Ömer Osman Durmuş, Asaf Varol, Esra Mutluay, "Veriden Bilgiye Ulaşmada Veri Madenciliğinin Önemi", Ulusal Teknik Eğitim, Mühendislik ve Eğitim Bilimleri Genç Araştırmacılar Sempozyumu (UMES 2007), 20-22 Haziran 2007, Kocaeli Üniversitesi, Kocaeli, Bildiriler Kitabı, S. 108-112

4.46. VERİDEN BİLGİYE ULAŞMADA VERİ MADENCİLİĞİNİN ÖNEMİ

[1] Ömer Osman DURSUN **[2] Asaf VAROL** **[3]Esra MUTLUAY**

[1,2,3] Fırat Üniversitesi Teknik Eğitim Fakültesi,

Elektronik-Bilgisayar Eğitimi Bölümü, Elazığ

[1]e posta: omerdursun23@yahoo.com

[2]e posta: avarol@firat.edu.tr

3e posta: esra.mutluay@hotmail.com

ÖZET

Senelerce her alanda birçok kayıt tutulmuştur. 1960'lı yıllardan itibaren bu kayıtlar veri tabanları ve veri ambarlarında depolanmaya başlanmıştır. Tutulan kayıtların veri tabanlarındaki sorgu, raporlama yöntemleri ile işlenmesi, insanların ihtiyaçlarını karşılamamaya başlamıştır. Sürekli artan verinin içinde gerekli, gereksiz, boş bilgilerin ve tutarsız verilerin bulunması, sistemleri hantallaştırmıştır.

Mevcut kayıtların insanların belli isteklerine cevap verip amaçlarına hizmet edebilmesi için, kayıtlar veri olmaktan çıkıp bilgi düzeyine erişmelidir. Uygun yazılımların gelişimi ve toplanan verilerin bilgiye çevrilme isteği, verileri işleyerek veri içindeki kullanılabilir ve ilginç ilişkilerin ortaya çıkarılması, gerekli hale gelmiştir. Bunu yapabilecek olan bilgisayar, onu yönlendirip programlayacak olan da insanın beyin gücüdür. Böylelikle verilerin içinden anlamlı ve amaca hizmet edebilecek örüntülerin çıkarılması için veri madenciliği kullanılmalıdır.

Bu çalışmada, Elazığ iline ait sıcaklık, çiğ noktası, nem oranı ve basınç verileri, veri tabanından alınarak ANFIS (Adaptive Network Based

Ömer Osman Durmuş, Asaf Varol, Esra Mutluay, "Veriden Bilgiye Ulaşmada Veri Madenciliğinin Önemi", Ulusal Teknik Eğitim, Mühendislik ve Eğitim Bilimleri Genç Araştırmacılar Sempozyumu (UMES 2007), 20-22 Haziran 2007, Kocaeli Üniversitesi, Kocaeli, Bildiriler Kitabı, S. 108-112

Fuzzy Inference System) ve YSA (Yapay Sinir Ağı) sınıflandırıcılarından ayrı ayrı geçirildikten sonra hem hava tahmininin yapılması hem de bu sınıflandırıcıların performanslarının değerlendirilmesi yapılarak, öneriler sunulmuştur.

Anahtar kelimeler:

Veri Madenciliği, Bilgi Keşfi, Veri Tabanı, Veri Analizi.

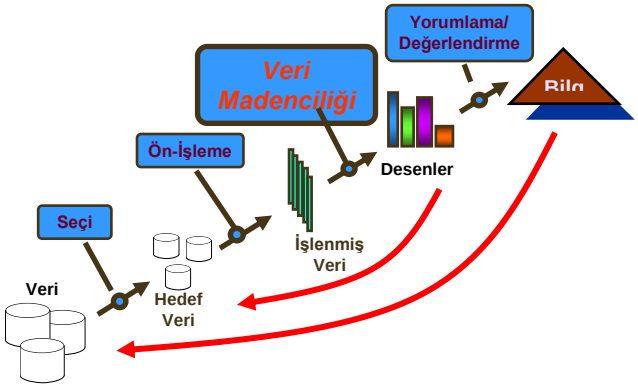
1. GİRİŞ

Dünyada veri miktarı her geçen saniye katlanarak artmaktadır. Artan veri miktarına paralel olarak insanların bilişim dünyasından beklentileri ve ihtiyaçları da artmakta işlemci hızları ve belleklerin kapasitesi artıp fiyatları düşmektedir.

1995 yılında birincisi düzenlenen Knowledge Discovery in Databeses konferansındaki bildiri kitabının sunuşunda şöyle bir paragraf bulunmaktadır [1].

"Dünyadaki enformasyon miktarının her 20 ayda ikiye katlandığı tahmin edilmektedir. Bu ham veri seli ile ne yapmamız gerekmektedir. İnsan gözü bunun ancak çok küçük bir kısmını görebilir. Bilgisayarlar bilgelik pınarı olmayı vaad etmekte ancak veri seline sebep olmaktadır."

Veri madenciliği istatistik ve matematik teknikleri ile birlikte örüntü tanıma teknolojilerini kullanarak depolama ortamlarında saklı bulunan veri yığınlarının elenerek anlamlı yeni korelasyon, örüntü ve eğilimlerin keşfedilme sürecidir [2].



Şekil 1: Veri Madenciliği Süreçleri [3].

2.VERİ MADENCİLİĞİNİN KULLANIM ALANLARI

Veri Madenciliği endüstriden astrolojiye birçok alanda kullanılmaktadır. Veri Madenciliğinin alanları, aşağıdaki şekilde gruplanabilir [4].

►Veri analizi ve karar destek sistemleri

- Pazar araştırmaları ve yönetimi

  - Hedef pazarlama, müşteri ilişkileri yönetimi, sepet analizli, çapraz satış, pazar gruplama.

  Örneğin sepet analizinde markette alışveriş yapanlardan zeytin ve peynir 'in toplam fiş hareketlerinde birlikte görülme olasılığı %3.67'dir. Zeytin alan müşterilerin %70.11 olasılıkla peynir de aldığı söylenebilir [5].

- Risk analizi ve yönetimi

  - Tahmin müşteri memnuniyeti, kalite kontrol, rekabet analizi.

Örneğin iyi analiz edilmiş bir müşterinin özellikleri, hangi ürünlere genelde ihtiyaç ve ilgi duyduğu gibi bilgilere sahip olunması, sunulacak ürünü belirlemeye böylelikle müşteri memnuniyetinin artmasına sebep olur.

- Sahtekarlık ve yaygın olmayan şablonların yakalanması

  Örneğin bankadan kredi isteyen müşterinin özelliklerine bakılır, önceden iyi sonuç alınmayan ödemesini geciktiren veya yapmayan müşterilerin özellikleri ile karşılaştırılır. Buna göre kredi verilip verilmeyeceğine karar verilir.

►Diğer uygulamalar

- Metin madenciliği
- DNA ve biyolojik (Bir hastada x ameliyatından sonra ilk üç gün içinde y enfeksiyonunun oluşması ihtimali.)[6].
- Web madenciliği (Web üzerinden faydalı bilgiyi keşfetmek ve yeni çıkarımlar oluşturmak)

3.VERİ MADENCİLİĞİ İŞLEM ADIMLARI

3.1 Verilerin Elde Edilmesi

Birçok alanda kullanımı olan veri madenciliğinde problem belirlendikten sonra bu verilerin elde edilme ve depolanma yöntemleri vardır. Bunlar veri ambarları, veri dosyaları, veri tabanları ve internet olabilir.

Alınan bu verilerin amaca uygun şekilde hazırlanması kullanacağımız enerji ve zamanın yarısından fazlasını alır. Yapılması gereken ilk iş tanımlanan problem için gerekli olduğu düşünülen verilerin

alınmasıdır. Alınan veriler arasında uyumsuzluklar göz önüne alınarak veriler tekrardan düzenlenmeli ve tek bir havuzda toplanmalıdır. Kurulacak modele göre veri seçimi, verinin görselleştirilmesi sağlanmalıdır. Gerekli görülen yerde veriler kullanılacak algoritma ya da model için uygun görülen dönüşümler yapılmalıdır.

Bilgi bulma süreci bittikten sonra elde edilen veriler ilgili alan açısından yorumlanarak problem üzerindeki etkisi belirlenir. Tanımlanan problem için hazırlanan verileri en uygun yöntemi kullanarak bilgiye çevirmelidir.

3.2 Veri Madenciliği Modelleri

Kullanılacak model, veri türüne ve probleme bağlıdır. Veri madenciliğinde tanımlı iki tür model ve her bir modele ait birçok algoritma vardır.

Bu modellerden ilki bilinen verilerden yorumsuz gerçeğe ulaşmadır. Kümeleme birliktelik ve açık zamanlı örüntüler bu modelin algoritmalarıdır. Örneği en çok satılan ürün var olan verilerden yorum eklenmeden çıkarılan sonuçtur.

Diğer model ise bilinen verileri kullanarak yorumlamak, tahmini yeni verilere ulaşmaktır. Mesela altının değeri her ayın ortalarında artıyorsa, bu ayın ortalarında da artacağı çıkartılabilir. Dikkat edilmesi gereken nokta, buradaki verinin süreklilik gösteren bir değere sahip olmasıdır. Bu alanda kullanılan algoritmalar Genetik Algoritma, Yapay Sinir Ağları, Bayes Yöntemi, Karar ağaçları, Bellek Tabanlı Modelleme gibi algoritmalardır.

3.3 Veri Madenciliğinde Kullanılan Algoritmalar

Genetik Algoritma: İlk olarak populasyon diye tabir edilen bir çözüm (kromozomlarla ifade edilir) seti ile başlatılır. Bir populasyondan alınan sonuçlar bir öncekinden daha iyi olacağı beklenen yeni bir

populasyon oluşturmak için kullanılır. Yeni populasyon oluşturulması için seçilen çözümler uyumluluklarına göre seçilir. Bu istenen çözüm sağlanıncaya kadar devam ettirilir [7].

Yapay Sinir Ağları: Genel anlamda YSA, beynin bir işlevini yerine getirme yöntemini modellemek için tasarlanan bir sistem olarak tanımlanabilir. YSA, yapay sinir hücrelerinin birbirleri ile çeşitli şekilde bağlanmasından oluşur ve genellikle katmanlar şeklinde düzenlenir. Donanım olarak elektronik devrelerle yada bilgisayarlarda yazılım olarak gerçeklenebilir [8].

Karar Ağaçları: İstatistiksel yöntemlerde veya yapay sinir ağlarında veriden bir fonksiyon öğrenildikten sonra bu fonksiyonun insanlar tarafından anlaşılabilecek bir kural olarak yorumlanması zordur. Karar ağaçları ise veriden oluşturulduktan sonra ağaç kökten yaprağa doğru inilerek kurallar (IF-THEN rules) yazılabilir. Bu şekilde kural çıkarma (rule extraction), veri madenciliği çalışmasının sonucunun doğrulanmasını sağlar. Bu kurallar uygulama konusunda uzman bir kişiye gösterilerek sonucun anlamlı olup olmadığı denetlenebilir. Sonradan başka bir teknik kullanılacak bile olsa, karar ağacı ile önce bir kısa çalışma yapmak, önemli değişkenler ve yaklaşık kurallar konusunda analiste bilgi verir ve daha sonraki analizler için yol gösterici olabilir [9].

Kümeleme Modelleri: Kümeleme modellerinde amaç üyelerinin birbirlerine çok benzediği, ancak özellikleri birbirlerinden çok farklı olan kümelerin bulunması ve veri tabanındaki kayıtların bu farklı kümelere bölünmesidir. Kümeleme analizinde; veri tabanındaki kayıtların hangi kümelere ayrılacağı veya kümelemenin hangi değişken özelliklerine göre yapılacağı, konunun uzmanı olan bir kişi tarafından belirtilebileceği gibi veri tabanındaki kayıtların hangi kümelere ayrılacağını geliştirilen bilgisayar programları da yapabilmektedir [6].

Birliktelik Kuralları ve Ardışık Zamanlı Örüntüler : Bir alışveriş sırasında veya birbirini izleyen alışverişlerde müşterinin hangi mal veya hizmetleri satın almaya eğilimli olduğunun belirlenmesi, müşteriye daha fazla ürünün satılmasını sağlama yollarından biridir. Satın alma eğilimlerinin tanımlanmasını sağlayan birliktelik kuralları ve ardışık zamanlı örüntüler, pazarlama amaçlı olarak pazar sepeti analizi (Market Basket Analysis) adı altında veri madenciliğinde yaygın olarak kullanılmaktadır. Bununla birlikte bu teknikler, tıp, finans ve farklı olayların birbirleri ile ilişkili olduğunun belirlenmesi sonucunda değerli bilgi kazanımının söz konusu olduğu ortamlarda da önem taşımaktadır [10].

Bellek Tabanlı Yöntemler: Bellek tabanlı veya örnek tabanlı bu yöntemler (memory-based, instance-based methods; case-based reasoning) istatistikte 1950'li yıllarda önerilmiş olmasına rağmen, o yıllarda gerektirdiği hesaplama ve bellek yüzünden kullanılamamış ama günümüzde bilgisayarların ucuzlaması ve kapasitelerinin artmasıyla, özellikle de çok işlemcili sistemlerin yaygınlaşmasıyla, kullanılabilir olmuştur. Bu yönteme en iyi örnek en yakın k komşu algoritmasıdır (k-nearest neighbor) [10].

Bayes Teoremi: bir sınıflandırma sorununun olasılı terimleriyle açıklanabileceği varsayımına dayanır. Bayes kuralı bir veri grubunda bir özelliğin olasılığını tahmin etme yöntemidir; belirli bir veri değerinde çeşitli varsayımların olasılığını araştırır [11].

4.VERİ MADENCİLİĞİNE KATKISI OLAN DİĞER YAKLAŞIMLAR

Veri madenciliği teorik ve sezgisel yaklaşımı birleştirir. **Makine öğrenmesinde** daha çok sezgisel öğrenmenin başarımını artırmaya çalışılır. **İstastistik** bir varsayımın doğruluğunu araştırır [12]. **Veri tabanı** yönetilebilir, güncellenebilir, taşınabilir, birbirleri arasında tanımlı ilişkiler bulunabilen bilgiler kümesidir.

Veri tabanı ile Veri Madenciliği arasındaki farklılığı aşağıdaki örnekle daha iyi anlayabiliriz.

Veri tabanı

- Bebek bezi alan müşterileri bul
- İlgili tarih aralığındaki kredi kartı taleplerini bul
- Devamsızlığı 18 günü aşan öğrenciyi bul

Veri madenciliği

- Bebek bezi ile satılan ürün
- Kredi kartı limitini ödeyebilecek talepleri bul
- Devamsızlık alışkanlıkları benzer öğrencileri bul.

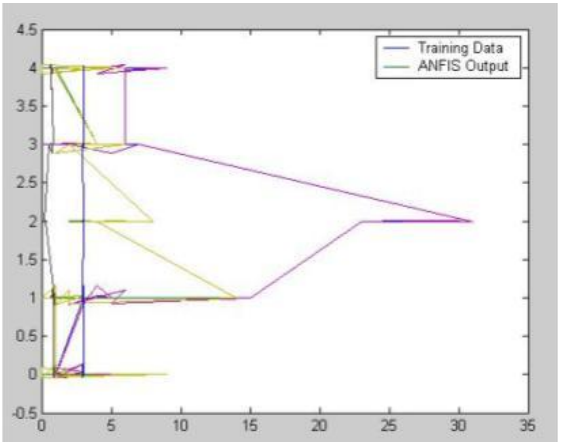
5. VERİ MADENCİLİĞİ SORUNLARI

Farklı zamanlarda ve farklı yerlerde alınan veriler arasında uyumsuzluklar oluşabilir. Uyumsuzluk sebebi fiziki koşulların yanında veri türünden kaynaklanabilir. Ayrıca veri tabanına eklenen verinin eksik, boş olması veya kesin olmaması da bir sorundur. Bunların takip edilmesi çok zordur. Veri deposuna eklenen veriler sürekli güncellenir, yeni veriler eklenir, değişiklikler yapılır. Bunlar eklenirken veya bu verilerin eldesi için çalışılırken, verilerin gürültüden etkilenmemesi imkansızdır. Gürültü ortadan kaldırılamaz ama en aza indirgenirse bu sorunun etkisi azaltılmış olur. Farklı veri tabanlarına aynı verinin eklenmesi veri tutarsızlığını oluşturur eklenen verilerin özelliğine, boş olup olmadığına, tutarlılığına ya da amaca uygunluğuna bakılmadan alınması, ilgili veri deposunun hantallaşmasına sebep olur.

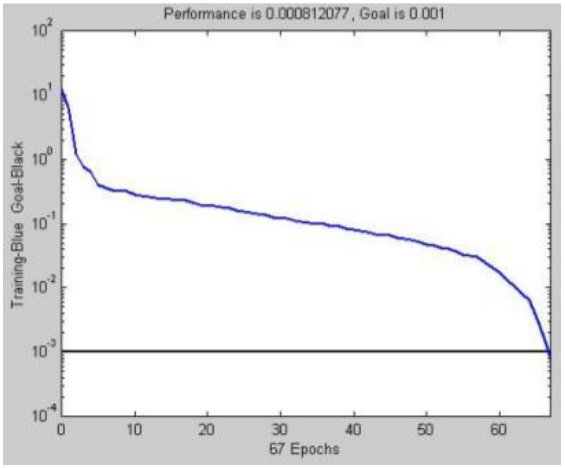
Ayrıca veri madenciliğinde yapılacak analiz ve çıkarımlar için kişisel bilgilere ihtiyaç duyulur ve büyük çoğunlulukla bu olay kişilerden izinsiz yapılır, bunun kanuni sorumluluğu vardır.

6. ELAZIĞ'DAKİ METEOROLOJİK VERİLERİN VERİ MADENCİLİĞİ İLE DEĞERLENDİRİLMESİ

Elazığ iline ait sıcaklık, çiğ noktası, nem oranı ve basınç verileri, veri tabanından alınarak ANFIS (Adaptive Network Based Fuzzy Inference System) ve YSA (Yapay Sinir Ağı) sınıflandırıcılarından ayrı ayrı geçirildikten sonra hem hava tahmininin yapılması hem de bu sınıflandırıcıların performanslarının değerlendirilmesi sağlanmıştır. 4 sınıftan oluşan 40 tane giriş verisi, toplam veri sayısı 160 olan veriler sınıflandırıcılara verilerek, eğitim sürecinden geçirilmiştir. Eğitim sürecinden geçirilen bu verileri bilgisayarın algılayıp hangi karar sınıfına ait olduğunun yorumlaması sağlanmıştır. Eğitim işlemi tamamlandıktan sonra yaklaşık 4 sınıftan oluşan 100 veri toplam 400 veri test süreci için sınıflandırıcılara verilmiştir. Elde edilen sonuçlar aşağıdadır (Şekil 2-3) [13]
.



Şekil 2: ANFIS sınıflandırıcısıyla eğitilen verilerin grafiği [13].



Şekil 3: YSA sınıflandırıcısıyla eğitilen verilerin grafiği [13]

7. SONUÇ

Bu çalışmada, meteorolojik hava tahmini uygulaması gerçekleştirilmiştir. Bu uygulama için 40 günlük meteorolojik veriler alınmıştır. Alınan sıcaklık, çiğ noktası, nem, basınç gibi anlamsız meteorolojik veriler toplanarak geniş bir veri tabanı oluşturulmuştur. Oluşturulan veri tabanındaki sıcaklık, çiğ noktası, nem ve basınç verileri normalizasyon işleminden geçirilmiştir. Normalize edilen sıcaklık, çiğ noktası, nem, basınç değerleri YSA ve ANFIS sınıflandırıcılarına giriş değeri olarak verilmiştir. Sınıflandırıcıdan doğru çıkış alabilmek için girişlerin eğitilerek bilgisayarın öğrenmesi sağlanmıştır. Böylece sonraki günün hava tahmini olayı gerçekleştirilmiştir. Hata tahmini aralığını azaltabilmek için ANFIS sınıflandırıcısının kural tabanı sayısı 16'dan 625' e çıkarılmıştır. Böylece hata oranının azaldığı görülmüştür [13].

14.02.2003 – 09.02.2004 tarihleri arasında rasgele alınan 100 günlük veri kullanılarak hava tahmini yapılmıştır. Gerçekleştirilen bu hava tahmininde ANFIS sınıflandırıcısının kural tabanı 16 iken 54 gününün hava

tahmini doğru olarak bulunmuştur. ANFIS sınıflandırıcısının kural tabanı 625'e çıkarıldığında, 65 günün hava tahmini doğru olarak bulunmuştur. YSA sınıflandırıcısı ile yapılan tahminde ise 55 günün hava durumu doğru olarak tahmin edilmiştir [13].

Matlab ile geliştiren bir veri madenciliği programı sayesinde, hava tahminlerinde %50'nin üzerinde çok daha doğru tahminlerin yapılabileceği görülmüştür. Yukarıda belirtilen sonuçlardan hareket ederek, ANFIS sınıflandırıcısının kural tabanının 625'e çıkarılarak, %65 gibi doğru hava tahminlerinin yapılabilmektedir. Bu amaçla geliştirilen veri madenciliği yazılımı kullanılarak, bundan böyle Elazığ için daha isabetli hava tahminleri yapılabilecektir [13].

Veri madenciliği verilerin tutarlı hale getirilmesinde hızlı ve tutarlı veri bütünlüğü sağlayan, personel ve zaman açısından kazanç sağlayan, verinin kalitesini artıran ve veriden bilgi elde etme yolları hedef alınmalıdır. Bu yöntemi kullanırken gerçekler ve gerçeklere bağlı olan tahminlerden yararlanılarak uygun model seçilmeli ve modele uygun algoritma hazırlanmalıdır.

Meteoroloji alanında yapılan çalışmada, veri madenciliği işleminde başarıya ulaşabilmek için, bol miktarda kaliteli veriye ihtiyaç olduğu saptanmıştır. Eksik veri olması durumunda, sağlıklı sonuçlar çıkartılamamıştır. Veri madenciliği bu bağlamda bir adım değil süreçtir. Bir kişi ile değil konuya hakim olan uzman bir ekiple gerçekleştirilir.

KAYNAKLAR:

1. Data Mining An Introduction www.spss./.../clem_health carel.htm
2. More on What Data Mining is –and isn't www.spss.com/datamine/what2.htm
3. U. Fayyad et al(1995), "From knowledge Discovery to Data Mining: An Over view," Advances in Knowledge Discovery and Data Mining, U. Fayyad et al (1995), AAAI/MIT Pres
4. http://web.sakarya.edu.tr/~ademiriz/slides/DmHafta2.ppt
5. Dolgun M.Ö, Zor İ. Bir Alışveriş Merkezinden Yapılan Satışlar İçin Sepet Analizi
6. Akpınar, H.2000, Veri Tabanlarında Bilgi keşfi ve Veri Madenciliği
7. Kurt M., Semetay C. Genetik Algoritma ve Uygulama alanları
8. C. Bishop 1996, Neural Networks for Pattern Recognition Oxford Univ. Pres
9. T Mitchell 1997 Machine Learning McGrow_Hill
10. Eker H. Veri Madenciliği veya Bilgi Keşfi
11. Baykal, N. Veri Tabanı ve Veri Madenciliği
12. www.cs.itu.edu.tr/~gunduz/courses/verimaden
13. Dursun, Ö. O., Meteorolojik Verilerin Veri Madenciliği ile Değerlendirilmesi, Fırat Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, (Danışman: Prof. Dr. Asaf Varol), 2005