaltmak gerekebilecektir. Bazı durumlarda bir sahada yer alan farklı değerlerin sayısı çok fazla olabilir. Sahanın içeriği sayısal ise aralıklar belirleyerek, içerik metin biçiminde ise kod numaraları vererek farklı değer sayısını azaltmak da gerekebilir. Bunun yanı sıra elemanları tekil olarak tanımlayacak olan saha veya sahalar (key attribute) seçilir. Eğer veriler birden çok kaynaktan elde ediliyorsa konu ile ilgili olanlar alınarak birleştirilmelidir.

Veri madenciliği

Bu aşamanın VTBK nin temel basamağı olduğunu söyleyebiliriz. Veri madenciliğinin hedefi veriler arasında gizli kalmış bilgilere dahi ulaşmamızı sağlayacak uygun, kullanılabilir ve anlamlı karar kalıplarının elde edilmesidir. Bu işlem için kullanılan VM yöntemleri, genel olarak, "Tanımsal" (Descriptive) ve "Tahminsel" (Predictive) diye ikiye ayrılır. Her yöntem, kendi içinde kullanılan algoritmalara bağlı olarak sınıflara ayrılır. Salkımlama, özetleme, eşleştirme veya ilişkilendirme, bağlantılılık kuralı ve ardıllık keşfi tanımsal yöntemler, sınıflandırma, zaman serileri analizi ve tahminleme ise tahminsel yöntemler içinde yer alır.

Yukarıda isimleri verilen yaklaşımlar ayrıca "Yönlendirilmiş" (Supervised) ve "Yönlendirilmemiş" (Unsupervised) diye ikiye ayrılabilir. Yönlendirilmiş yöntemlere sınıflandırma, diğerine ise salkımlama örnek gösterilebilir.

Yönlendirilmiş algoritmalarda sınır şartları ve eşik değerleri kullanıcı tarafından belirlenir. Hakkında inceleme yapılacak sahalar da kullanıcı tarafından seçilir. Bu olanak bir kullanıcı arayüzü (User Interface) oluşturularak kullanıcıya sunulur. Bu arayüz kullanıcı ile program arasında iletişimi ve etkileşimi sağlayarak işlemlerin yinelenmesinde de yardımcı olacaktır.

VM'de bilgiye ulaşmak için bu konu ile ilgili başka alanlar ve bilim dallarındaki bilgiler de kullanılır. Mantık, veri depoları kullanımı, OLAP, İstatistik, yapay zeka, yapay sinir ağları bu alanlardan bazılarıdır. Tüm bu alanlar kullanılarak oluşturulan, verileri işleyen, inceleyen ve sonuç ürüten yani sınıfları ve kalıpları keşfeden algoritmalar bütününe "Veri Madenciliği Motoru" (Data Mining Engine) adı verilmektedir.

Sonuçların sunumu ve yorumlanması

VM sonucunda elde edilen kalıplar kullanıcıların anlayabileceği bir şekle dönüştürülerek sunulmalıdır. Mümkünse sonuçların yorumları da kullanıcıya sunulur. Bu sonuncusunun nedeni, kullanıcıların her zaman, konunun uzmanı olmalarının beklenmemesidir. Ayrıca, kullanıcı elde ettiği sonuçları, kendi bilgisi çerçevesinde yorumlayabilir.

VERİ MADENCİLİĞİNİN YER BİLİMLERİNE UYGULANMASI

Yer bilimlerinde zemin hakkında bilgi edinmek için çeşitli yöntemler bulunabilir. Burada, kuyu açılarak elde edilen zemin verileri esas alınmıştır. Bunun nedeni, sadece Seyitömer Kömür Havzası bilgilerine ulaşma olanağı bulunabilmesidir. Aşağıda, verilerin incelenmesi sonucu karşılaşılan sorunlar ve önerilen çözümler anlatılmaktadır.

İlk sorun, kullanılacak veri tabanının tipinin seçilmesidir. Veri tabanı tüm bilgilerin kullanılabileceği bir şekilde oluşturulmalıdır. Çok gerekli olmadıkça, bir sahadaki değerlerin tekrarlanması istenmez. Bu durum kalıpların oluşturulmasında zorluk yaratabilir. Ayrıca bellekte gereksiz yer kaplar. Benzer nedenler ile "BOŞ" (NULL) değerlerin bulunması istenmeyen bir durumdur. Bu açıdan bakıldığında ilişkisel veri tabanı kullanımının uygun olacağı görülmüştür. Nesne yönelimli veri tabanı veya veri küpü gibi seçeneklerde boş hücre sayısı çok olacaktır. Tek bir büyük tablo kullanımı özetleme, sınıflandırma ve salkımlama işlemlerinde kolaylık sağlamaktadır. Ancak, en fazla tekrarlanan değer ve boş hücre sayısı bu tabloda bulunacaktır. İlişkisel veritabanında ortak sahalara sahip olan veriler küçük boyutlu tablolarla gösterilerek bu sakıncalar giderilebilir. Tablolar arası ilişkiler kuyu ve katman numaraları ile sağlanabilir. Ayrıca standart sorgulama dili (Standard Query Language-SQL) kullanılarak basit sorgular ve gereken yeni tablolar oluşturmak çok kolaylaşacaktır.

Diğer önemli sorun verilerin ifade ediliş biçimleridir. Günlük hayatta "kil, yeşil" ile "yeşil kil" aynı anlama gelmesine rağmen, bilgisayar bunları farklı veriler olarak algılayacaktır. Bu anlamsız kalıpların oluşmasına neden olacaktır. Dolayısı ile bu konuda bir standardın oluşması gerekmektedir. Bu sağlanamıyor ise program, kullanıcının kendi standardını oluşturmasına izin vermelidir. Sayısal verilerde de katman kalınlık ve derinliklerinin çok çeşitli olması sorun yaratacaktır. Bu sorun belirli aralıklar seçilerek bir ölçüde giderilebilir. Belirli bir değerden küçük olan katman kalınlıkları da bir eşik değeri seçilerek elenebilir. Ayrıca kullanıcıya, önemli görmediği katmanları aynı isim ve kod numarası altında toplama olanağı verilmelidir. Böylece kalıp ve sınıf sayıları kabul edilebilir düzeyde tutulabilir.

Kuyulardaki katmanların sayıları, tipleri, kalınlık ve derinlikleri çok büyük farklılıklar gösterdiğinden salkımlama gibi yönlendirilmemiş bir teknik kullanılırsa kuyu sayısına eşit sayıda grup oluşabilir. Yönlendirilmiş tekniklerin daha uygun olacağı açıktır. Kullanıcı istediği sahalara, bir veya daha fazla sahadaki değerlere göre kuyuları sınıflandırabilir, özet bilgiler oluşturabilir. Bu bilgiler tekrar incelenmek üzere saklanmalıdır. Böylece çeşitli seçeneklere göre elde edilen sonuçları karşılaştırabiliriz. Bunu sağlayabilmek için veriler üzerinde yapılan düzenlemeler yaratılan geçici veritabanlarında saklanmalı, asıl veritabanını oluşturan tablolar ilk halini korumalıdır. Bu sayede, elde edilen yeni veriler veritabanına eklenebilir ve incelemeler son duruma göre tekrar edilebilir. Elde edilen sonuçlar uygun tablolara dönüştürülüp kullanıcıya sunulmalıdır. Kullanıcı bunlara kendi yorumunu ekleyebilecektir.

Esas veri tabanı kullanılarak ek faydalar sağlanabilir. İlk akla gelenleri şöyle sıralamak mümkündür:

a-Her kuyunun düşey kesitinin çizimi bilgisayara çizdirilebilir. Bu, zamandan ve insan gücünden ekonomi sağlayacaktır.

b-Kuyu koordinatları kullanılarak, havzada kuyuların yerleri belirlenebilir ve bir harita oluşturulabilir.

c-Bu harita üzerinden seçilen kuyuların düşey kesitleri görüntülenerek katmanların eğilimleri saptanabilir. Bu bilgi açılması düşünülen kuyuların sayısında azalma sağlayabilir.

Bu tip çalışmalara konunun uzmanları da katılırsa beklentilere ve gereksinimlere daha uygun programlar oluşturmak mümkün olacaktır.