VERİ MADENCİLİĞİNDE YENİ YAKLAŞIMLAR

Bilal Alataş, Erhan Akın
Fırat Üniversitesi, Bilgisayar Mühendisliği Bölümü, 23119, Elazığ

Özet: Bu makale, veri madenciliği (VM) için yapay zeka ve yumuşak hesaplama tekniklerinden bulanık mantık, genetik algoritma ve yapay sinir ağları yaklaşımlarını; ayrıca mühendislik problemlerinin çözümünde de yeni yeni kullanılmaya başlanan yapay bağışıklık sistemleri, karınca koloni optimizasyon algoritmaları, kaos ve destek vektör makineleri yöntemlerini içermektedir. Bu yöntemler, VM teknikleri bakımından karşılaştırılmış, bunların VM'de yeni ve etkili algoritma, teknik ve sistemlerin geliştirilmesi için kullanımı hedeflenmiş ve özellikle dağıtık ve paralel VM'de oldukça etkili olacağı vurgulanmıştır.

Anahtar Kelimeler: *Veri Madenciliği, Yapay Zeka, Yumuşak Hesaplama*

NEW APPROACHES IN DATA MINING

Abstract: This paper includes fuzzy logic, genetic algorithms, and artificial neural networks, techniques of artificial intelligence and soft computing; furthermore, artificial immune systems, ant colony optimization algorithms, chaos, and support vector machines, recently been started to be used in engineering problems, for data mining. These techniques have been compared from the point of view of data mining techniques, been aimed to be used in developing new and efficient algorithms, techniques, and systems; and specially it has been emphasized that they can be quite efficient in distributed and parallel data mining.

Keywords: *Data Mining, Artificial Intelligence, Soft Computing*

1. Giriş

Veri madenciliği (VM), eldeki verilerden üstü kapalı, çok net olmayan, önceden bilinmeyen ancak potansiyel olarak kullanışlı bilginin çıkarılmasıdır. Başka bir deyişle, VM, verilerin içerisindeki desenlerin, ilişkilerin, değişimlerin, düzensizliklerin, kuralların ve istatistiksel olarak önemli olan yapıların yarı otomatik olarak keşfedilmesidir. VM'de keşfedilecek kurallar veritabanın özelliklerine ve kuralların kullanılışına göre farklı tekniklerle bulunur. Bunlardan bazıları sınıflama, kümeleme, birliktelik kuralları, ardışık örüntüler, zaman serisi analizi, tahmin etme, tanımlama ve görselleştirme gibi tekniklerdir.

Gitgide mükemmele yakınlaşma isteği ve doğanın belki de bir gün aynısının yapay yollarla ortaya çıkarılmaya çalışılması yapay zeka ve yumuşak hesaplamanın ortaya çıkarılmasına neden olmuştur. Yapay zeka, insanın düşünme yapısını anlamak ve bunun benzerini ortaya çıkaracak bilgisayar işlemlerini geliştirmeye çalışmak olarak tanımlanır. Yani programlanmış bir bilgisayarın düşünme girişimidir. Daha geniş bir tanıma göre ise, yapay zeka, bilgi edinme, algılama, görme, düşünme ve karar verme gibi insan zekasına özgü kapasitelerle donatılmış bilgisayarlardır. Yumuşak hesaplama, insan aklının verdiği sezgi ve düşüncelerin gerçekleştirilmesini kullanır. Geleneksel hesaplama yöntemleri ile çözülemeyen veya oldukça karmaşık olup tam modelleri tanımlanamayan gerçek dünya problemlerinin birçoğunda insan sezgilerinin ve tecrübelerinin kullanılması çok faydalı olmaktadır. Böylece yumuşak hesaplamanın amacı insan karar verme prosedüründe başarılı, basit, gerçekleştirilebilir ve düşük maliyetli çözümlerle belirsizlikleri ortadan kaldırmaktır. Bu makalede de, VM için yapay zeka ve yumuşak hesaplama temelli yeni ve etkili olabilecek yöntemlerden bahsedilmektedir.

2. Veri Madenciliğinde Yapay Zeka ve Yumuşak Hesaplama Yaklaşımları

Sağladığı en büyük faydası "insana özgü tecrübe ile öğrenme" olayının kolayca modellenebilmesi ve belirsiz kavramların bile matematiksel olarak ifade edilebilmesine olanak tanımak olan bulanık mantığın insan düşünme tarzına yakın olması, uygulanışının matematiksel modele ihtiyaç duymaması ve yazılımın basit olması dolayısıyla ucuza mal olması gibi avantajları olduğunu söyleyebiliriz. Bulanık mantık, VM tekniklerinden kümeleme (Joshi (1998)), sınıflama (Mendes ve diğ. (2001)), zaman serileri (Huarng (2001)) ve ardışık örüntülerde (Chen ve diğ. (2001)) kullanıldığı gibi birliktelik kural keşfinde de kullanılmıştır (Wong ve diğ. (2001)).

Doğal seçim ilkelerine dayanan etkili bir global arama ve optimizasyon yöntemi olan genetik algoritmalar VM'de de etkili şekilde kullanılabilir. Genetik algoritmalar daha çok sınıflama (Alataş ve

Arslan (2003a)), kümeleme (Cole (1998)) ve zaman serisi analizi (Cortez (2001)) gibi VM tekniklerinde kullanılmıştır. Birliktelik kuralları keşfinde GA'nın kullanıldığı iki çalışma vardır. Biri ikili değerler üzerinde çalışan algoritma (Alataş ve Arslan (2003b)), diğeri ise nicel değerler için önerilen algoritmadır (Vázquez ve diğ. (2002)).

Beynin bir işlevi yerine getirme yöntemini modellemek için tasarlanan bir sistem olarak tanımlanabilen yapay sinir ağları (YSA) ile VM çalışması özellikle sınıflama ve kümeleme teknikleri için yoğunlaşmıştır (Zhou ve diğ. (2000), Vlajic ve Card (1998)). Ayrıca zaman serisi analizi (Giles ve diğ. (2001)), birliktelik kural keşfi (Gaber ve diğ. (2000)) gibi VM tekniklerinde kullanılmıştır. Optimize edilmiş farklı ağ yapıları ve parametrelerle daha etkili VM için YSA çalışmaları devam etmektedir.

3. Veri Madenciliğinde Yeni Yaklaşımlar

3.1. Yapay Bağışıklık Sistemi

Bağışıklık sistemindeki etkileşimleri daha iyi anlayabilmek için bağışıklık sisteminin bir modelini oluşturmak ve sistemdeki olayları hesapsal araç olarak kullanabilmek amacıyla ortaya atılmıştır

Bilgisayar bilimcileri, mühendisler, matematikçiler, filozoflar ve diğer araştırmacılar, karmaşıklığı beyne benzer olan bu sistemin özellikle yetenekleri üzerine ilgi duymaktadırlar. Yapay bağışıklık sistemi ile VM tekniklerinden sınıflama ve kümeleme için yeni yeni birkaç çalışma yapılmıştır (Carvalho ve Freitas (2001), Nasraoui ve diğ. (2002a)). Birliktelik kurallarının keşfi için de tek bir çalışma vardır (Nasraoui ve diğ. (2002b)). Ancak diğer teknikler için de çalışmaların olacağı açıktır. Özellikle sistemin sahip olduğu özelliğiyle dağıtık ve paralel VM'de etkili olarak kullanılabilecektir.

3.2. Karınca Koloni Optimizasyonu

Karıncalar, yuvalarından bir gıda kaynağına giden en kısa yolu, herhangi görsel ipucu kullanmadan bulma yetisine sahiptirler. Koloni halinde yaşayan karıncalar yiyecek bulmak için ilk olarak öncü karıncaları tek başına gönderirler. Bu öncüler çevreyi araştırarak uygun yiyecek kaynağını bulmaya çalışır. Öncüler yiyecek bulursa, kolooninin olduğu yere geri dönerken arkalarında özel bir koku izi bırakarak ilerler. Bu iz sayesinde diğer karıncalar da bu yiyecek kaynağını bulabilirler. Araştırmacılar bu arkalarında iz bırakarak ilerleyen öncü karıncaların uyguladığı yöntemi "sanal karıncalar" oluşturarak bilgisayarlarla simüle ettiklerinde çoğu problemin daha kolay çözülebileceğini gösterdiler.

Karınca koloni optimizasyon algoritması VM'de sınıflama tekniğinde yeni yeni kullanılmaya başlanmıştır (Parpinelli ve diğ. (2002)). Elde edilen sonuçlar tatmin edicidir. Daha optimize parametrelerle dağıtık ya da paralel gerçeklemelerle çok daha iyi sonuçların alınacağı kesindir. Diğer tekniklerde de bu yaklaşım kullanılabilir.

3.3. Destek Vektör Makineleri

Destek vektör makineleri yeni bir öğrenme metodudur. Çekirdek tabanlı doğrusal olmayan sınıflandırıcıların sinyal işleme, yapay öğrenme ve VM alanındaki pratik problemlerde iyi sonuçlar verdiği bulunmuştur.

V. Vapnik tarafından önerilen destek vektör makineleri (DVM) ileri yönde beslemeli yeni bir ağ kategorisidir. İstatistiksel öğrenme teorisinde iyi şekilde kurulmuş bir teoriye sahiptir ve sınıflandırma problemlerine yaklaşım için uygundur. Özellikle iki sınıf sınıflandırma probleminde, DVM iki sınıf arasındaki sınırı büyükleyen optimal ayırt etme yüzeyini belirlemekte, yani eğitim kümesi ile ayırt etme yüzeyine en yakın noktaların arasındaki mesafeyi en büyüklemektedir.

Kısaca DVM, doğrusal olmayan bir şekilde ayrılabilen öbekler için optimal hiper-düzlemi bulmaya çalışır. Bu yüzden DVM'nin VM'deki uygulamaları özellikle sınıflama tekniğinde ortaya çıkmıştır. Elde edilen sonuçlar bu yöntemin sınıflama tekniğinde oldukça başarılı olduğunu göstermiştir (Fung ve Mangasarian (2002)).

3.4. Kaos

Kaos kelimesi insanda pek de hoş olmayan çağrışımlar yapar. Karmaşıklık, belirsizlik ve hatta anarşi. Bilimde ise kaos kelimesi belirlenemezlik olarak kabul edilir. Yani günlük yaşamda kullanımı ile bilimde kullanımı oldukça farklıdır.

Kaos teorisi engin uygulama alanına sahip olan bir yaklaşımdır. Her türlü alanda uygulanabilme yeteneğinden dolayı, kaos teorisinin bilim dallarını birbirinden soyutlayan engelleri aştığı söylenebilir. Çok küçük görünen bir nedenin kendisinden çok daha büyük sonuçlara yol açabileceği mantığından hareket eden kaos kuramı, düzensizlik ve karmaşadan çok, bu düzensizlik içerisinde belli bir düzeni, düzenli düzensizliği anlamaya yöneliktir.

Endüstriyel alanlarda çoğu işlemlerin doğrusal olmamasından dolayı kaotik işlemlerin tahmini yapılmaya çalışılmaktadır. VM'de kaos kuramı, kümelemede (Angelini ve diğ.) ve zaman verisinin de kullanıldığı durumlarda birliktelik kural keşfinde kullanılmıştır (Barbara (1999)).

4. Sonuç

Çoğu gerçek dünya problemlerinde etkili sonuçlar veren yapay zeka ve yumuşak hesaplama metotları veri madenciliğinde de kullanılabilir. Yeni veri tiplerinin madenciliği, geniş hacimli ve çok boyutlu VM için bu yöntemler henüz yeni yeni kullanılmaktadır. Özellikle dağıtık ve paralel VM'de bu yöntemlerin oldukça etkili olacağı kesindir.

Kaynaklar

Alataş, B. ve Arslan, A., Mining of interesting prediction rules with uniform two-level genetic algorithm, *International Journal of Computational Intelligence (IJCI)* Proceedings - International XII. Turkish Symposium of Artificial Intelligence and Neural Networks, Volume 1, Number 1, 65-70, 2003a.

Alataş, B. ve Arslan, A., Association rule mining with genetic algorithms, *Mühendislik Bilimleri Genç Araştırmacılar 1. Kongresi MBGAK'2003*, İstanbul, 81-88, 2003b.

Angelini, L. ve diğerleri, Clustering Data by Inhomogeneous Chaotic Map Lattices, *PHYSICAL REVIEW LETTERS*, Vol. 85, No. 3, 554-557.

Barbara D., Chaotic Mining: Knowledge Discovery Using the Fractal Dimension, *SIGMOD DMKD Workshop*, Philadelphia, PA, 1999.

Carvalho, R. ve diğerleri, Discovery of fuzzy sequential patterns for fuzzy partitions in quantitative attributes, *ACS/IEEE International Conference on Computer Systems and Applications AICCSA'01*, 144, 2001.

Carvalho, R. ve Freitas, A.A., An immunological algorithm for discovering small-disjunct rules in data mining. *GECCO-2001*, 401-404. San Francisco, CA, USA, 2001.

Cole, R.M., Clustering with genetic algorithms, http://www.cs.uwa.edu.au/pub/robvis/theses/RowenaCole.ps.gz, 1998.

Cortez, P. ve Neves, M.R.J., Genetic and evolutionary algorithms for time series forecasting. *IEA/AIE*, 393-402, 2001.

Fung, G. ve Mangasarian, O. L., Incremental Support Vector Machine Classification *Second SIAM International Conference on Data Mining*, 2002.

Gaber, K. ve diğerleri, Parallel mining of association rules with a Hopfield type neural network*, 12th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'00)*, 2000.

Giles C.L. ve diğerleri, Noisy Time Series Prediction using a Recurrent Neural Network and Grammatical interface, *Machine Learning*, Vol.44, Num. ½, 161-183.

Huarng, K., 2001, Heuristic models of fuzzy time series for forecasting fuzzy sets and systems 123, 369-386, 1998.

Joshi ve Krishnapuram R., Robust fuzzy clustering methods to support web mining, *Workshop in Data Mining and Knowledge Discovery, SIGMOD*, 15-1--15-8.

Mendes, R.R.F. ve diğerleri, Discovering fuzzy classification rules with genetic programming and co-evolution, *LNAI*, 2168, Berlin, Springer-Verlag, 314-325, 2001.

Nasraoui, O. ve diğerleri, A Novel Artificial Immune System Approach to Robust Data Mining, GECCO'2002, 2002a

Nasraoui O. ve diğerleri, The Fuzzy Artificial Immune System: Motivations, Basic Concepts, and Application to Clustering and Web Profiling. *IEEE International Conference on Fuzzy Systems*, 711-716, Hawaii, HI, 2002b.

Parpinelli, R.S. ve diğerleri, An Ant Colony Algorithm for Classification Rule Discovery. *In: H. Abbass, R. Sarker, C. Newton. (Eds.) Data Mining: a Heuristic Approach*, 191-208. London: Idea Group Publishing, 2002.

Wong, C. ve diğerleri, Mining fuzzy association rules for web access case adaptation, *Workshop Program at the 4th International Conference on Case-Based Reasoning*, 2001.

Vázquez, J.M. ve diğerleri, Discovering numeric association rules via evolutionary algorithm. *PAKDD 2002*, 40-51, 2002.

Vlajic N. ve Card H., An adaptive Neural Network Approach to Hypertext Clustering. University of Manitoba, 1998.

Zhou, H. ve diğerleri, A General Neural Framework for Classification Rule Mining, *International Journal of Computers, Systems and Signals*, 1(2): 154-168, 2000.