RİSK ANALİZİNDE VERİ MADENCİLİĞİ UYGULAMALARI

Peral Toktaş, Melek Başak Demirhan
Yeditepe Üniversitesi, Sistem Mühendisliği Bölümü, 34755, İstanbul

Özet: Büyük miktarda veriden anlamlı bilgilerin çıkarılması olarak tanımlanan veri madenciliği, birçok uygulamada olduğu gibi risk analizinde de sıklıkla kullanılan bir yaklaşımdır. Finansal risk analizi kapsamında, doğru ve etkin kredi kararı verebilme, kredi geri ödemesi yapmamaya meyilli müşterileri belirleme, risk derecelendirme, finansal işlemlerde sahtekarlığa yönelik eğilimleri izleme, ekonomik ve finansal yatırımları kararlaştırma, iflas ya da başarısızlık tahmini gibi alanlarda veri madenciliği uygulamalarına rastlamak mümkündür. Bu çalışmada, finansal risk analizi alanında sıkça kullanılan veri madenciliği tekniklerine değinilmiş; uygulama olarak, Türkiye'de faaliyet gösteren ve 34 tanesi 1997-2003 döneminde başarısız olmuş olan 77 ticaret ve kalkınma-yatırım bankasını kapsayan bir finansal başarısızlık tahmin çalışmasına yer verilmiştir. Söz konusu bankalar için başarısızlığı bir yıl öncesinden tahmin edecek ve gerekli önlemlerin önceden alınmasına olanak verecek bir erken uyarı modelinin geliştirilmesi amaçlanmıştır. Model, bir veri madenciliği tekniği olan sinir ağları kullanılarak kurulmuştur. Sermaye yeterliliği, aktif kalitesi, likidite, karlılık ve gelir-gider yapısı, bir bankanın ileride başarılı olup olmayacağını işaret eden göstergeler olarak bulunmuştur.

Anahtar Kelimeler: *Risk Analizi, Veri Madenciliği, Sinir Ağları, Banka Başarısızlık Tahmini*

DATA MINING APPLICATIONS IN RISK ANALYSIS

Abstract: Data mining, which can be defined as extracting or mining knowledge from large amounts of data, is a frequently used approach in risk analysis as in many applications. In the scope of financial risk analysis, it is possible to experience data mining applications in areas such as giving correct and effective credit decisions, determining the customers who have a tendency not to payback the credits, rating risk, following the fraud-oriented tendencies, deciding economic and financial investments, and predicting bankruptcy or failures. This study mentions the data mining techniques frequently used in financial risk analysis; and a financial failure prediction study that covers 77 commercial and development-investment banks in Turkey, 34 of which failed during 1997-2003, is included in the application part. It is aimed to develop an early warning model, which will predict failures one year prior to failure for the named banks and enable taking the necessary precautions in advance. The model is developed using neural networks, which is a data mining technique. The capital adequacy, asset quality, liquidity position, profitability, and income-expenditure structure of a bank at any time are found to be the indicators of its likelihood of failure at a posterior time.

Keywords: *Risk Analysis, Data Mining, Neural Networks, Bank Failure Prediction*

1. Giriş

Veritabanlarında bilgi keşfi, veri içerisindeki geçerli, yeni, yararlı ve sonuç olarak anlaşılabilir örüntülerin çıkarılması sürecidir. Bu süreç, uygulama alanının öğrenilmesi ile başlar ve uygulamanın amaçları doğrultusunda hedef veri seti seçilir. Daha sonra, gürültülü ve tutarsız verilerin çıkarıldığı veri temizleme ve ön işleme basamağı gelir. Gerekli durumlarda veri, madenciliğe uygun bir forma dönüştürülür. Beşinci basamak olan *veri madenciliği*, zeki yöntemler aracılığıyla büyük miktarda veriden anlamlı bilgilerin çıkarılması sürecidir. Daha sonra, çıkarılan örüntüler, içlerinden yararlı olanların belirlenmesi için değerlendirilir. Veritabanlarında bilgi keşfinin son basamağı ise, elde edilen bilginin görüntüleme ve bilgi gösterimi yöntemleri kullanılarak kullanıcıya sunulmasıdır.

Veri madenciliği, birçok alanda olduğu gibi risk analizinde de sıklıkla kullanılan ve gelişmekte olan bir yaklaşımdır. *Risk*, istenmeyen bir sonucun ortaya çıkma olasılığı olup; *risk analizi*, risklerin saptanması, tanımlanması, ölçülmesi ve değerlendirilmesi olarak tanımlanabilir.

2. Risk Analizi ve Veri Madenciliği

Risk analizinde, yöneylem yönetimi, finans, pazarlama, mühendislik, çevre vb. alanlarda veri madenciliği uygulamalarına rastlamak mümkündür. Finansal risk analizi kapsamında ise, doğru ve etkin kredi kararı verebilme, kredi geri ödemesi yapmamaya meyilli müşterileri belirleme, risk derecelendirme, finansal işlemlerde sahtekarlığa yönelik eğilimleri izleme, ekonomik ve finansal yatırımları kararlaştırma,

iflas ya da başarısızlık tahmini gibi alanlarda veri madenciliği yaygın olarak uygulanmaktadır. Bu alanlarda sıkça kullanılmakta olan veri madenciliği teknikleri, Bayes kanı ağları, birliktelik kuralları, bulanık küme yaklaşımları, genetik algoritmalar, k-en yakın komşu algoritması, karar ağaçları, kural çıkarsama, regresyon analizi ve sinir ağları (yapay sinir ağları) şeklinde sıralanabilir.

Bayes kanı ağları, genellikle sınıflama amacıyla kullanılır ve belirsiz muhakeme ile ilgilenir. Bir kanı ağı, iki bileşenden oluşur. Birinci bileşen güdümlü döngüsüz bir grafik olup, düğümler rassal değişkenleri, arklar da olasılıkları temsil eder. İkinci bileşen ise her bir değişken için bir koşullu olasılık tablosundan oluşur. Ağ içerisindeki bir ya da daha fazla düğüm, çıkış düğümü olarak seçilir. Ağ üzerinde sonuç çıkarma algoritmaları uygulanabilir. Sınıflama süreci, tek bir sınıf adı vermekten ziyade, her bir sınıfın olasılığını tahmin eder.

Esas olarak bağ analizi için kullanılan *birliktelik kuralları*, bir veri kümesi içindeki parçalar arasındaki anlamlı ilişkileri araştırır. Pazar sepet analizi, birliktelik kurallarının yaygın bir uygulama alanı olup, birlikte alınan ürünleri araştırarak müşterilerin satınalma alışkanlıklarını inceler. Birliktelik kurallarını uygulamak için farklı algoritmalar mevcuttur. Bu algoritmalardan en yaygın olanı ise Apriori algoritmasıdır.

Bulanık mantığa dayanan *bulanık küme yaklaşımları*, belirsizlik ile ilgilenir. Bulanık mantık, karmaşık, iyi tanımlanmamış ya da matematiksel olarak kolay analiz edilemeyen sistemlerin davranışlarını tanımlamak için hemen hemen doğru olan ve etkin yöntemler sunar. Diğer bir deyişle, bulanık mantık, belirsiz ve kesin olmayan bilginin kullanılması için bir platform oluşturur. Kategoriler arasında kesin bir ayrımdan ziyade, 0-1 arasında bir doğruluk değeri kullanır. Bulanık küme yaklaşımları, veri madenciliği uygulamalarında özellikle sınıflama amacıyla kullanılır.

Genetik algoritmalar, zor matematiksel problemlerin çözümünde en sık kullanılan evrimsel algoritmalar olup, doğal evrim felsefesi üzerine kurulmuşladır. Genetik algoritmalardaki temel fikir, rekabet ve kontrollü değişim süreci içinde bilgi yapıları popülasyonunu muhafaza etmektir. Popülasyondaki her bir yapı somut problem için aday bir çözüm temsil eder ve uygunluk değerlerine göre rekabet sürecinde hangi yapıların yenilerini oluşturmak için yetkinliğinin olduğu belirlenir. Yeni yapılar, çaprazlama ve mutasyon gibi genetik operatörler kullanılarak oluşturulur. Yeni popülasyon oluşturma süreci, popülasyon evrim geçirene kadar devam eder. Genetik algoritmalar, optimizasyon problemlerinin yanı sıra sınıflama amacıyla da kullanılmaktadır. Ayrıca, veri madenciliğinde, diğer algoritmaların uygunluğunu değerlendirmek amacıyla da kullanılabilirler.

Veri madenciliğinde sınıflama amacıyla kullanılan bir diğer teknik ise örnekseme yoluyla öğrenmeye dayanan *k-en yakın komşu algoritması*dır. Bu teknikte tüm örneklemler bir örüntü uzayında saklanır. Algoritma, bilinmeyen bir örneklemin hangi sınıfa dahil olduğunu belirlemek için örüntü uzayını araştırarak bilinmeyen örnekleme en yakın olan *k* örneklemi bulur. Yakınlık Öklid uzaklığı ile tanımlanır. Daha sonra, bilinmeyen örneklem, *k* en yakın komşu içinden en çok benzediği sınıfa atanır. k-en yakın komşu algoritması, aynı zamanda, bilinmeyen örneklem için bir gerçek değerin tahmininde de kullanılabilir.

Karar ağaçları da genellikle sınıflama amacıyla kullanılan bir veri madenciliği tekniğidir. Karar ağacı akış diyagramına benzer bir ağaç yapısında olup, her bir dal bir testin sonucunu, yaprak düğümleri ise sınıfları temsil eder. Bilinmeyen bir örneklemi sınıflamak için örneklemin nitelik değerleri karar ağacı karşısında test edilir. Kökten, o örneklem için sınıf tahminini içeren yaprak düğümüne kadar bir yol izlenir. ID3 ve C4.5 sıklıkla kullanılan karar ağacı algoritmalarıdır. Bu algoritmaların yanı sıra, doğruluğu arttırmak için gürültülü verileri yansıtan dalları çıkaran budama algoritmaları da mevcuttur. Karar ağaçları kolaylıkla sınıflama kurallarına dönüştürülebilir.

Kural çıkarsama, şart-eylem kuralları, karar ağaçları ya da benzer bilgi yapılarını kullanarak uygulamanın hedefleri ile ilgili örüntülerin çıkarılmasını sağlayan hedef-odaklı bir tekniktir. Bu teknikte performans öğesi, örnekleri, karar ağacının dallarından aşağı doğru sıralar ya da şartları örnekle eşleşen ilk kuralı bulur. Sınıflar ya da tahminlerle ilgili bilgi, kuralların eylem kısmında ya da ağacın yapraklarında saklanır. Öğrenme algoritmaları, bilgi yapısı içerisine dahil edilecek nitelikleri seçmek için genellikle istatistiksel bir değerlendirme fonksiyonu kullanır.

Regresyon analizi istatistiksel bir teknik olup, iki ya da daha fazla nicel değişken arasındaki ilişkiyi kullanarak bir değişkenin diğer(ler)i aracılığıyla tahmin edilmesini sağlar. Regresyon analizinin değişik türleri vardır. Doğrusal regresyon analizinde, bir bağımlı değişken, bir ya da daha fazla bağımsız değişkenin doğrusal fonksiyonu olarak modellenir. Eğer bağımlı ve bağımsız değişkenler arasındaki ilişki doğrusal değilse doğrusal olmayan regresyon kullanılabilir. Lojistik regresyon ya da Poisson regresyon gibi genelleştirilmiş doğrusal modeller, kategorik bağımlı değişkenlerin tahmininde kullanılır. Tüm niteliklerin kategorik olduğu log-doğrusal modeller ise kesikli çok-boyutlu olasılık dağılımlarını

yaklaştırır. Regresyon analizinin değişikleri türleri, veri madenciliği uygulamalarında tahmin ve sınıflama amacıyla sıkça kullanılmaktadır.

Son olarak, tahmin ve sınıflama amacıyla kullanılan bir başka veri madenciliği tekniği ise *sinir ağları*dır. Düğüm ve oklardan oluşan bir sinir ağında, düğümler nöronları, oklar ise sinyal akışının yönüyle beraber nöronlar arasındaki bağlantıları temsil eder. Nöronlar, giriş ve çıkış katmanlarında ve eğer varsa gizli katman(lar)da bulunur. Sinir ağları, nöronlar arasındaki sinaptik bağlantıları ayarlamak suretiyle, girdi ile hedef çıktı eşleşecek şekilde eğitilir. Bu şekilde, ağ, veri içinde gömülü olan bilgiyi keşfeder. Sinir ağlarının gücü, şartlara ve çevreye intibak yetenekleri ve kendi kendilerini düzenleme özelliklerinden ileri gelir.

3. Bir Finansal Başarısızlık Tahmin Çalışması

Bu çalışmada, Türkiye'deki bankalar için başarısızlığı bir yıl öncesinden tahmin edecek bir finansal başarısızlık tahmin modeli geliştirilmiştir. Örneklemde, Türkiye'de faaliyet gösteren ve 34 tanesi 1997-2003 döneminde başarısız olmuş olan 77 ticaret ve kalkınma-yatırım bankası bulunmaktadır. Bu dönem içinde kapanan, Tasarruf Mevduatı Sigorta Fonu'na devredilen, bir başka bankaya devredilen ya da bir başka banka bünyesinde birleştirilen bankalar başarısız olarak alınmıştır. Başarılı bankalar, başarısız banka sayısının en yüksek olduğu 2001 yılından seçilmiştir.

Finansal başarısızlık tahmin modelinin kurulmasında, bağımsız değişken olarak bankaların mali tablolarından türetilen 19 finansal oran; bağımlı değişken olarak ise, başarısız bankalar için 0, başarılı bankalar için de 1 değerini alan 0/1 değişkeni kullanılmıştır. Veriler, başarı ya da başarısızlık yılından bir yıl önce yayınlanan finansal tablolardan elde edilmiştir.

Finansal oranlar yapıları gereği birbirleri ile ilişkili olduğu için, ham verinin kullanılması sonuçların yanlı çıkmasına yol açar. Bu nedenle, çoklu bağlantıyı yok etmek için öncelikle *faktör analizi* uygulanmış ve elde edilen faktör puanları kullanılarak model kurulmuştur. Faktör analizi çoklu bağlantıyı yok etmek amacıyla uygulandığından, faktör sayısı bağımsız değişken sayısına eşit alınmıştır.

Modelin kurulmasında, bir veri madenciliği tekniği olan *sinir ağları* kullanılmıştır. Modelde 19 nörondan oluşan bir giriş katmanı, beşer nörondan oluşan iki gizli katman ve iki nörondan oluşan bir çıkış katmanı bulunmaktadır. Sinir ağı modeli, *tek-basamaklı sekant algoritması* kullanılarak eğitilmiştir. Modelin doğru sınıflandırma oranı %100,0 olarak bulunmuştur. *Çapraz-geçerlilik yöntemi* kullanılarak gerçekleştirilen geçerlilik testine göre de modelin geçerlilik oranı %84,5 olarak hesaplanmıştır.

4. Sonuçlar

Veri madenciliği teknikleri, risk analizi alanında da kullanılmakta ve iyi sonuçlar vermektedir. Bu çalışmadaki sinir ağı modelinin tahmin gücü oldukça yüksektir. Modelden elde edilen sonuçlara göre, sermaye yeterliliği, aktif kalitesi, likidite, karlılık ve gelir-gider yapısı, bir bankanın bir yıl sonra başarılı olup olmayacağını işaret eden göstergeler olarak bulunmuştur.

Kaynaklar

Chen, Z., *Data Mining and Uncertain Reasoning: An Integrated Approach*, John Wiley & Sons, Inc., Canada, 2001.

De Jong, K. A., "Evolutionary Computation for Discovery", *Communications of the ACM*, 42(11): 51-53, 1999.

Evans, J. R. ve Olson, D. L., *Introduction to Simulation and Risk Analysis*, Prentice-Hall, Inc., Upper Saddle River, New Jersey, 2002.

Fayyad, U. ve diğerleri, "The KDD Process for Extracting Useful Knowledge from Volumes of Data", *Communications of the ACM*, 39(11): 27-34, 1996.

Fu, L., "Knowledge Discovery Based on Neural Networks", *Communications of the ACM*, 42(11): 47-50, 1999.

Han, J. ve Kamber, M., *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, Inc., San Francisco, California, 2001.

Langley, P. ve Simon, H. A., "Applications of Machine Learning and Rule Induction", *Communications of the ACM*, 38(11): 55-64, 1995.

Neter, J. ve diğerleri, *Applied Linear Statistical Models*, The McGraw-Hill Companies, Inc., Boston, Massachusetts, 1996.