GENETİK PROGRAMLAMAYA DAYALI SINIFLANDIRMA YAKLAŞIMI: MEPAR-MINER

Lale ÖZBAKIR
Adil BAYKASOĞLU

ÖZET

Genetik programlama evrimsel gelişime dayalı optimizasyon algoritmaları arasında yer almaktadır. Genetik programlama bireyleri değişken boyuta sahip hiyerarşik ağaç yapısındadır. Genetik programlama metodolojisinin temeli genetik algoritmaya dayanmakla birlikte, kromozom gösterimi ve genetik operatörler açısından farklılık göstermektedir. Genetik programlama değişken boyuttaki ağaç yapısında bireyler üzerinde evrimsel gelişimi sağlamak amacıyla genetik operasyonları gerçekleştirir. Genetik programlamanın avantajlarının yanı sıra kodlanması, ağaç yapısından dolayı uygun olmayan bireylerin ortaya çıkması, ağaç derinliğinin genetik operatörler ile aşırı artması ve buna bağlı olarak çözüm süresinin yüksek olması gibi dezavantajları vardır. Bu dezavantajları aşmak amacıyla doğrusal gösterime sahip değişken uzunlukta bireyler türeten algoritmalar geliştirilmiştir. Çoklu denklem programlama bu algoritmalar arasında yer almaktadır. Bu çalışmada sembolik regresyona yönelik geliştirilen çoklu denklem programlama yaklaşımının, sınıflandırma kuralları türetmek üzere geliştirilmesi ile ortaya konulan MEPAR-Miner yaklaşımı üzerinde durulmuştur. Etkin bir kromozom gösterimi ile genetik programlamanın avantajlarını içerisinde barındıran ve değişken uzunlukta sınıflandırma kuralları türeten bu yaklaşım, genetik operatörlerle uygun olmayan bireylerin türetilmesi, kural boyutunun aşırı büyümesi gibi dezavantajları da ortadan kaldırmıştır. Bu çalışmada geliştirilmiş olan MEPAR-Miner algoritması detaylı açıklanarak, etkinliği karşılaştırmalı olarak ortaya konulmuştur.

Anahtar Kelimeler: Genetik programlama, Sınıflandırma, Çoklu denklem programlama.

ABSTRACT

Genetic programming (GP) is an evolutionary optimization algorithm. Genetic programming individuals are represented as variable sized hierarchical tree structures. Although the main idea behind the GP is similar to genetic algorithm, they have different chromosome representations and genetic operators. GP applies genetic operators on different sized and shaped tree structures in order to obtain evolutionary improvement. Besides the advantages of GP, it has the disadvantages in implementation of genetic operators, generation of infeasible individuals, bloating tree sized and increased CPU times correspondingly. In order to cope with these difficulties, different algorithms which generate variable sized linear chromosomes have been developed. Multi-expression programming which is devoted to symbolic regression appears into these algorithms. In this study, a new Multiple Expression Programming based method named MEPAR-Miner for the derivation of classification rules in data mining applications is addressed. MEPAR-Miner involves advantages of genetic programming due to linear representation of variable sized chromosomes while preventing the generation of infeasible individuals and bloating of tree sizes. In this study MEPAR-Miner algorithm is explained in detail and comparative results are presented in order to analyze the performance of the proposed approach. The experimental results are compared with decision tree and other evolutionary algorithms.

Key Words: Genetic programming, Classification, Multi-expression programming.

1. GİRİŞ

Genetik programlama (GP), zor problemlerin çözümünde kullanılan evrimsel bir çözüm tekniğidir. GP bireyleri genellikle ağaç yapıları ile doğrusal olmayan bir şekilde gösterilir ve işlem görürler. Yakın zamanda GP'nın birçok doğrusal gösterime sahip biçimleri için farklı yapılar önerilmiştir. Bunlardan bazıları gramere dayalı evrim, doğrusal genetik programlama ve gen denklem programlamadır (GEP). Bu farklı yapılardaki gösterimlerin amacı GP'nın performansını artırmak, aynı zamanda da programlanmasını kolaylaştırmaktır. Bu GP yaklaşımlarının ortak özelliği, doğrusal olmayan GP yapısının, doğrusal bireyler olarak ifade edilmesidir.

Genetik programlama, bireyleri programlardan oluşan bir popülasyona genetik algoritma operasyonlarının uygulanmasıdır Uygulama alanı geniş olmakla birlikte sembolik regresyonda iyi sonuçlar verdiği ortaya konulmuştur [1]. Genetik algoritmadan en önemli farklılığı, çözüm dizisinin değişken uzunlukta olma özelliğini taşımasıdır. Bireylerin ağaç yapısındaki gösterimleriyle birlikte çaprazlama ve mutasyon operatörleri genetik algoritmadan farklı olarak uygulamaya geçirilir. Genetik programlamanın çeşitli uygulamaları ile elde edilen sonuçları, Koza [2] ve Langdon [3]'de ayrıntılı bir biçimde yer almaktadır.

Yakın zamanda yapılan çalışmalarla genetik programlama, üretim problemlerinin optimizasyonunda uygulanmaya başlanmıştır. Ancak genetik programlama uygulamaları, kodlanmasının zor olması, ağaç yapısında mutasyon ve çaprazlamadan dolayı uygun olmayan çözümlerin ortaya çıkması yüzünden yaygın bir kullanıma sahip olamamıştır [3]. Ayrıca bu problemler sabit uzunlukta kromozomlarla genetik algoritma tarafından kolayca ifade edilebildiğinden ve çözümün uygunluğu daha kolay kontrol altında tutulabildiğinden genetik algoritma tercih edilmiştir. Bu konuda Dimopoulos ve Zalzala [4]-[6] genetik programlama ile klasik tek makine çizelgeleme problemlerinde uygun sıralama denkleminin belirlenmesine yönelik bir çalışma yapmışlardır. Çalışmada problem yapısına göre toplam gecikmeyi en küçükleyecek şekilde, uygun çizelgeleme kuralını belirlemek üzere farklı problem boyutlarında sıralama kurallarından oluşan dokuz farklı denklem türeterek etkinliklerini karşılaştırmışlardır. Dimopoulos ve Mort [7], hücresel üretim sistemleri için, hiyerarşik yapıda hücrelerin belirlenmesi ve hücrelere atanacak ürün ailelerinin oluşturulması için genetik programlama temelli bir yaklaşım önermişlerdir.

Son yıllarda üretim ve hizmet sektöründe bilgi sisteminin yoğunluğu, teknolojik gelişmelere bağlı olarak tutulan veri hacmini hızla artırmaktadır. Bu büyük miktarlarda veriyi, kullanılabilir ve anlaşılabilir bilgiye dönüştürme süreci ise bilgi keşfi olarak adlandırılmaktadır. Veri madenciliği ise bilgi keşfi sürecindeki aşamalardan birisidir. Veri madenciliği, yapay zekâ, istatistik, veri tabanı sistemleri gibi farklı disiplinleri içerisinde barındırmakta olup, pek çok alanda başarıyla uygulanabilmektedir. Hızlı veri artışından kaynaklanan kullanılmayan veriyi, elverişli bilgiye dönüştürme süreci, son yıllarda araştırmacıların odağı haline gelmiştir. Buna bağlı olarak bilim camiasında veri madenciliğine yönelik algoritmaların ve yöntemlerin hızla geliştirildiği bir süreç yaşanmaktadır. Bu çalışmalar arasında, özellikle veri madenciliği adımlarından en önemlisini teşkil eden sınıflandırma problemleri ve bu problemlere sezgisel yöntemlerle geliştirilen kural tabanlı sistemler ön plana çıkmaktadır. Freitas [8] evrimsel algoritmaların, özellikle genetik algoritma ve genetik programlamanın veri madenciliği ve bilgi keşfindeki kullanımını tartışmıştır. Zhou ve ark. [9] sınıflandırma kurallarının türetilmesine yönelik olarak gen denklem programlama temelli yeni bir yöntem önermişlerdir. DeFalco ve ark. [10] sınıflandırma kurallarının keşfine yönelik bir genetik programlama yaklaşımı önermişlerdir. Tan ve ark. [11] tıbbi veri madenciliğine yönelik sınıflandırma problemleri üzerinde yoğunlaşarak, iki aşamalı melez bir evrimsel yaklaşım geliştirmişlerdir. Bojarczuk ve ark. [12] sınıflandırma problemi için kısıtlı-sözdizim genetik programlama yaklaşımı geliştirmişlerdir. Weinert ve Lopes [13], doğrusal genetik programlama yaklaşımlarından gen denklem programlamaya dayalı bir sınıflandırma kural çıkarımı yöntemi önermişlerdir. Wang ve ark. [14] sınıflandırma kural kümeleri üreten bir parçacık sürü optimizasyonu algoritması geliştirmişlerdir. Literatürde yer alan çalışmalar incelendiğinde özellikle genetik algoritma, genetik programlama gibi evrimsel algoritmalar ile sürü zekasına dayalı optimizasyon yaklaşımlarının sınıflandırma problemleri üzerinde uygulamaları gerçekleştirilmiş ve performansları değerlendirilmiştir.

Bu çalışmada, Baykasoğlu ve Özbakır [15] tarafından önerilen, çoklu-denklem programlamaya dayalı, etkin sınıflandırma kuralları çıkarımında kullanılan MEPAR-miner yaklaşımı ele alınarak açıklanmıştır.

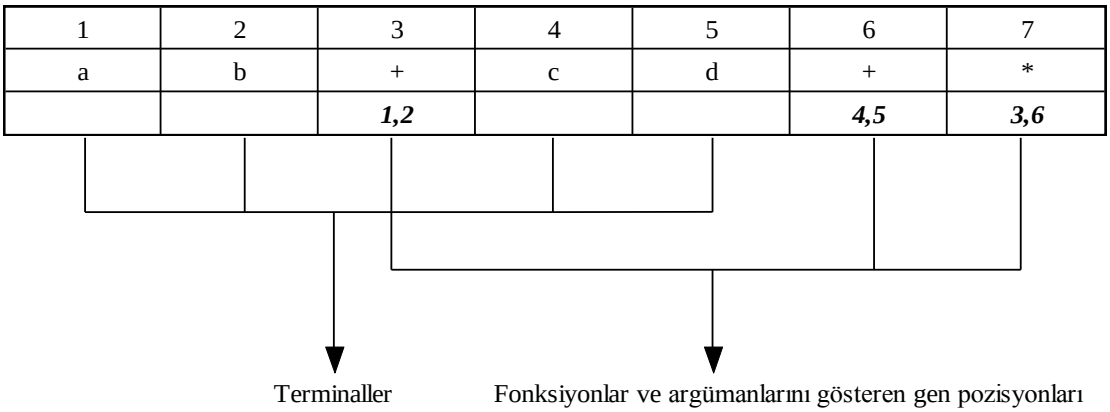
Çalışmanın ikinci bölümünde Oltean ve Dumitrescu [16] tarafından geliştirilen çoklu denklem programlama (ÇDP) yaklaşımı kısaca açıklanmıştır. Üçüncü bölümde ise çoklu denklem programlamaya dayalı geliştirilen MEPAR-Miner algoritması ele alınmıştır. Dördüncü bölüm, deneysel çalışmaları içermekte olup, sonuç değerlendirme beşinci bölümde yer almıştır.

2. ÇOKLU DENKLEM PROGRAMLAMA

Oltean ve Dumitrescu [16] tarafından ortaya konulan, Oltean ve Grosan [17]'ın evrimsel algoritmaların geliştirilmesinde uyguladıkları standart çoklu denklem programlama algoritması, rastgele seçilmiş bireylerden oluşan bir popülasyonla başlar. Mevcut popülasyondaki her birey, probleme bağlı olarak belirlenen uygunluk fonksiyonuna göre değerlendirilir. Belirli sayıda seçilen en iyi bireyler, bir sonraki jenerasyona geçirilir (elitizasyon). Eşleşme havuzu ikili turnuva seçimi ile doldurulur. Daha sonra bu havuzdaki bireyler rastgele eşleştirilerek çaprazlanır. İki ebeveynin çaprazlanması sonucu iki yeni birey elde edilir. Elde edilen bireyler mutasyona uğrar ve bir sonraki jenerasyona girerler. Algoritma belirli sayıda jenerasyon için tekrarlanır.

2.1. Çoklu Denklem Programlama Kromozom Gösterimi

Çoklu denklem programlama genleri, değişik uzunluklardaki alt dizilerle ifade edilir. Kromozomdaki gen sayısı sabittir ve kromozom uzunluğunu ifade eder. Her gen bir terminal ya da fonksiyon sembolü içerir. Fonksiyon içeren bir gen, fonksiyonun argümanlarına işaret eden bir işaretçi taşır. Fonksiyon parametrelerinin, her zaman bu fonksiyonun kendisinin bulunduğu pozisyondan daha küçük pozisyon indislerine sahip olması gerekmektedir. Önerilen gösterim, kromozom deşifre edilirken hiçbir şekilde döngü ortaya çıkarmaz. Bu gösterim tarzına göre kromozomun ilk sembolü mutlaka bir terminal sembolü olmalıdır. Ancak bu şekilde, dizilimi doğru programlar elde edilebilir. Örnek bir kromozom yapısı Şekil 1'deki gibi oluşturulabilir. İlk satırdaki sayılar genlerin kromozom içerisindeki pozisyonlarını ifade etmektedir.



Şekil 1. ÇDP Kromozom Yapısı

Bu ÇDP kromozomu incelenecek olursa 1,2,4,5 pozisyonlu genler birer terminaldir. 3, 6 ve 7 numaralı genler ise bir denklemi ifade eden fonksiyonları ve bu fonksiyonların argümanlarını göstermektedir.

1	2	3	4	5	6	7
a	b	***a+b***	c	d	***c+d***	***(a+b)*(c+d)***

Şekil 2. ÇDP Kromozomunun Denklem Gösterimi

Şekil 2'de görüldüğü gibi ÇDP kromozomları genellikle aynı anda birden fazla denklemi kodlayabilmektedir. Buna karşılık genetik programlama kromozomu tek bir denklemi kodlayabilmektedir. ÇDP kromozomunun kodladığı birden fazla denklemin herhangi birisi kromozomu temsil etmek üzere seçilebilir. Her bir denklemin uygunluk fonksiyonu değeri hesaplanarak en iyi olanı kromozomu temsil edecek şekilde belirlenir. Bazı uygulamalarda bağlantı fonksiyonları (+,*) ile denklemler birbirlerine bağlanabilir.

2.2. Seçim

Standart ÇDP algoritması evrimsel bir algoritma olduğundan, seçim süreci uygunluk fonksiyonu en yüksek olan bireyin bir sonraki jenerasyona aktarılmasını sağlayacak şekilde tasarlanmıştır. Bunun için q-turnuva seçimi yöntemi kullanılmaktadır. Rasgele seçilmiş q adet bireyin içerisinden en iyi (uygunluk fonksiyonu en yüksek olan) birey seçilir. Oltean ve Dumitrescu [16], yaptıkları çalışmada ikili turnuva seçiminin sembolik regresyonda diğer seçim yöntemlerine göre daha iyi sonuç verdiğini ortaya koymuşlardır. İkili turnuva seçiminde mevcut popülasyondan seçilen 2 birey arasından uygunluk fonksiyonu yüksek olan birey alınır.

2.3. Genetik Operatörler

ÇDP algoritması içerisinde kullanılan arama operatörleri çaprazlama ve mutasyondur. Uygulanan arama operatörleri kromozom yapısını koruyacak ve uygun olmayan birey türetmeyecek şekilde tasarlanmıştır.

Çaprazlama: Standart ÇDP algoritmasında üç farklı tip çaprazlama tekniği uygulanmıştır. Probleme bağlı olarak bu tekniklerden herhangi birisi veya tamamı kullanılabilmektedir.

Bu teknikler:

Tek-nokta çaprazlama: Tek nokta çaprazlamada, iki ebeveyn bireyden rastgele bir nokta seçilir ve bireylerin belirlenen noktadan sonraki genleri birbirleri ile yer değiştirilerek yeni bireyler elde edilir.

İki-nokta çaprazlama: İki-nokta çaprazlama da tek-nokta çaprazlamaya benzemekle birlikte, aralarındaki fark, tek nokta yerine iki noktanın rastgele seçilerek çaprazlamanın gerçekleştirilmesidir.

Düzenli çaprazlamadır: Düzgün çaprazlamada çocuklar her iki bireyden rastgele genlerin seçilmesi ile oluşturulur.

Mutasyon: Kromozom içerisindeki her bir gen mutasyona tabi tutulabilir. Mutasyon operatörünü uygularken, mutasyon olasılığı (p_m) göz önüne alınmaktadır. Standart ÇDP algoritmasında iki farklı tip mutasyon uygulanmaktadır;

- Standart mutasyon
- Düzgün mutasyon

Standart Mutasyon*:* Mutasyonla kromozom içerisindeki bazı genler değişikliğe uğrar. Burada dikkat edilecek nokta ilk genin her zaman terminal olmasının sağlanmasıdır. Diğer genler için herhangi bir kısıtlama yoktur. Eğer seçilen gen bir terminal sembolse, yine bir terminal sembole veya bir fonksiyon sembole dönüştürülebilir. Eğer seçilen gen bir fonksiyonsa, bir terminal sembole veya başka bir fonksiyona değiştirilebilir ve argümanları da yeni fonksiyona göre belirlenir.

Düzgün Mutasyon: Fonksiyon sembollerinin değiştirilmesinde düzgün mutasyon kullanılabilir. Düzgün mutasyon, gen içerisinde yer alan her sembolü (fonksiyon sembolü veya fonksiyonun argüman pozisyonları) belirlenmiş bir olasılıkla (p_{sm}) değiştirir.

Eğer gen içerisindeki fonksiyon iki argümanlı bir fonksiyon ise, *psm*=0.33 olarak tanımlandığında gen içerisindeki her bir pozisyona eşit değerde değişiklik oranı düşmektedir.

3. MEPAR-MINER ALGORİTMASI

Oltean ve Dumitrescu [16] tarafından ortaya konulan ÇDP yaklaşımı, sınıflandırma kuralı türetmek üzere değiştirilmiş ve yazılımı gerçekleştirilmiştir. ÇDP kromozom gösteriminin temel yapısı kullanılmış ancak fonksiyon ve terminal kümesindeki değişikliklerle mantıksal ifadelerin elde edilmesi sağlanmıştır. Bir mantıksal ifade sınıfı sınıflandırma kuralı şeklinde gösterilebilir ve *n* sınıflı bir problem için her bir sınıfa ait bir yada daha fazla mantıksal ifade bir araya getirilebilir. Aşağıdaki ifade bir kural kümesini göstermektedir.

EĞER Şart 1 İSE Sınıf$_1$
DEĞİLSE EĞER Şart 2 İSE Sınıf$_2$
……
DEĞİLSE Sınıf$_{varsayılan}$

Kural kümesinin değerlendirilmesine ilk kuraldan başlanır ve örneği sağlayan kural bulunana kadar devam edilir. Örnek hiçbir kural tarafından sağlanmamışsa, o örneğin sınıfı varsayılan sınıf olarak belirlenir. Varsayılan sınıf genel olarak örnekler içerisinde en sık rastlanan sınıf olarak atanır.

3.1. Fonksiyon ve Terminal Kümesi

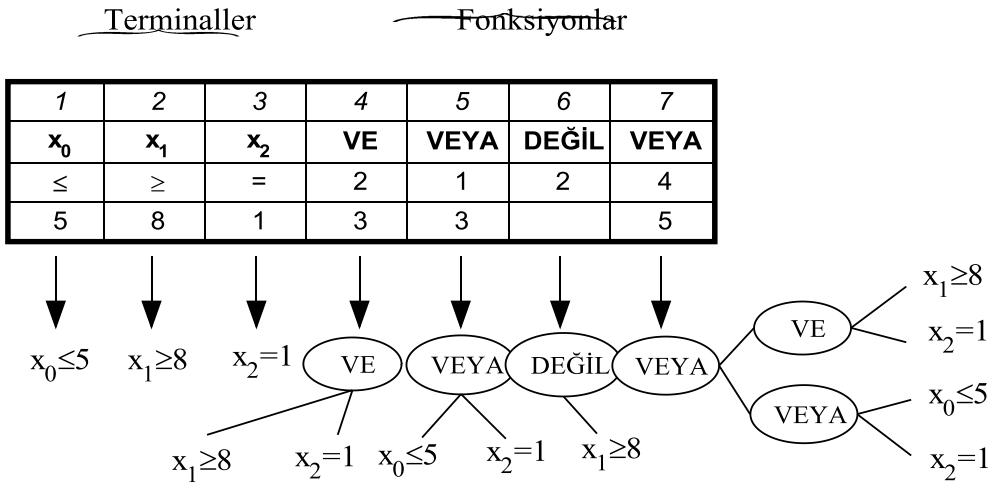
Bir sınıflandırma problemi için fonksiyon ve terminal kümesinin belirlenmesi gerekmektedir. ÇDP algoritması sembolik regresyon denklemleri üretmek amacıyla geliştirildiğinden fonksiyon kümesi matematiksel operatörlerden oluşur. MEPAR-Miner algoritması sınıflandırma için mantıksal ifadeler üretmek üzere tasarlandığından fonksiyon kümesinde (*F*) mantıksal operatörler yer almaktadır. Terminal kümesi ise mantıksal ifadelerdeki özellik-değer ilişkisini yansıtacak şekilde düzenlenmiştir. Dolayısıyla terminal kümesi, özellik-ilişkisel operatör-değer üçlüsünden oluşan elemanlar içermektedir. Değer, o özelliğe ait değer aralığından belirlenmektedir. İlişkisel operatör ise özelliğin sürekli veya kategorik olma durumuna göre belirlenmektedir. Kategorik özellikler için "=", sürekli değer alan özellikler için "≤, ≥" ilişkisel operatörleri kullanılmaktadır. Her bir gen, kendi içerisinde özelliği barındırırken, bu özelliğe ilişkin ilişkisel operatör ve değeri ifade eden işaretleyicileri de içermektedir. MEPAR-Miner algoritmasında kullanılan fonksiyon ve terminal kümeleri Tablo 1'de yer almaktadır.

Tablo 1. Fonksiyon ve terminal kümeleri

X_i	i. özellik
İlişkisel operatör	Özelliğin türü
=	Kategorik özellikler
≤, ≥	Sürekli özellikler
V_{xi}	i. özelliğin değer alanı
Terminal Kümesi	{x_0 – İO - V_{xo} , x_1 – İO - V_{x1} , .., x_n – İO - V_{xn}}
Function Set	{VE, VEYA, DEĞİL}

Aynı durumu ifade eden anlamsız mantıksal ifadelerin ortaya çıkmasını engellemek amacıyla, fonksiyon kümesinden seçilmiş olan elemanlardan oluşan her bir genin argümanlarının birbirinden farklı indisleri işaret etmesi gerekmektedir. Aşağıdaki fonksiyon, terminal ve ilişkisel operatörlerle oluşturulmuş bir örnek kromozom yapısı Şekil 3'te gösterilmektedir.

F →{VE, VEYA, DEĞİL}, İO → {≤, ≥, =}, T→ { x_0, x_1, x_2}



Şekil 3. Sınıflandırma Kuralı İçin Örnek Kromozom Yapısı

Şekil 3.'de yer alan kromozomun 7. Geni aşağıdaki sınıflandırma kuralını ifade etmektedir.

EĞER (($x_1 \geq 8$) VE ($x_2=1$)) VEYA (($x_0 \leq 5$) VEYA ($x_2=1$)) İSE $Sınıf_i$

Bu kuralın sınıfının belirlenmesi aşamasında sınıfın önceden sabitlenmiş olması, kural kalitesini etkileyeceğinden, her bir sınıf için aynı kural değerlendirilmiştir. Elde edilen kuralın doğruluğunu en yüksek uygunluk değeri ile ifade eden sınıf, kuralın sınıfı olarak atanmıştır. Uygunluk değerlendirmesi sonucu, kromozom içerisinde yer alan her gen kendi uygunluk değerine ve sınıfına sahip olmaktadır. Bir kromozomun uygunluk değeri ise, o kromozom içerisinde yer alan genlerden en yüksek uygunluk değerine sahip olan genin uygunluk değeri olarak belirlenmektedir. Aynı zamanda bu gen, kromozomu ifade eden en iyi kural haline gelmektedir. Mevcut kodlama biçimi, Michigan kodlama yaklaşımı olarak belirlenmiştir [8]. Bu yapıda, popülasyon içerisindeki her bir kromozom diğerlerinden bağımsız tek bir kuralı ifade etmektedir. Dolayısıyla popülasyon farklı sınıflar için kurallar içeren bireylerden oluşmaktadır. Her bir sınıfı en yüksek doğrulukla ifade eden kurallar seçilerek kural kümesi oluşturulmaktadır. Kural kümesinin boyutu örnek veri kümesindeki sınıf sayısına ilave olarak kalan örnekleri varsayılan sınıfa atamaya ilişkin bir ifade de içermektedir.

3.2. Uygunluk Fonksiyonu

Bir eğitim örneğini sınıflandırmak için kural değerlendirildiğinde dört farklı durumdan birisi ile karşılaşılır [18]: doğru pozitif, doğru negatif, yanlış pozitif ve yanlış negatif. Doğru pozitif ve doğru negatif, uygun sınıflandırmayı, yanlış pozitif ve yanlış negatif de hatalı sınıflandırmayı ifade etmektedir.

Doğru pozitif (TP) : Kural, sınıfı evet olarak belirler, örneğin sınıfı gerçekte evet'tir.
Doğru negatif (TN) : Kural, sınıfı hayır olarak belirler, örneğin sınıfı gerçekte hayır'dır.
Yanlış pozitif (FP) : Kural, sınıfı evet olarak belirler, örneğin sınıfı gerçekte hayır'dır.
Yanlış negatif (FN) : Kural, sınıfı hayır olarak belirler, örneğin sınıfı gerçekte evet'tir.

Duyarlılılk (Se) tüm veri kümesi içerisindeki "evet" sınıfı örneklerinden doğru sınıflandırılanların oranını hesaplar.

$$S_e = TP/(TP+FN) \quad (1)$$

Özgüllük (Sp) tüm veri kümesi içerisindeki "hayır" sınıfı örneklerinden doğru sınıflandırılanların oranını hesaplar.

$$S_p = TN/(TN + FP) \quad (2)$$

Sınıflandırma kurallarının etkinliğini ölçen bu iki kavram ile uygunluk fonksiyonu aşağıdaki gibi tanımlanmıştır [19];

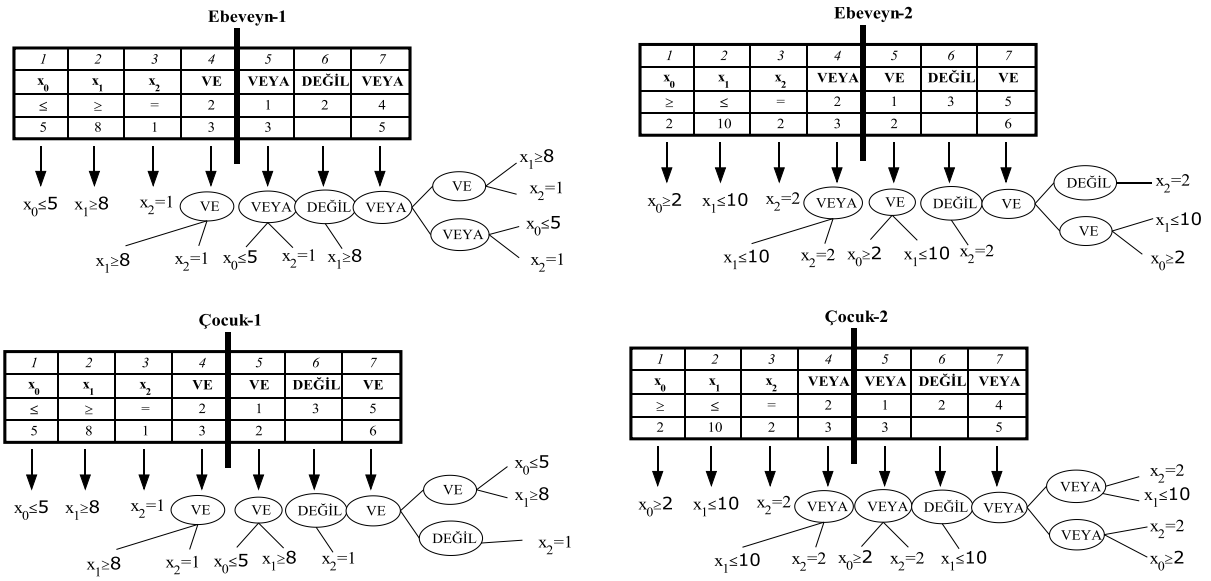
$$Uygunluk = S_e \times S_p \quad (3)$$

Uygunluk fonksiyonunun değeri 0-1 aralığında yer almaktadır. Bütün örnekler doğru sınıflandırıldığında uygunluk değeri 1'dir.

3.3. Genetik Operatörler

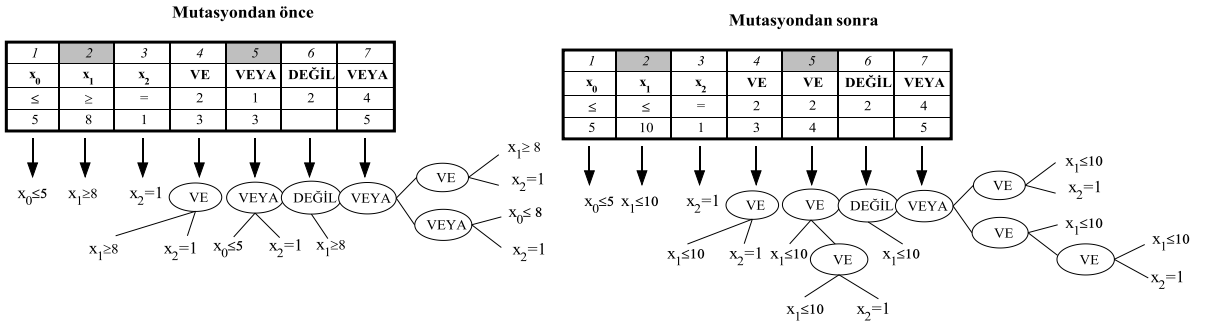
MEPAR-miner algoritmasında kullanılan genetik operatörler çaprazlama ve mutasyondur. Her iki operatör de kromozomun yapısını bozmamaktadır. Bu operatörlerin uygulanması sonucunda ortaya çıkan yeni bireyler anlamlı mantıksal ifadeler içermektedirler. Genetik operatörler uygulanmadan önce popülasyondaki en iyi birey sonraki jenerasyona aktarılır. Seçim sürecinde eşleşme havuzu, ikili turnuva seçimi ile belirlenen bireylerden oluşturulur. Popülasyondan rastgele iki birey seçilir ve bunlardan uygunluk değeri daha yüksek olan eşleşme havuzuna aktarılır.

Çaprazlama: Çaprazlama operatörü, belirli bir çaprazlama olasılığına göre (p_c) iki ebeveyn bireyin eşleşme havuzundan seçilmesi ve bu bireylerden yeni iki birey elde edilmesini sağlar. Bu çalışmada tek nokta çaprazlama operatörü uygulanmıştır. Rastgele bir çaprazlama noktası seçilerek, iki ebeveyn bireyin çaprazlama noktasından önceki ve sonraki genleri çaprazlanarak yeni bireyler elde edilir (Şekil 4).



Şekil 4. MEPAR-Miner Çaprazlama

Mutasyon: Mutasyon operatörü kromozomdaki terminal ve fonksiyon genlerine, önceden belirlenmiş mutasyon olasılığına göre (p_m) uygulanır. Kromozomda rastgele mutasyon noktaları belirlenerek bu noktalardaki genlerin terminal veya fonksiyon olma durumuna göre farklı şekillerde uygulanır. Eğer seçilen gen terminal ise bu gende yer alan ilişkisel operatör ve değer değiştirilir. Eğer seçilen gen fonksiyon ise, yeni bir fonksiyon ve bu fonksiyona ait işaretçiler seçilerek bu gen değiştirilir. Seçilen işaretçiler mevcut genden daha düşük genlerin indislerini içermelidir. Şekil 5'te 2. ve 5. noktalardaki terminal ve fonksiyon genlerinin mutasyonu gösterilmektedir.



Şekil 5. MEPAR-Miner Mutasyon

4. DENEYSEL ÇALIŞMA

MEPAR-miner algoritmasının performansını analiz etmek için 6 farklı veri kümesi UCI makine öğrenme veritabanından seçilmiştir (http://www.ics.uci.edu/~mlearn/MLRepository.html). Veri kümelerinin temel özellikleri olan örnek sayıları, kategorik ve sürekli özellik sayıları ile sınıf sayıları Tablo 2'de özetlenmiştir.

Tablo 2. Veri Kümelerinin Özellikleri

Veri Kümesi	Örnek sayısı	Kategorik özellik sayısı	Sürekli özellik sayısı	Sınıf sayısı
WBC	683	-	9	2
LBC	282	9	-	2
Tic-Tac-Toe	958	9	-	2
CRX	690	9	6	2
Nursery	12960	8	-	5
Adult	45222	8	6	2

MEPAR-miner algoritmasının performansı, kullanılan veri kümeleri için farklı çalışmalarda ortaya konulan sonuçlar ile karşılaştırılmıştır.

1. Tan ve ark. [11] tarafından geliştirilen iki aşamalı hibrid bir evrimsel algoritma,
2. Carvalho ve Freitas [20,21] tarafından önerilen hibrid karar ağacı-genetik algoritma,
3. Parpinelli ve ark. [19] tarafından geliştirilen karınca koloni optimizasyonuna dayalı sınıflandırma yaklaşımı
4. Bojarczuk ve ark. [12]'nın önerdiği genetik programlama yaklaşımı,
5. Weinert ve Lopes [13] tarafından ortaya konulan gen denklem programlamaya dayalı sınıflandırma yaklaşımı,
6. Wang ve ark.[14]'nın geliştirdiği parçacık sürü optimizasyonuna dayalı sınıflandırma yaklaşımı bu çalışmada sonuç değerlendirme ve karşılaştırma amacıyla kullanılan çalışmalardır.

Sonuç karşılaştırmaları, eğitim verisi kullanılarak elde edilen sınıflandırma kurallarının test verisi üzerindeki performansı üzerinden gerçekleştirilmiştir. Test verisindeki tahminleme doğruluğu denklem 4'teki formülasyon ile hesaplanmaktadır.

$$Dogruluk = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

Adult veri kümesi büyük boyutlu veri olmasından dolayı eğitim ve test verisi olarak ikiye bölünmüştür. Diğer veri kümeleri ise çapraz doğrulama yöntemi kullanılarak 10 farklı eğitim ve test veri grubu ile analiz edilmiştir. 10'lu çapraz doğrulamada, veri kümesi 10 alt gruba bölünerek her seferinde 1 alt grup test, kalan 9 alt grup eğitim amaçlı kullanılmıştır. Böylece her bir veri kümesi için 10 farklı çalıştırma gerçekleştirilmiştir. Tahminleme doğruluklarının ortalaması ve standart sapmaları hesaplanmıştır.

4.1. MEPAR-Miner Parametreleri

Jenerasyon Sayısı: MEPAR-miner algoritmasının bir sınıflandırma kuralı türetmek için tekrarlı çalışma sayısını ifade etmektedir. Belirlenen jenerasyon sayısına ulaşıldığında en iyi birey, bu bireye ait sınıf ile birlikte bir sınıflandırma kuralını teşkil eder. Kalan sınıflar için algoritma aynı sayıda tekrarlı çalıştırılır.

Popülasyon Sayısı: Her bir jenerasyonda değerlendirilen birey sayısını ifade eder. Başlangıç popülasyonu rastgele oluşturulur.

Kromozom uzunluğu: Bir kromozomdaki gen sayısını ifade eder. MEPAR-miner algoritmasında gen sayısı aynı zamanda o kromozomda yer alan farklı kural sayısını da ifade eder.

p_c : Çaprazlama olasılığı
p_m: Mutasyon olasılığı

Tablo 3. MEPAR-Miner Parametre Değerleri

Parametre	Değer
Jenerasyon Sayısı	250
Popülasyon Sayısı	250
Kromozom uzunluğu	100
p_c	0.9
p_m	0.2

Algoritmanın işleyişi esnasında belirlenen parametre değerleri Tablo 3'te verilmiştir. MEPAR-miner algoritması C++ programlama dilinde kodlanmış ve her bir veri kümesi için 10 farklı çalıştırma sonucunda elde edilen ortalama tahminleme doğruluğu, minimum, maksimum değerleri ve standart sapması Tablo 4'te verilmiştir.

Tablo 4. MEPAR-Miner Tahminleme Doğruluğu

	Data Sets					
Tahminleme Doğruluğu(%)	CRX	Nursery	Adult	LBC	Tic-Tac-Toe	WBC
Maks	100	98.76	96.89	100	96.87	100
Ortalama	**96.96**	**95.83**	**95.27**	**90.63**	**94.47**	**99.41**
Min	91.30	93.20	91.82	85.72	93.62	98.53
Standart Sapma	2.50	1.80	1.55	4.48	1.31	0.76

MEPAR-miner algoritmasından elde edilen sonuçlar literatürde yer alan çalışmalarda sunulan sonuçlar ile tahminleme doğruluğu açısından karşılaştırılmış ve Tablo 5'te özetlenmiştir.

Tablo 5. Karşılaştırma Sonuçları

Veri Kümesi	MEPAR-miner	Ant-Miner[1]	GP[2]	BGP[2]	EvoC[3]	PSO[5]	GEPCLASS[6]
LBC	**90.63**±4.48	75.28±2.24	71.8±4.68	63.85±5.67	-	77.58±0.27	68.5±12.73
WBC	**99.41**±0.76	96.04±0.93	93.5±0.79	89.3±4.37	95.26±0.25	97.95±0.68	93.8±2.89
Tic-tac-toe	**94.47**±1.31	73.04±2.53	-	-	-	98.84±0.24	-
Veri Kümesi	MEPAR-miner	C4.5[4]	Double C4.5[4]	C4.5/GA[4]			
CRX	**96.96**±2.50	91.79±2.1	90.78±1.2	91.66±1.8	-	-	-
Nursery	**95.83**±1.80	95.4±1.2	97.23±1.0	96.77±0.7	-	-	-
Adult	**95.27**±1.55	78.62±0.5	76.06±0.5	80.04±0.1	-	-	-

1. Parpinelli ve ark. [19]
2. Bojarczuk ve ark. [12]
3. Tan ve ark. [11]
4. Carvalho ve Freitas [21]
5. Wang ve ark.[14]
6. Weinert ve Lopes [13]

Tablo 5'te MEPAR-miner ve diğer algoritmaların ortalama tahminleme doğrulukları ile standart sapmaları yer almaktadır. MEPAR-miner algoritması LBC, WBC, CRX ve Adult veri kümelerinde diğer çalışmalara göre daha yüksek tahminleme doğruluğuna ulaşmıştır. Tic-tac-toe veri kümesinde PSO algoritması, Nursery veri kümesinde ise Carvalho ve Freitas [21] tarafından önerilen hibrid C4.5/GA yaklaşımı MEPAR-miner'dan daha iyi sonuç vermiştir. Tablo 6 her bir veri kümesi için elde edilen en yüksek uygunluk değerine sahip kural kümelerini içermektedir. MEPAR-miner algoritması bu sonuçlara her sınıf için bir kural içeren kural kümesi ile ulaşmıştır. Dolayısıyla her sınıf için çoklu kural üreten algoritmalara göre kural kümesinde yer alan kural sayısı ve tahminleme doğruluğu açısından oldukça etkin bir sınıflandırma algoritması olarak ön plana çıkmaktadır. Algoritmanın tahminleme doğruluğu her bir sınıfa ait kural sayısındaki artışla doğru orantılı olarak yükselecektir. Ancak kural kümesinin basit ve anlaşılır olması uygulama açısından her zaman tercih edilir bir durumdur. Genel olarak değerlendirilecek olursa MEPAR-miner algoritması ele alınan bütün veri kümelerinde test verisi üzerinde oldukça yüksek doğrulukta tahminleme yapabilmektedir.

SONUÇ

Bu çalışmada sınıflandırma problemleri için çoklu denklem programlamaya dayalı bir kural çıkarımı yöntemi olan MEPAR-miner ele alınmıştır. Çoklu denklem programlama sembolik regresyon için geliştirilmiş bir doğrusal genetik programlama yaklaşımıdır. Bu yöntemin kromozom gösterimi, terminal ve fonksiyon tanımlamaları sınıflandırma problemleri için EĞER-İSE kuralları türetecek şekilde yeniden yapılandırılmıştır. MEPAR-miner algoritmasında kromozomlar doğrusal gösterime sahip olmasına rağmen değişik uzunlukta kurallar elde edebilecek şekilde tasarlanmıştır. Bu özelliğinden dolayı genetik programlamanın avantajlarını içermekle birlikte doğrusal kromozom yapısından dolayı kolay yazılımı ve genetik operatörlerle uygun bireyler elde edebilmesi de söz konusudur. Bu çalışmada MEPAR-miner ile 6 farklı veri kümesi üzerinde analizler gerçekleştirilmiştir. Bu veri kümeleri sürekli ve kesikli özellikleri içermesi, ikili yada çoklu sınıfa sahip olmaları, özellikle Nursery ve Adult veri kümelerinin yüksek sayıda örnekten oluşması, algoritmanın farklı yapıda veri kümeleri üzerindeki performansını değerlendirebilmek amacıyla belirlenmiştir. Oluşturulan kural kümesinin her bir sınıfa ait örnekler için tek bir kural içermesi ve bu kural kümesinin yüksek tahminleme doğruluğuna sahip olması, MEPAR-miner algoritmasının sınıflandırma kuralları geliştiren algoritmalar içerisinde önemli bir yere sahip olduğunu ortaya koymaktadır. Genetik algoritma, parçacık sürü optimizasyonu, karınca koloni optimizasyonu ve gen denklem programlamaya dayalı genetik programlama yöntemleri ile elde edilen sonuçlar karşılaştırılmalı olarak ele alınmıştır.

Tablo 6. Veri Kümeleri İçin Bulunan En İyi Kural Kümeleri ve Uygunluk Değerleri

Veri Kümeleri	Se	Sp	Uygunluk Fonksiyonu	Kural Kümesi
CRX	*0.94* *0.81*	*0.80* *0.93*	*0.75* *0.75*	EĞER (((((x2>=2.00))ve((x0=2.00)))ve((x6=1.00)))ve((x11=1.00)))veya(((((x10<=3.00))veya(((x13<=2.00)) veya((x5=4.00))))ve((x4=3.00)))veya(((x8=1.00))veya((x12=3.00)))) İSE Sınıf 1 DEĞİLSE EĞER DEĞİL((((x4=3.00))ve((x10<=1.00)))veya((x8=1.00)))İSE Sınıf 0 DEĞİLSE Sınıf 0
Nursery	*1.00* *1.00* *1.00* *1.00* *0.93*	*1.00* *0.99* *0.79* *0.50* *0.65*	*1.00* *0.99* *0.79* *0.50* *0.60*	EĞER DEĞİL(DEĞİL((x7=3.00)))İSE Sınıf 0 DEĞİLSE EĞER ((((x7=1.00))ve((x5=1.00)))ve((x1=1.00)))ve(DEĞİL(DEĞİL((x4=1.00))))İSE Sınıf 1 DEĞİLSE EĞER (((x4=3.00))veya(DEĞİL(DEĞİL(((x4=3.00))veya(DEĞİL((x0=3.00))))))) ve(((x7=1.00))ve((DEĞİL((x1=5.00)))veya((x4=3.00))))İSE Sınıf 2 DEĞİLSE EĞER DEĞİL((x7=3.00))İSE Sınıf 3 DEĞİLSE EĞER DEĞİL((DEĞİL(DEĞİL(((x7=3.00))))veya((x1=2.00)))İSE Sınıf 4 DEĞİLSE Sınıf 0
Adult	*0.87* *0.66*	*0.65* *0.89*	*0.57* *0.59*	EĞER (((x6=6.00))veya((x5=1.00)))ve((x4>=2.00))İSE Sınıf 1 DEĞİLSE EĞER DEĞİL((((((x11<=10.00))veya((x1=8.00)))ve((((x10>=16.00))veya((x5=1.00)))ve((x13=15.00)))) veya(((x10>=16.00))veya((x5=1.00))))veya(((x10>=16.00))veya((x5=1.00))))İSE Sınıf 0 DEĞİLSE Sınıf 0
LBC	*0.67* *0.72*	*0.74* *0.66*	*0.50* *0.47*	EĞER (((x7=3.00))veya(((x5=3.00))veya((x2=7.00))))ve(((x0=6.00))veya(((((x1=3.00))veya((x8=1.00)))ve((((x7=3.00)) ve(((x4=1.00))veya(DEĞİL((x5=3.00)))))veya(((x5=3.00))veya((x2=7.00)))))veya((x2=7.00))))İSE Sınıf 11 DEĞİLSE EĞER ((DEĞİL((x8=1.00)))veya((x2=4.00)))ve(DEĞİL((x5=3.00)))İSE Sınıf 0 DEĞİLSE Sınıf 1
Tic-Tac -Toe	*0.70* *0.80*	*0.70* *0.66*	*0.49* *0.53*	EĞER DEĞİL((((x2=1.00))ve((x0=1.00)))veya((x4=1.00)))İSE Sınıf 1 DEĞİLSE EĞER (((x8=1.00))ve((x2=1.00)))veya((((x6=1.00))ve((x8=1.00)))veya((x4=1.00)))İSE Sınıf 0 DEĞİLSE Sınıf 0
WBC	*0.97* *0.98*	*0.96* *0.92*	*0.93* *0.90*	EĞER (DEĞİL(((DEĞİL(((x1<=6.00))ve(((x0<=6.00))ve(((x4<=8.00))ve(DEĞİL((x5>=7.00)))))))veya(DEĞİL((x2<=4.00))))) veya((((x0<=6.00))ve(((x4<=8.00))ve(DEĞİL((x5>=7.00)))))ve(DEĞİL(((x3>=3.00))ve((x4<=8.00)))))İSE Sınıf 0 DEĞİLSE EĞER DEĞİL((((x0<=2.00))veya((x3<=9.00)))ve(((((x0<=2.00))veya((x5<=2.00))) veya(DEĞİL(((x7<=4.00))ve((x8<=5.00)))))ve((((x7<=4.00))ve((x8<=5.00)))ve((((x0<=2.00))veya((x3<=9.00))) veya(((x0<=2.00))veya((x5<=2.00)))))))İSE Sınıf 1 DEĞİLSE Sınıf 1

KAYNAKLAR

[1] SETTE, S., BOULLART, L., "Genetic Programming: Principles and Applications", Engineering Applications of Artificial Intelligence, vol.14, 727-736, 2001.

[2] KOZA, J.R., "Evolution of Emergent Cooperative Behavior Using Genetic Programming", in: Paton Ray (Ed.), Computing with Biological Metaphors, Chapman&Hall, 280-297, 1994.

[3] LANGDON, W.B., "Genetic Programming and Data Structures: Genetic Programming+Data Structures=Automatic Programming", The Kluwer International Series In Engineering and Computer Science, 1998.

[4] DIMOPOULOS, C., ZALZALA, A.M.S., "Evolving Scheduling Policies through a Genetic Programming", in: Proceedings of the Genetic and Evolutionary Computation Conference, GECCO-99, Florida, USA, 1231-1232, 1999a.

[5] DIMOPOULOS, C., ZALZALA, A.M.S., "A Genetic Programming Heuristic for the One Machine Total Tardiness Problem", in: Proceedings of the Congress on Evolutionary Computation, CEC'99, New York: IEEE Press, vol. 3, 2207-2214, 1999b.

[6] DIMOPOULOS, C., ZALZALA, A.M.S., "Investigating the Use of Genetic Programming for A Classic One – Machine Scheduling Problem", Advances in Engineering Software, vol. 32, 489-498, 2001.

[7] DIMOPOULOS, C., MORT, N., "A Hierarchical Clustering Methodology Based on Genetic Programming for the Solution of Simple Cell-formation Problems", International Journal of Production Research, vol. 39(1), 1-19, 2001.

[8] FREITAS, A.A., "A Survey of Evolutionary Algorithms for Data Mining and Knowledge Discovery", in: A. Ghosh, S. Tsutsui (Ed.), Advances in Evolutionary Computation, Berlin, Springer, 2002.

[9] ZHOU, C., XIAO, W., TIRPAK, T.M., NELSON, P.C., "Discovery of Classification Rules by Using Gene Expression Programming", in: Proceedings of the 2002 Int. Conf. on Artificial Intelligence, IC-AI'2002, Las Vegas, 2002.

[10] DE FALCO, I., CIOPPA, A.D., TARANTINO, E., "Discovering Interesting Classification Rules with Genetic Programming", Applied Soft Computing, vol. 1, 257-269, 2002.

[11] TAN, K.C., YU, Q., HENG, C.M., LEE, T.H., "Evolutionary Computing for Knowledge Discovery in Medical Diagnosis", Artificial Intelligence in Medicine, vol. 27, 129-154, 2003.

[12] BOJARCZUK, C.C., LOPES, H.S., FREITAS, A.A., MICHALKIEWICZ, E.L., "A Constrained-Syntax Genetic Programming System for Discovering Classification Rules: Application to Medical Data Sets", Artificial Intelligence in Medicine, vol. 30, 27-48, 2004.

[13] WEINERT, W.R., LOPES, H.S., "GEPCLASS: A Classification Rule Discovery Tool Using Gene Expression Programming", X. Li, O.R. Zaiane and Z. Li (eds.):ADMA 2006, LNAI 4093, Springer-Verlag, Berlin, 871-880, 2006.

[14] WANG, Z., SUN, X., ZHANG, D., "Classification Rule Mining Based on Particle Swarm Optimization", G. Wang et al. (Eds.):RSKT 2006, LNAI 4062, Springer-Verlag, Berlin, 436-441, 2006.

[15] BAYKASOĞLU, A., ÖZBAKIR, L., "MEPAR-Miner: Multi-Expression Programming for Classification Rule Mining", European Journal of Operational Research, vol. 183, 767-784, 2007.

[16] OLTEAN, M., DUMITRESCU, D., "Multi Expression Programming", Department of Computer Science, Technical Report, Romania, 2002.

[17] OLTEAN, M., GROSAN, C., "Evolving Evolutionary Algorithms Using Multi Expression Programming", in: Proceedings of the 7th European Conference on Artificial Life, ECAL, Springer Berlin, 651-658, 2003.

[18] PARPINELLI, R.S., LOPES, H.S., FREITAS, A.A., "An Ant Colony Based System for Data Mining: Applications to Medical Data, Proc. Genetic and Evolutionary Computation Conf. (GECCO-2001), Morgan Kaufmann, San Francisco, California, 791-798, 2001.

[19] PARPINELLI, R.S., LOPES, H.S., FREITAS, A.A., "Data Mining with An Ant Colony Optimization Algorithm", IEEE Trans. On Evolutionary Computation, vol. 6(4), 321-332, 2002.

[20] CARVALHO, D.R., FREITAS, A.A., "A Genetic Algorithm with Sequential Niching for Discovering Small-Disjunct Rules", Proc. Genetic and Evolutionary Computation Conf. (GECCO-2002), NewYork, 1035-1042, 2002a.

[21] CARVALHO, D.R., FREITAS, A.A., "New Results for A Hybrid Decision Tree/Genetic Algorithm for Data Mining", Proc. 4th Int. Conf. on Recent Advances in Soft Computing (RASC-2002), 260-265, 2002b.

ÖZGEÇMİŞ

Lale ÖZBAKIR

1971 yılı Kayseri doğumlu Yrd. Doç. Dr. Lale Özbakır, Lisans öğrenimini 1992 yılında Bilkent Üniversitesi Bilgisayar ve Enformatik Mühendisliği Bölümünde tamamlamıştır. Erciyes Üniversitesi Sosyal Bilimler Enstitüsü Yönetim ve Organizasyon Anabilim Dalında 1997 yılında yüksek lisans derecesini, Üretim Yönetimi ve Pazarlama Anabilim Dalında 2004 yılında doktora derecesini almıştır. Erciyes Üniversitesi Endüstri Mühendisliği Bölümüne 1997 yılında araştırma görevlisi, 2004 yılında Yardımcı Doçent olarak atanmış olup halen aynı bölümde Öğretim Üyesi olarak görev yapmaktadır. Yazarın uluslararası bilimsel dergilerde 30'un üzerinde, ulusal ve uluslararası kongrelerde 50'nin üzerinde bilimsel yayını bulunmaktadır. Yrd. Doç. Dr. Lale Özbakır çok sayıda ulusal ve uluslararası dergide hakemlik görevi yapmakta olup, çalışma alanları içerisinde veri madenciliği, yapay zeka ve meta-sezgisel yaklaşımlar, evrimsel algoritmalar, yöneylem araştırması yer almaktadır.

Adil BAYKASOĞLU

Prof. Dr. Adil Baykasoğlu Isparta Teknik Lisesi Makina bölümünden mezun olduktan sonra Lisans ve Yüksek Lisans derecelerini Makina Mühendisliği alanında 1993 ve 1995 yıllarında Gaziantep'te, doktora derecesini ise YÖK bursu ile gittiği Nottingham Üniversitesinden 1999 yılında Endüstri Mühendisliği alanında almıştır. 1993-2010 yılları arasında Gaziantep Üniversitesi Endüstri Mühendisliği Bölümünde çalışan Prof. Baykasoğlu halen Dokuz Eylül Üniversitesi Endüstri Mühendisliği bölümünde çalışmaktadır. Prof. Baykasoğlu ulusal ve uluslar arası bilimsel dergi ve kongrelerde 300 civarında bilimsel makale yayımladı. Yazarın ayrıca üç adet yayımlanmış kitabı, düzenleyip editörlüğünü yaptığı çeşitli ulusal ve uluslar arası kongre kitapları bulunmaktadır. Yazarın çalışma alanları genelde yöneylem araştırması, bilişimsel yapay zekâ, zeki etmenler, lojistik ve üretim sistemleri yönetimi/tasarımı, bilgisayar destekli üretim, kalite ve benzetim konuları üzerinde yoğunlaşmaktadır. Prof. Baykasoğlu çok sayıda uluslararası dergide hakem ve yayın kurulu üyesi olarak görev yapmakta olup aynı zamanda Turkish Journal of Fuzzy Systems dergisinin eş-editörlüğünü yürütmektedir. Prof. Baykasoğlu'na 2007 yılında Türkiye Bilimler Akademisi Üstün Başarılı Genç Bilim İnsanı ödülü, 2008 yılında ODTÜ M. Parlar araştırma teşvik ödülü, 2010 yılında ise Tübitak Teşvik ödülü verilmiştir.