KOCAELİ ÜNİVERSİTESİ * FEN BİLİMLERİ ENSTİTÜSÜ

VERİ MADENCİLİĞİNDE GENETİK ALGORİTMALAR

YÜKSEK LİSANS

Bilgisayar Müh. Özlem Evrim GÜNDOĞDU

Anabilim Dalı: Bilgisayar Mühendisliği
Danışman: Yrd. Doç. Dr. Nevcihan DURU

KOCAELİ, 2007

KOCAELİ ÜNİVERSİTESİ * FEN BİLİMLERİ ENSTİTÜSÜ

VERİ MADENCİLİĞİNDE GENETİK ALGORİTMALAR

YÜKSEK LİSANS TEZİ

Bilgisayar Müh. Özlem Evrim GÜNDOĞDU

Tezin Enstitüye Verildiği Tarih: 04.06.2007
Tezin Savunulduğu Tarih: 28.06.2007

Tez Danışmanı	Üye	Üye
Yrd. Doç. Dr. Nevcihan DURU	Prof. Dr. Hülya YILDIRIM	Prof.Dr. Oya KALIPSIZ

KOCAELİ, 2007

ÖNSÖZ ve TEŞEKKÜR

Günümüz teknolojileri arasında önemli bir yeri olan veri madenciliği bir çok farklı alanda kullanılmakta olup; geliştirdiği çözümlerle bundan sonra da veri madenciliği için yapılan çalışmalarda ilerlemelerin kaydedilmesine katkıda bulunacaktır. Tez kapsamında hedeflenen veri madenciliğinin literatürdeki örneklerinin incelemesi ve veri madenciliği yöntemlerinden olan genetik algoritmalar ile öğrenci verilerinden oluşturulmuş veri tabanı kullanılarak bir analiz aracının geliştirilmesidir.

Bu tez çalışmasında desteği ve bilgi birikimi ile yardımlarını esirgemeyen tez danışmanım Kocaeli Üniversitesi Bilgisayar Mühendisliği Bölümü Öğretim Üyesi Yrd. Doç. Dr. Nevcihan Duru'ya teşekkür ederim.

Tez uygulamasında kullanılmış verilerin temini için KO.Ü. Bilgi İşlem Daire Başkanlığına yardımlarından dolayı teşekkür ederim.

Ayrıca güven ve desteğini her zaman hissettiğim eşime teşekkür ederim.

İÇİNDEKİLER

ŞEKİLLER DİZİNİ

SEMBOLLER

T : Popülasyondaki tüm bireylerin uygunluk fonksiyon değeri toplamı
r : Rasgele seçilmiş bir tamsayı
K : Toplam ebeveyn sayısı
Fitness(i) : Uygunluk fonksiyonu
Acc(i) : Kuralın kestirim oranı
Surp(i) : Kuralın ilginçlik ölçütü
N : Bireyin o sınıftaki her bir bireyle aynı olan nitelik sayıları toplamı
M : Bireyin ait olmadığı sınıftaki her bir bireyle aynı olan nitelik sayıları toplamı
c : Sabit değer

Kısaltmalar

AHA : Adaptative Hypermedia for All
ÇDH : Çoklu Dizi Hizalama
GA : Genetik Algoritmalar
KDD : Knowledge Discovery From Databases
KO.Ü : Kocaeli Üniversitesi
OLAP : Online Analytical Processing
ÖSS : Öğrenci Seçme ve Yerleştirme Sınavı
ÖSYM : Öğrenci Seçme ve Yerleştirme Merkezi
RAGA : Rule Acquisition with a Genetic Algorithm
SQL : Structured Query Language

VERİ MADENCİLİĞİNDE GENETİK ALGORİTMALAR

Özlem Evrim GÜNDOĞDU

Anahtar Kelimeler: Veri Madenciliği, Genetik Algoritmalar, Öğrenci Verileriyle Veri Madenciliği

Özet: Bilgi teknolojilerinin doğal gelişim sonucu olan veri madenciliği büyük veri yığınları içerisinden anlamlı veri birlikteliklerinin yakalanabilmesi için; akıllı metotlar yardımıyla bu birlikteliklerin çekilmesi işlemidir. Başka bir deyişle, büyük veri yığınları içerisinden veriyi madenleme olarak da tanımlanabilir. Veri madenciliği pazarlama, bankacılık ve finans, tıp ve ilaç sektörü, biyoloji, genetik, endüstri ve mühendislik, eğitim gibi bir çok alanda bulunan verilerden anlamlı sonuçların çıkartılabilmesi için kullanılmaktadır.

Bu tez kapsamında hedeflenen veri madenciliği ve genetik algoritmaların incelenmesi öğrenci verileri kullanılarak bu iki yöntemin birleştirilmesi sonucu ortaya çıkacak olan kuralların analiz edilmesidir. Genetik algoritma hızlı çalışan ve büyük veri kümelerinde iyi sonuçlar üretebilen bir sınıflandırma algoritması olduğu için tercih edilmiştir.

Yapılan uygulamayla, genetik algoritmalar ile ortaya çıkan kurallar ve parametreler aracılığıyla yapılan bazı değişikliklerin sonuçlara olan etkisi incelenmiştir. Veri madenciliğinde sınıflandırma algoritması olarak kullanılan genetik algoritmaların performansı ve hızı gözlenmeye çalışılmış olup; algoritmanın veri madenciliğinde kullanım şekli iyileştirilmeye çalışılmıştır.

Çalışmada KO.Ü öğrenci bilgi sisteminden alınan 2003 ve 2004 girişli öğrencilerin verileri kullanılmıştır. Öğrenci verilerinin niteliklerinin incelenmesi için geliştirilen bu çalışma da öğrencilerin durumları ile ilginç kurallar yakalanmaya çalışılmış ve ilerisi için kullanılabilir sonuçlar ortaya çıkartılması amaçlanmıştır.

THE GENETIC ALGORITHMS IN DATA MINING

Özlem Evrim GÜNDOĞDU

Keywords: Data Mining, Genetic Algorithms, Data Mining in Education

Abstract: Data mining is a method that extracting meaningful knowledge from large amounts of data using intelligent methods. Data mining can be viewed as a result of the national evolution of information technogies and describe as mining the information. Data mining had been using in many areas effectively like marketing, banking and insurance, medicine, biology, genetic, industry and engineering, education etc. to obtain meaningful results.

The objective of this study was to examine data mining and genetic algorithms and to analyze student's database by using the software application which was included this algorithm.

In this study, rules are analyzied that are results of genetic algortihms. Genetic algorithms are choiced because it is a fast and gives good results in large amounts of datasets.

In this study, student's database is obtained from Kocaeli University Student Information System. This study is developed to analyze student's informations. The purpose of analyzing rules that poduced with genetic algorithms, using them for taken decisions for the future.

1.GİRİŞ

Günümüz şartları ve gelişen teknolojileri insanoğlunun birçok alanda geçmişe yönelik deneyimlerini hatırlaması gerekliliğini ortaya koymuştur. Gün geçtikçe karmaşıklaşan yaşam ve hayat standartları yaşanılan ve elde edilen deneyimlerin kayıt altına alınmaya başlanmasıyla beraber insanoğlunun geleceğe yönelik anlamlı, kullanılabilir bilgi edinme ihtiyacı artmıştır. Bu durum sadece bireysel yaşamda değil toplum yaşamında da kendini göstermiş; sürekli gelişen, yeni atılımların peşinde olan bilim ve iş alanlarında da yerini almıştır.

Hızla gelişen bilgisayar, ağ ve donanım teknolojileri sayesinde geçmişe yönelik artan veri birikimi yukarıda bahsedilen deneyimlerin anlamlandırılmasını sağlamıştır. Şu an, sadece dünya gözlem uydularında günde 1 terabayt mertebelerinde veri üretilir duruma gelinmiştir [1]. Bunun gibi birçok alanda veri birikimi artmış; günümüz rekabetçi toplumunda örneğin bankacılık sektöründe bir müşterinin bu ay hesabına ne kadar para yatırdığı gibi basit SQL cümlecikleri yazımıyla alınabilecek cevaplar sektör gelişimlerinde yeterli olmamaya başlamıştır. Şu ana kadar hızlı ve iyi gelişmiş olan veri tabanı yönetim sistemleri artık sektörel bazda bir atılım ya da farklılık yaratabilmek için yeterli olmamaktadır. Bu durum sadece iş sektöründe değil tıp, eğitim, biyoloji, genetik gibi bir çok bilim dalında da farkedilmiştir.

Bunun yanında son zamanlardaki sistemler, analiz metotları için yeni olarak gereksinim duyulan çok geniş miktarlardaki veriyi tutabilmektedir. Günümüz veri tabanı uygulamaları temelde veri giriş-çıkış işlemlerini hızlı ve verimli bir şekilde gerçekleştirmeye yönelik olarak tasarlanmaktadır. Veri tabanlarından analiz amaçlı olarak çok sayıda bilginin çekilmesi ise farklı yönde bir teknik, eniyileme gerektirmektedir. Diğer bir anlamda, raporlama ya da analize yönelik tasarımlar veri giriş-çıkışlarını yavaşlatmakta, diğer taraftan giriş-çıkış işlemlerinin hızlı ve

verimli bir şekilde yapılmasına yönelik tasarımlar ise raporlama ya da analizi olumsuz etkilemektedir [20].

Bu sebeplerden dolayı ortaya çıkan ve hızla gelişen veri madenciliği bu bağlamda bilgi teknolojilerinin doğal gelişim süreci içerisinde de görülebilir. Veri madenciliği, büyük veri yığınları içerisinden anlamlı veri birlikteliklerinin yakalanabilmesi için; akıllı metotlar yardımıyla bu birlikteliklerin çekilmesi işlemidir. Başka bir deyişle, büyük veri yığınları içerisinden veriyi madenleme olarak da tanımlanabilir.

Veri madenciliği birçok alanda ilerlemeye yol açmıştır. Bankacılık sektörüne ilişkin kredi kartı uygulamaları, sigorta işlemleri gibi birçok konuda müşteri davranışlarının analizleri veri madenciliği yöntemleri sayesinde yapılabilmekte ve çalışmalar anlamlı sonuçlar üretebilmektedir. Örneğin bir müşterinin kredi kartı bilgilerinden yararlanılarak bu kişinin hayat sigortası yaptırmak isteyip istemeyeceği gibi analizler yapılabilmekte; bu da müşteri gruplandırılması, ileriki günlerde ne kadar harcama yapabileceği, yılbaşı gibi özel günlerde kampanyalar düzenlendiğinde katılıp katılmayacağı gibi ticari kaygı ve riskleri azaltabilecek sonuçlar üretilebilmektedir.

Bunun yanında veri madenciliği tıp, biyoloji ve genetik gibi birçok bilim dalında da kullanılabilir durumdadır. Ancak bu bilim dallarında yeterince kullanılamamasının en önemli nedeni veri kayıtlarının düzenli ya da gerektiği kadar bilgisayar ortamına geçirilememesi olabilir. Örneğin, tıp alanında yapılan bir çok çalışmada ilk sorun veri kayıtlarındaki düzensizlik, boşluklar daha da önemlisi bu verilerin sadece kağıt üzerinde tutuluyor olmasıdır. Bu nedenle tıbbi veri tabanları üzerinde yapılan çalışmalara literatürde çok fazla rastlanamamaktadır. Fakat yapılmış olan örneklerinden veri madenciliği yöntemleriyle bu alandaki gelişmelere büyük katkıda bulunulacağı açıktır. Örneğin, 2006 yılında Chen ve Hsu tarafından yapılan çalışmada, bir veri madenciliği yöntemi olan genetik algoritma tabanlı bir yaklaşımla göğüs kanseri örnekleri değerlendirilmektedir [14]. Daha önceleri yapılan çalışmalarda göğüs kanseri tanılarında istatistikle desteklenmiş yöntemler kullanılmıştır. Göğüs kanseri tanıları lineer olmayan tipte olmaktadır; bu nedenle istatistiksel yaklaşım kullanarak, bağımsız değişkenlerin içinden önemli olanını alıp kapsamlı bir model geliştirmek çok zordur. Son zamanlarda sinir ağlarıyla yapılan çalışmaların, bilinen istatiksel yaklaşımlardan daha güvenilebilir sonuçlar verdiğini göstermektedir. Sinir ağlarının kullanımının literatürde faydalı olduğu gösterilmiştir; fakat en büyük engel, kullanılan modelde ya da yapıda, sınıflandırma kurallarının

farkedilme zorluğunun olmasıdır. Bu çalışmada alınan sonuçlar ticari bir veri madenciliği yazılımıyla karşılaştırılmış ve deneysel olarak görülmüştür ki, modelin basitliğini artırmak ve kestirim oranını iyileştirmek için kural çekme yaklaşımı kurulmuştur. Bu çalışmadaki kural çıkarma sistemi, göğüs kanseri olma potansiyelini tespit edebilen diğer uzman sistemler kadar yetenekli olduğu görülmüştür.

Veri madenciliği görüldüğü üzere birçok alanda kullanılıp verimli sonuçlar elde edilmesinde etkili bir yöntem olup; her geçen gün hızla gelişen bir teknoloji olarak literatüre geçmiştir. Farklı alanlarda, farklı ihtiyaçlara cevap verilmesi ise veri madenciliğinin kendi içinde yöntemler geliştirilmesi ihtiyacını ortaya çıkarmıştır. Veri madenciliği kapsamında, Apriori, K-en yakın komşu, Genetik algoritmalar gibi algoritmalar geliştirilmiştir. Bu algoritmalardan biride genetik algoritmalardır.

Genetik algoritmalar ilk olarak Michigan Üniversitesi'nden psikoloji ve bilgisayar bilimi uzmanı John Holland tarafından ortaya atılmıştır [3]. Genetik algoritmaların tanımlanmasından uzun yıllar sonra tekrar irdelenmiş ve bir çok problemde uygulanabilecek yapıda olduğu fark edilmiştir. Önceleri sadece bir eniyileme yöntemi olarak görülen genetik algoritmalar, çok çeşitli problemlere çözüm üretebilir durumdadır. Genetik algoritmalar, Darwin'in evrim teorisinden yola çıkılarak, 'en iyi uyumu sağlayan bireyin hayatta kalacağı' ilkesi temel alınarak geliştirilmiş olup; algoritmanın temel yapısının oluşturan operatörleri çaprazlama, mutasyon, seçim ve bunların yanında kullanılan parametreler sayesinde doğru sonuçlara gidilmesini sağlamaktadır. Genetik algoritmaların yapısı oluşturulurken aşağıdaki kriterlere dikkat edilmelidir: Bireylerin gösterimi doğru bir şekilde yapılmalı, uygunluk fonksiyonu etkin bir şekilde oluşturulmalı, doğru genetik işlemciler seçilmeli [9].

Genetik algoritmalar veri madenciliği alanında kendine sınıflandırma ve arama algoritması olarak yer bulmuştur. Genetik algoritmaların veri madenciliğinde kullanılan bir çok örneği bulunmaktadır.

Veri madenciliğinde genetik algoritmaların kullanımı sırasında şunlara dikkat edilmelidir:

1.Kullanıcıya açıklanması ve anlatılması zor olabilir,

2.Sorunu soyutlamak ve bireyleri temsil etmek için kullanılan modeller zordur,

3.Uygunluk fonksiyonunu belirlemek zordur,

4.Çaprazlama ve mutasyon işlemlerinin nasıl yapıldığına dair sorunun çözümü zordur [5].

Veri madenciliğinde genetik algoritmalar kullanılarak yapılan birçok çalışma mevcuttur. Tıp eğitim, diğer algoritmalarla performans değerlendirmesi gibi konularda literatürde çalışmalara rastlamak mümkündür.

Alataş ve Arslan tarafından 2005 yılında yapılan bir çalışmada öğrenci verileri kullanılarak, {alındı, alınmadı} ya da {var yok} şeklinde ikili değerler dışında kategorik ve nicel değerler de içeren veri tabanlarında birliktelik kurallarının keşfi için yapay zeka ve zeki hesaplama tekniklerinden genetik algoritma bulanık mantık tabanlı etkili, yeni bir yöntem geliştirilmiştir. Genetik algoritmalarda başlangıç popülasyonunu gelişigüzel üretmek yerine, bunu çözüm uzayına düzgün dağıtan düzenli popülasyon yöntemi kullanılmıştır. Genelde kullanılan yöntemlerin aksine yüksek destek ve güven değerlerine sahip birliktelik kuralları yoğun nesne kümeleri üretilmeden direk olarak ve her veri tabanı için belirlenmesi güç olan minimum güven ve minimum destek eşiklerine ihtiyaç duyulmadan keşfedilmiştir. İlginç birliktelik kurallarını bulmak için uyarlamalı mutasyon ve elitizm stratejisi uygulanmıştır. Bu şekilde genetik algoritmanın son popülasyonu ilginç birliktelik kurallarını temsil etmiştir. Önerilen yöntem hem yapay bir veri tabanında hem de Fırat Üniversitesi Elektrik-Elektronik Mühendisliği lisans öğrencilerinin ders not kayıtlarında denenmiş, kullanışlı ve ilginç kurallar etkili şekilde bulunmuştur [13].

Aynı alanda Romero, Ventura, Bra ve Castro tarafından yapılan bir başka çalışmada da öğrenci bilgileri kullanılarak farklı kestirim kurallarının nasıl bulunabileceği gösterilmeye çalışılmış ve bunlar kullanılarak web üzerinden yapılan kurslarda düzeltilmesi gereken noktalar aranmıştır. Eğitim alanındaki gücünü arttırmak için AHA (Adaptative Hypermedia for All)' da birçok değişiklikler yapılmıştır. Kullanılabilecek veriler arasındaki ilişkiyi bulmak için AHA' da kütükte tutulmuş bilgiler kullanılmıştır (okuma zamanları seviye zorlukları ve test sonuçları). En ilgi çekici olan ilişkiler öğretmene gösterilmiştir. Böylece kursun daha verimli olabilmesi için gereken değişikliklerin farkedilebilirliği kolaylaştırılmıştır [17].

Romao, Freitas ve Gimenes tarafından 2004 yılında genetik algoritmalar kullanılarak bilim ve teknoloji verilerinden ilginç kuralların çekilebilmesi için bir çalışma yapılmıştır. Genetik algoritmaların başka algoritmalarla birlikte kullanıldığı bu çalışmada; veri madenciliğinde en sık kullanılan verilerin gösterim biçimi olan "İf-Then" kuralları formundaki verilerin keşfi konu alınmıştır. Bu bağlamda, ilginç olan bulanık kestirim kurallarının keşfi için bir genetik algoritma (GA) tasarlanmıştır. GA, kullanıcı için yeni ve sürpriz olabilecek kestirim kurallarını arar. Ayrıca bulanık mantık, GA ile elde edilmiş kuralların anlaşılabilirliğini kolaylaştırır; çünkü terimler sözlük anlamlarıyla kullanılır. Örnek gerçek bir bilim ve teknoloji veritabanında uygulanmıştır. GA ve J4.8 algoritmaları bu örnekte karşılaştırmalı olarak verilmiştir. Deneylerde, kestirim oranı ve her iki algoritmada elde edilmiş kuralların ilginçlik derecelerine bakılmıştır. Bu çalışmada kurallar GA ile elde edilmiştir ve en iyi kuralların elde edilmesi için de J4.8 kullanılmıştır [12].

Jourdan, Dhaenens ve Talbi tarafından tıp verileriyle yapılan bir veri madenciliği uygulamasında da genetik algoritmalar nitelik seçimi için kullanılmış olup; genetik özelliklerin ve çevresel faktörlerin obezite ve diabet hastalığı gibi birden fazla faktöre bağlı olan hastalıklar üzerindeki etkisini incelemişlerdir. Deneyler Lille Biyolojik Enstitüsünün verileriyle yapılmıştır. Çok büyük sayıda veri olduğundan, keşifsel (heuristic) yaklaşım seçilmiştir. İlk aşamada nitelik seçimi için genetik algoritmalar kullanılmıştır. Bu problemin çözebilmek için genetik algoritmalarda tanımlanmış bazı yöntemlere başvurulmuştur; bu yöntemler paylaşım, göç, genetik operatörler gibi. İkinci aşamada, bir önceki aşamada seçilen niteliklerin

sınıflandırılmasına çalışılmıştır. Bunun içinde en popüler sınıflandırma algoritması olan k-means kullanılmıştır [15].

Tıp verileriyle 2003 yılında yapılan bir başka çalışmada da Toprak, Ganiz, Toprak, Arslan tarafından genetik algoritmalar kullanılmıştır. Genetik algoritmalar (GA) doğadaki evrimsel süreçleri model olarak kullanan bilgisayara dayalı problem çözme teknikleridir. Makine öğrenmesi ise deneyim ile otomatik olarak kendisini geliştiren bilgisayar programlarını nasıl yapabiliriz sorusuyla ilgilenir. Bu makalede, GA'ların makine öğrenmesinde kullanımı irdelenmiş ve konuyla ilgili olarak tıp alanında bir gerçek hayat problemi olan "cerrahi müdahale geçirecek hastada kardiyak riskinin belirlenmesi " ele alınmıştır. Bu problem için makine öğrenmesine uygun bir hipotez uzayı tanımlanarak eğitim örnekleri hazırlanmıştır. Bu eğitim örnekleri kullanılarak GA ile makine öğrenmesi uygulaması için gereken alt yapı sağlanmıştır [26].

Genetik algoritmalarla yapılan çalışmalarda farklı yöntemlerin de katkısıyla hibrit yöntemler geliştirilmeye çalışılmıştır. 2001 yılında Cattral, Oppacher, Deugo tarafından yayınlanan bir makalede RAGA (Rule Acquisition with a Genetic Algorithm) isimli bir çalışma anlatılmıştır. RAGA, eğitimli (supervised) ve eğitimsiz (unsupervised) veri madenciliği alanları için genetik algoritmayla genetik programlamanın karma kullanımı (hibriti) şeklinde tanımlanabilir [22].

Veri madenciliğinde kullanılan bir çok yöntem olmasının doğal sonucu olarak bu yöntemlerin performans karşılaştırılmalarının yapıldığı çalışmalarda önem kazanmıştır. Werner, Fogarty tarafından yapılan bir çalışmada sınıflandırma tekniği olan k-means ile iki farklı genetik algoritma yaklaşımı örnek bir veri kümesi kullanılarak karşılaştırılmıştır. İki algoritma arasında yapılan karşılaştırmada genetik algoritmayla elde edilen sonuçların daha iyi olduğu görülmüştür [6].

Defalce, Cioppa, Tarantino' da diğer tekniklerle karşılaştırmalı olarak veri madenciliğindeki çok büyük boyuttaki verilerden alışılmamış ya da ilginç olan bilgileri keşfedebilme probleminin çözüm yöntemine ilişkin bir uygulama yapmışlardır. Bu tür problemler genellikle alınmış örneklerin standart sorgulama mekanizmalarında ya da klasik istatiksel metotlarda kullanım zorlukları ortaya

çıktığı zaman sezgisel olarak çözülürdü. Bu makalede, otomatik kural keşfi süreci insanlar tarafından kolaylıkla anlaşılabilen bir genetik programlama yapısı sunulmuştur. Diğer tekniklerle de karşılaştırıldığında sonuçların başarılı olduğu görülmüştür. Ayrıca, elde edilmiş olan bazı kurallar gösterilmiş ve ayrılmış olan değerler ispatlanmıştır [19].

Genetik algoritmalar, Apriori algoritmasıyla beraber kullanılan başka bir çalışmada da sınıflandırma amaçlı kullanılmıştır. Genellikle, genetik programlama kullanılarak oluşturulmuş sistemlerin öğrenme hızlarının yavaş olduğu görülmüş. Bu nedenle, çevreye koşullarına göre ayarlanmış olan yüksek önceliğe sahip bir öğrenme sistemi kurulabilir; çünkü yapı da aynı zamanda oluşturulmaktadır.

Nimi, Tazaki tarafından yayınlanan bir makalede büyük veritabanları için kural oluşturulmasında kullanılan Apriori algoritmasıdır incelenmiştir. Apriori bir birliktelik kuralı algoritmasıdır. Apriori algoritması kural yapısı için iki değer kullanır: destek (support) ve güven (confidence). Her indeksin eşik değerine bağlı olarak arama uzayı küçültülebilir ya da aday olan birliktelik kurallarının sayısı çoğaltılabilir. Bununla birlikte etkili bir eşik değeri oluşturabilmek için deneyim gereklidir. Yukarıda bahsedilen her iki teknikte avantaj ve dezavantajlar içermektedir. Bu makalede, Apriori algoritmasıyla genetik programlama birleştirilerek, veritabanları için kural keşfinde kullanılacak tekniklerin bulunması amaçlanmıştır. Kurallarını böyle oluşturan bir öğrenme metodunun kullanılmasındaki amaç büyük veritabanlarındaki değişken kurallarının aranmasını sisteme oturtmaktır [18].

Genetik algoritmalar kural kümeleriyle ilgili yapılan bir çalışmada ise şöyle kullanılmıştır: Robotlarla ilgili yapılan uygulamalarda, başarılı çözüm teknikleri gerektiren birçok problem meydana gelir. Evrimsel Hesaplama, bu durumda çıkabilecek problemlerin bir kısmında başarı getiren bir yöntemdir. Evrimsel metotlar¸ ceşitli akıllı robot mimarilerinde uygulanmışlardır. Örneğin, evrimsel algoritmalar kural tabanlı otonom ajanların kural kümelerinin öğrenilmesinde, robot kontrolü için kullanılan sinir ağlarının ağırlıklarının ve topolojisinin

öğrenilmesinde, bulanık mantık kontrol sistemlerinde ve davranış tabanlı robotların kurallarında kullanılmaktadır [23].

Genetik algoritmalar, Kaya, Alhajj tarafından bulanık birliktelik kurallarının madenciliği için geliştirilmiş olan bir uygulamada kullanılmıştır. Bu amaçla, ilk önce sınıflandırma temelli olan genetik algoritmalar ikinci olarak da literatürde çok bahsedilen ve etkili bir yöntem olan örneklemeyle sınıflandırma yöntemi kullanılmıştır. Genetik algoritma tabanlı yaklaşımla literatürde geçen diğer yaklaşımlar karşılaştırılmıştır. Deneylerde, genetik algoritmayla bulunan ilginç kurallarının sayısının diğer metotlara göre daha fazla olduğu görülmüştür. Deneyler gerçek veri kümesiyle yapılmış olup;önerilen yaklaşımın anlamlı sonuçlar ürettiği verimli ve etkili olduğu görülmüştür [24].

Lin, Kuo tarafından OLAP (online analytical processing) sistemlerinde genetik algoritmalar kullanılarak bir analiz yapılmıştır. Çok yönlü veri analizi olarak da bilinen OLAP sistemlerindeki gibi uzun zamandır toplanmış verilerin, bir çok birleştirme fonksiyonunun işletilmesini gerektiren sistemler içinde genetik algoritma yöntemleri kullanılmıştır. Sorgulama zamanını azaltmak ve analistlerin sunulabilecek bakış açılarını arttırabilmek için; bu veriler genellikle veri küpleri olarak adlandırılan çok yönlü veri modeli şeklinde organize edilirler. Bir veri küpündeki her hücre, farklı yönler için farklı değerler taşır. Veri küpü seçimindeki problem kullanıcı sorguları ve depolanacak yer kısıtlarına rağmen, bakım ve/veya sorgulama değerlerini küçültebilmek için veri küpleri arasından gerçekleştirilmiş küplerin seçimidir. Bu problem NP-hard problemi olarak bilinir. Bu makalede, genetik algoritmaların bir uygulaması küp seçimi için incelemiştir. Aynı depolama kısıtlarına rağmen, sonuçta bakım ve sorgulama değerlerinde düşüş görülmüştür. Genetik algoritmaların her aşamada doğru, iyi sonuçlar verdiği görülmüştür [25].

Bu tezde veri madenciliği yöntemlerinden olan genetik algoritmalar ile öğrenci verileri kullanılarak bir uygulama geliştirilmiştir. Öğrenci verileri KO.Ü. Bilgi İşlem Daire Başkanlığından alınmış ve veri önişleme aşamalarından geçirilerek tablolara aktarılmıştır. Veriler öğrencilerin ÖSYM tarafından gönderilen bilgilerinden ve

öğrenci bilgi sisteminde bulunan öğrenci ders başarılarını içeren verilerin birleştirilmesiyle tablolara aktarılmıştır.

Veri madenciliği yöntemlerinden genetik algoritmaların seçilme nedenleri şöyle sıralanabilir:

1.Genetik algoritmaların iyi bir sınıflandırma ve arama algoritması olmaları.

2.Genetik algoritmalar hızlı çalışan ve arama uzayında yerel çözümlere takılmadan verimli çalışabilen bir algoritmadır.

3.Genetik algoritmalar uygulanacak probleme özgü geliştirilebilir ve genetik operatörler ve parametreler doğru seçildiği sürece her daldaki ve tipteki veriler ile çalışılabilirler.

4.Genetik algoritmalar başlangıç çözümünden bağımsız çalıştıkları için diğer yöntemlere göre daha uygun bir çözüm yöntemi olarak öne çıkmaktadır.

Bu tez çalışmasının amacı;

1.Veri madenciliği, genetik algoritmalar ve kural keşfi ile ilgili literatür çalışması yapılması, genetik algoritmalarla yapılmış olan çalışmaların incelenmesi ve konuyla ilgili yazı hazırlanması,

2.Sınıflandırma teknikleri içerisinde olan genetik algoritmalar ile öğrenci verilerinden oluşturulan veritabanı üzerinde algoritmayı incelemek ve performansını değerlendirilmesi,

3.Genetik algoritmalar kullanılarak öğrenci verilerinden oluşan veritabanını eğitim alanında ilginç ve yararlı olabilecek kuralların keşfi için kullanılmasını sağlayacak bir uygulama geliştirilmesi.

Çalışmanın literatür taramasında bilimsel makalelerden, konu ile ilgili kitap ve sempozyum bildirilerinden yararlanılmıştır. Uygulamada kullanılan veri tabanı KO.Ü. Bilgi İşlem Daire Başkanlığından alınmıştır.

Geliştirilen uygulama Matlab, SPSS gibi veri madenciliğinde sık kullanılan paket programlar kullanılmadan gerçekleştirilmiştir. Uygulamada kullanılan yazılım paket programlardaki gibi kullanıcı kısıtlamaları içermemektedir. Geliştirilen yazılımın kullanımı için kolay ve kullanıcı tarafından kolay anlaşılabilecek bir ara yüz tasarlanmıştır.

Bu tez çalışması beş bölümden oluşmaktadır. Tezin ikinci bölümünde veri madenciliğine giriş yapılmıştır. Bu bölümde veri madenciliğinin tanımı, diğer disiplinlerle olan ilişkileri, bilgi keşfi süreci ve veri madenciliğinde kullanılan yöntemlere yer verilmiştir. Veri madenciliği konusu anlatılırken yeri geldikçe bu konuda yapılmış olan çalışmalara da değinilmiştir.

Üçüncü bölümde veri madenciliği tekniklerinden olan genetik algoritmalar detaylı olarak anlatılmıştır. Genetik algoritmaların temelinin anlatıldığı bu bölümde, genetik algoritmaların nasıl ortaya çıktığı, hangi mantık üzerine oturtularak geliştirildiği, genetik algoritmalarda kullanılan operatörler ve parametreler, genetik algoritmaların uygulama alanlarına ayrıntılı olarak yer verilmiştir. Son olarak genetik algoritmaların veri madenciliğinde nasıl kullanıldığı örneklerle açıklanmış ve performansı değerlendirilmiştir.

Uygulamanın anlatıldığı dördüncü bölümde öncelikle uygulamanın amaçlarından ve sağlayacağı yararlardan bahsedilmiştir. Bu çalışmada kullanılan öğrenci verilerinin veri tabanına nasıl aktarıldığına, nitelik seçimlerindeki önceliklere ve veri madenciliğinin önemli bir aşaması olan veri önişleme işlemine yer verilmiştir. Uygulama için genetik algoritmalarda probleme uygun seçilmesi gereken operatör ve parametrelerin neden tercih edildiği anlatılmış olup; algoritmanın çalışma mantığına ayrıntılı olarak değinilmiştir. Verilerin analizinde izlenen yol ve analizlerin gösterim şekilleri açıklanmış, uygulamanın ara yüzlerinin nasıl kullanılacağı ayrıntılı bir biçimde anlatılmıştır.

Tezin sonuçlar ve öneriler bölümünde geliştirilen uygulama özetlenerek, genetik algoritmaların seçilme nedenleri ve sağladığı avantajlar anlatılmıştır. Genetik algoritmalar kullanılarak geliştirilen yazılımın sağladığı yararlardan bahsedilmiş ve performansı analiz edilmiştir. Uygulamada karşılaşılan zorluklardan bahsedilerek, daha sonra yapılabilecek çalışmalar için önerilerde bulunulmuştur.

2. VERİ MADENCİLİĞİ

2.1. Veri Madenciliği Nasıl Ortaya Çıkmıştır ?

Yaşadığımız yüzyıl içerisinde bilgisayar teknolojilerinde ve bununla bağıntılı olarak birçok alanda hızlı gelişmeler olmuştur. Teknolojideki hızlı gelişim süreci beraberinde aşılması gereken birçok probleme de yol açmıştır. Bunun sonuçlarından biri olarak da elimizde toplanan veri miktarındaki önemli artış örnek verilebilir. Şu an, sadece dünya gözlem uydularında günde 1 terabayt mertebelerinde veri üretilir duruma gelinmiştir [1].

Depolanan ve işlenmesi gereken birçok konuda büyük miktarlarda veri birikimi elde edilmiştir. Bu da daha gelişmiş veri tabanı yönetim sistemlerine, dolayısıyla veri toplama ve toplanan verileri saklama olanaklarının geliştirilmesi ihtiyacını ortaya çıkarmış bu konuda etkili yazılımların kullanılmasını gerektirmiştir. Verilerin veri tabanı yönetim sistemleriyle sayısal olarak saklanabilmesi ise detaylı ve doğru bilgiye bilgisayar ağları yardımıyla hızlı ve etkin bir biçimde ulaşma imkanı sağlamıştır.

Bilgisayar teknolojilerindeki, birbirini tetikleyen ve birbiriyle etkileşimli olarak meydana gelen bu ilerlemeler; beraberinde bu kadar bilginin etkili bir biçimde kullanılması gerekliliğini ortaya çıkarmıştır. Sonuçta, veri kendi başına değersizdir. İstediğimiz ise amacımız doğrultusundaki bilgilere ulaşmaktır. Bu nedenle veriyi, amaca yönelik işlenmiş bilgi olarak tanımlamak yerinde olur [2].

Veri tabanları bu kadar büyük ölçekli olduğundan; bu veriler arasından yararlı verilere ve bilgilere erişim ihtiyacı doğmuş; doğal olarak bu yararlı verilerin analizinin elle ve gözle yapılamayacağı anlaşılmıştır. Bu nedenle de, veri madenciliği güncel araştırma konularından biri olmuş; iş yönetimi, ürün kontrol sistemleri,

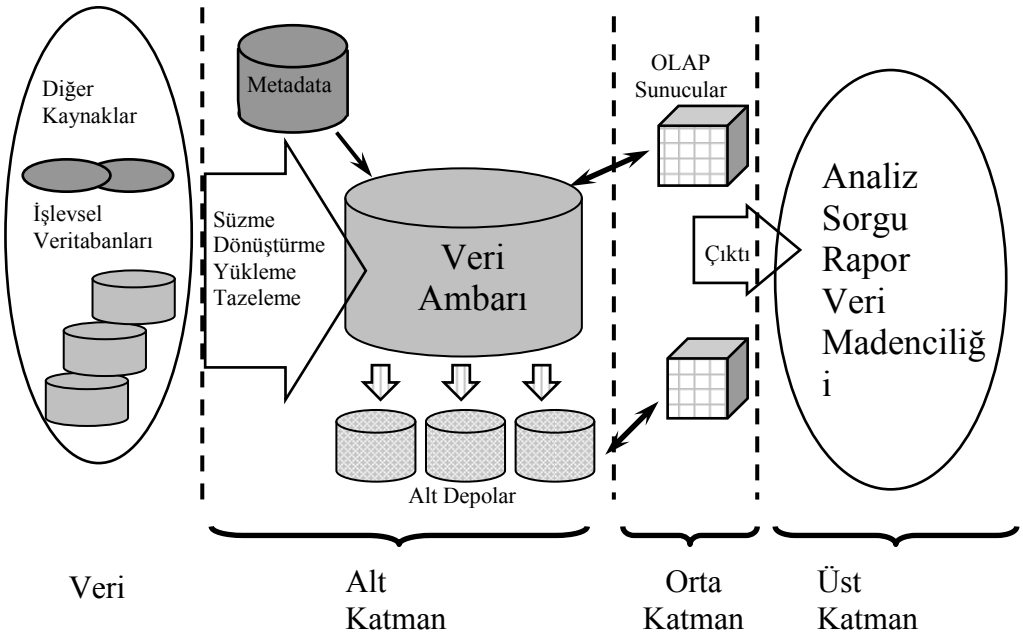
eğitim, pazarlama mühendislik gibi alanlarda veri madenciliği yöntemleriyle analizler yapılmaya başlanmıştır.

Veri tabanı teknolojilerinin gelişim sürecine bakıldığında her geçen yıl hızlı ilerleyen bir teknoloji olduğu görülmektedir.

1960'lar: Veri toplama, veri tabanlarının oluşturulması
1970'ler: İlişkisel veri modeli, ilişkisel veritabanı yönetim sistemleri (OLTP)
1980'ler: RDBMS, ileri veri modelleri ve uygulama kaynaklı veritabanı yönetim sistemleri (uzamsal, bilimsel, mühendislik, vb.)
1990'lar ve 2000'ler: Veri madenciliği ve veri ambarlama, çoklu ortam veri tabanları ve web veritabanları [5].

Veri madenciliği, bilgi teknolojilerinin doğal gelişim sürecinin sonucu olarak da değerlendirilebilir. Bu bağlamda veri analizi, veri ambarcılığı ve veri madenciliği aşamalarını gerektirir.

Veri ambarı günlük işlemlerin gerçekleştirildiği sistemlerin arkasındadır. Bu sistemlerde oluşan veriler işletmenin seçimine göre belirlenen periyotlarla veri ambarına aktarılırlar. Veri ambarları, veri temizleme, veri dönüştürme, veri yükleme, ve periyodik veri transferi işlemlerinden inşa edilmişlerdir [2].

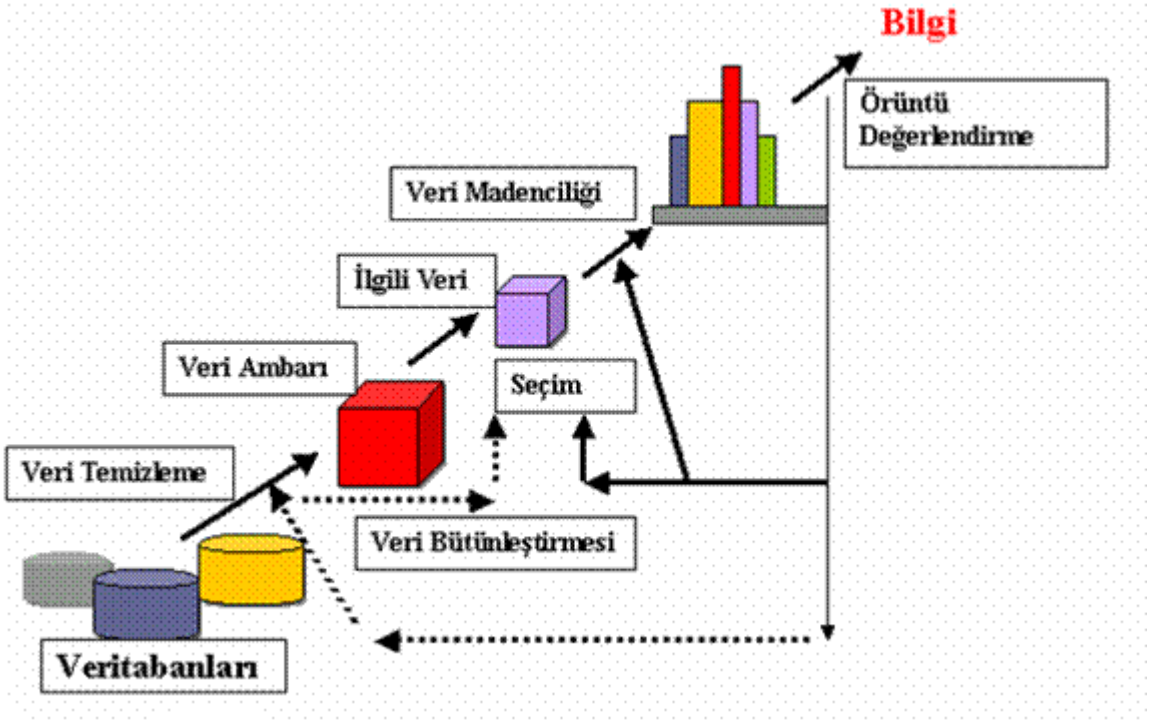


Şekil 2.1 Veri ambarı aşamaları [21]

Şekil 2.1'de gösterilen aşamalardan geçen veri, veri madenciliği sayesinde önümüze işlenmiş ve anlamlı örüntüler halinde gelir. Veri ambarları birden fazla kolda hizmet vermekte ve tekrar işlenmiş olan veriler veri ambarına geri dönmektedir.

2.2. Veri Madenciliği Nedir?

Veri madenciliği, büyük veri yığınları içerisinden anlamlı veri birlikteliklerinin yakalanabilmesi için; akıllı metotlar yardımıyla bu birlikteliklerin çekilmesi işlemidir. Başka bir deyişle, büyük veri yığınları içerisinden veriyi madenleme olarak da tanımlanabilir. Bununla birlikte veri madenciliği nitelendirmesi, literatürde başka deyimlerle de isimlendirilmiştir. Veri tabanlarında bilgi madenciliği (knowledge mining from databases), bilgi çıkarımı (knowledge extraction), data/pattern analysis (veri ve örüntü analizi), veri arkeolojisi gibi. Bunların arasında en çok kullanılanı veritabanlarından bilgi keşfi (Knowledge Discovery From Databases-KDD)' dir. Burada veri madenciliği bilgi keşfi sürecinin bir parçası olarak görülmektedir [2]. Şekil 2.2'de veritabanlarından bilgi keşfi aşamaları gösterilmiştir. Bu aşamalardan biri olan veri madenciliği örüntü değerlendirmeden bir önceki adımdır.



Şekil 2.2 Bilgi keşfinin süreçleri [5]

Veri madenciliği ve elektronik ticaretle ilgili yapılan bir çalışmada da, veri madenciliği şöyle tanımlanmıştır: Temel olarak veri madenciliği, veri kümeleri arasındaki desenlerin ya da düzenin, verinin analizi ve yazılım tekniklerinin kullanılması ile ilgilidir. Veriler arasındaki ilişkiyi, kuralları ve özellikleri belirlemekten bilgisayar sorumludur. Amaç daha önceden fark edilmemiş veri desenlerini tespit edebilmektir.

Veri madenciliğini istatistiksel bir yöntemler serisi olarak görmek mümkün olabilir. Ancak veri madenciliği, geleneksel istatistikten birkaç yönde farklılık gösterir. Veri madenciliğinde amaç, kolaylıkla mantıksal kurallara ya da görsel sunumlara çevrilebilecek nitel modellerin çıkarılmasıdır. Bu bağlamda, veri madenciliği insan merkezlidir ve bazen insan – bilgisayar ara yüzü birleştirilir [4].

Veri madenciliğine bütün olarak bakılacak olursa temel bileşenleri şunlardır:

1.Veri tabanı, veri ambarı ve diğer depolama teknikleri

2.Veri tabanı yada veri ambarı sunucusu

3.Bilgi Tabanı

4.Veri Madenciliği Motoru

5.Örüntü Değerlendirme

6.Kullanıcı Arayüzü

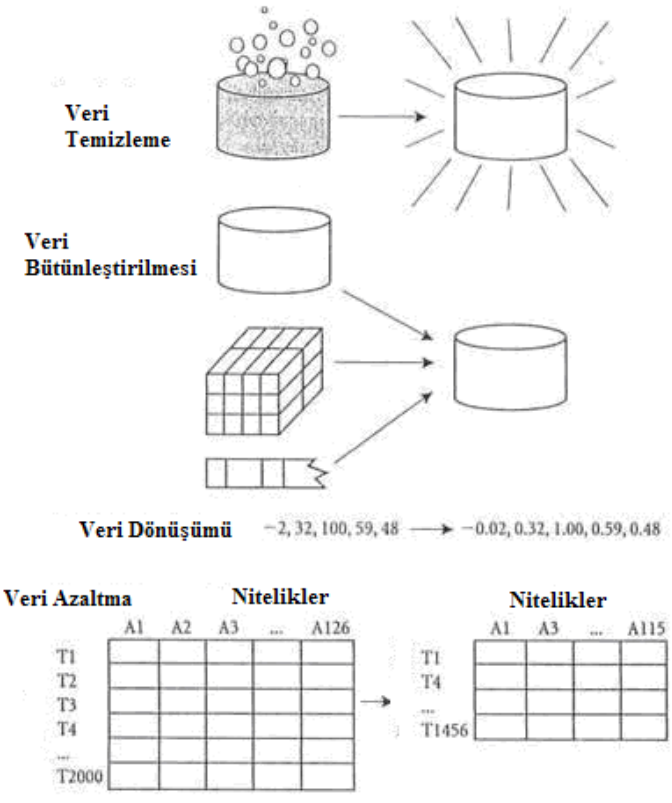
Veritabanı ve veri çekmeyle bu kadar iç içe olduğu için veri madenciliği; sorgulama işlemi, öğrenen sistemler ya da istatistiksel programlarla karıştırılıyor olabilir. Oysa veri madenciliği veritabanlarından veri çekmeyi değil, anlamlı ve gelecek için tahmin yapılabilecek birliktelik kurallarını çekme işlemidir. Bu yüzdendir ki, veri madenciliği veritabanı çözümlemesi ve karar desteği aşamalarında kullanılır.

2.3. Veri Madenciliğinde Veri Önişleme Aşamaları

Veri madenciliğinde çok büyük boyuttaki veriler üzerinde çalışıldığından daha önce bahsedildi. Ne kadar çok ve çeşitli veriye sahip olunursa, geleceğe yönelik anlamlı, işe yarar örüntüler çıkarma olasılığı, o orantıda artabilir.

Depolanan ya da kullanılacak olan verinin büyük boyutlarda olması; bu verilerin ne şekilde saklanacağı konusunu önemli bir sorun haline gelmiştir. Bu da veri madenciliğinden önce, veri ambarcılığı sürecinin içerisinde gerçekleştirilen veri önişleme işlemini zorunlu hale getirmiştir. Veri tabanındaki eksik veri birçok sebepten kaynaklanmış olabilir. İstenilen nitelikler her zaman hazır veya elde edilebilir düzeyde olmayabilir, örneğin satış işlemleri verisinde müşteri bilgileri bulunmayabilir. Kayıt sırasında önemli olabilecekleri düşünülmediği için diğer veriler tamamen hesaba katılmamış olabilir. Anlayışsızlık veya kötü çalışma nedeniyle uyumlu veya ilgili veriler kaydedilmemiş olabilir. Diğer kayıtlarlardaki verilerle çelişkili olan veriler silinmiş olabilir. Üstelik verinin geçmişi ve değişimi ile kayıtlar dikkate alınmamış olabilir. Bütün bunlar düşünüldüğünde veri önişleme işleminin, sonuçların daha kaliteli ve anlamlı çıkmasını sağlayacağı açıktır.

Veri önişleme süreci neleri içerir? Verinin, veri madenciliği sürecinden önce işlenmesi verinin içindeki gürültünün temizlenmesi, varsa bazı verilerdeki boşlukların doldurulması, tutarsız verilerin göz ardı edilmesi gibi aşamalardan geçmesini gerektirmektedir. Bu bölümde bu aşamalardan bahsedilecektir.



Şekil 2.3 Veri önişleme formları [2]

2.3.1 Veri Temizleme

Veri temizleme rutinleri kayıp değerleri doldurur, aykırı değerleri tanımlarken dış gürültüleri düzeltir ve verideki uyumsuzluğu düzeltir [2].

1.Kayıp Değerler

Kullanılan veritabanının tablolarında bazı kayıtların olmadığını varsayılsın. Bu boşlukların doldurulmasında aşağıdaki yöntemler kullanılabilir:

a.Tablo satırını yoksaymak: Sınıflandırma veya tanımlama gerektiren madencilik için kullanılır. Fakat bu metot çok fazla kayıp değer yoksa fazla etkili değildir.

b.Kayıp değerlerin el ile doldurulması: Bu yöntem çoğu zaman kullanılmaktadır; fakat veritabanı çok büyük boyutlardaysa uygun değildir.

c.Kayıp değerler için genel sabit kullanımı: Bütün kayıp değerler için aynı sabit değer yerleştirilir. Örneğin ’bilinmiyor’ yazılabilir. Bu durumun veri madenciliğinde genel bir kavram formu olarak algılanabileceği için pek tavsiye edilmez.

d.Kayıp değerleri doldurmak için ortalama özellikleri kullanmak: Burada da bir firmanın maaş değerlerinde eksiklik var ise; ortalama maaş değeri eksik verilerin yerine yazılabilir.

e.Tablo satırında verilen aynı sınıfa ait bütün değerler için aynı ortalama değeri kullanmak: Örneğin kredi kartı bilgilerinin tutulduğu bir tabloda müşterinin gelir değeri yerine maaş değerlerinin yazılması gibi.

f.Kayıp değerlerin doldurulması için en olası değerlerin yazılması: Örneğin müşterilerle ilgili diğer özellikleri kullanarak maaş değeri için ortalama bir değer atamak gibi. Burada ortalama maaş tespiti karar ağaçları kullanılarak yapılabilir.

Yöntem c ve f arasındaki metotlar veri odaklıdır. Değerin içi doğru olmayan verilerle doldurulabilir. Buna rağmen metot f en popüler metottur. Diğer metotlarla karşılaştırıldığında kayıp değeri tahmin etmek için şimdiki değer üzerinde oldukça bilgi kullanır [2].

2.Gürültülü Veri

Gürültü standart hatadır veya ölçülmüş değişkenlerin miktarındaki değişikliktir. Verilerdeki gürültüyü yok etmek için aşağıdaki yöntemler kullanılır:

a.Kutulama: Kutulama metodu birbirleriyle yakın şekilde sıralanmış değerleri yakınındaki "komşu" değerler aracılığıyla düzeltir. Yakın değerler birçok kova veya kutu içinde dağıtılmıştır. Çünkü kutulama metodunda komşu değerlere danışılır, bu metot yerel düzeltme gerçekleştirir. Fiyat verilerine dayanarak bir örnek verilecek olursa:

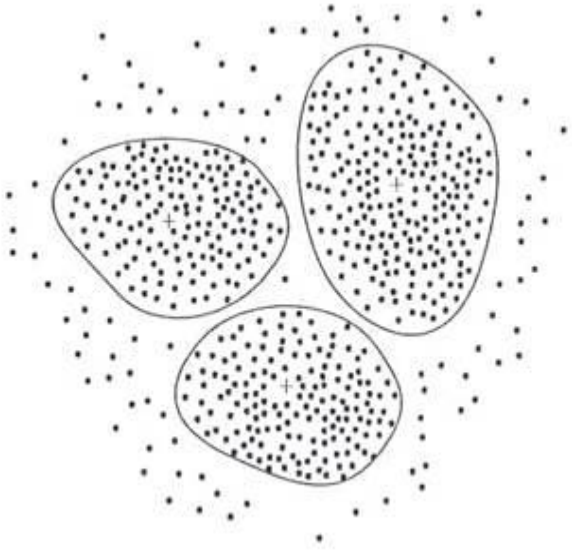
Fiyat için veri sıralaması: 4, 8, 15, 21, 21, 24, 25, 28, 34

Kutu ortalaması ile veri düzeltmesi:
Kutu 1: 9, 9, 9
Kutu 2: 22, 22, 22
Kutu 3: 29, 29, 29

Kutu sınırları ile veri düzeltmesi:
Kutu 1: 4, 4, 15
Kutu 2: 21, 21, 24
Kutu 3: 25, 25, 34

Bu örnekte kutulama 3 genişliğinde yapılmıştır. Kutu ortalaması yönteminde, her kutunun ortalama değeri bulunup; değerler bu ortalama değerlerle yer değiştirilir. Kutu sınırları ile veri düzeltimi yönteminde de her kutudaki değerler maksimum ve minimum değerlere en yakın olan değerle yer değiştirilir.

b.Kümeleme: Kümeleme ile aykırı değerler saptanabilir, burada benzer değerler kümeler veya gruplar içinde organize edilirler. Sezgisel olarak sınıf seti tarafından kapsanmayan değerler aykırı değer olarak düşünülebilir.



Şekil 2.4 Kümeleme [2]

c.Birleşik bilgisayar ve insan denetimi: Aykırı değerler, insan ve bilgisayar denetiminden geçirilerek bulunabilirler. Değişik karakter versiyonları ya da çöp (etiketlenemeyen karakterler) olarak tanımlanan veriler, örüntüler halinde bir listeye çıkış olarak verilebilirler. Bir insan gerçekte çöp olarak tanımlanan örüntüler listesinden birbirlerine yakın olanları seçebilir. Bu metot, bütün veri tabanı içindeki gözle aramadan daha hızlı bir yöntemdir.

d.Gerileme: Gerileme fonksiyonları gibi fonksiyonlarla veri uygunlaştırılarak düzeltilebilir. Doğrusal gerileme iki değişken arasındaki "en iyi" çizgiyi bulmayı gerektirir. Bu sebeple bir değişken diğerini tahmin için kullanılabilir. Çoklu doğrusal gerileme doğrusal gerilemenin genişletilmişidir. Çoklu doğrusal gerilemede ikiden fazla değişken gerekmekte ve veri bir yüzey üzerinden uydurulmaktadır. Veriyi uydurmak için kullanılan matematik eşitlikler bulunurken, gerileme kullanılması gürültüyü yok etmeye yardım eder [2].

2.4. Veri Madenciliği Teknikleri

Bu bölümde genel başlıklarla anlatılan veri madenciliğinde kullanılan yöntemlerin ana hatları gösterilmiş olup, bu yöntemlerde kullanılan algoritmalar alt başlıklar halinde verilmektedir.

Veri madenciliği teknikleri birbirinden kullandıkları veri yapıları ve örüntüleriyle farklılık gösterirler. Yapılan birçok çalışmada veri madenciliği tekniklerinin gösterimi birbirinden farklıdır. Burada en çok kullanılan gösterim yöntemi tercih edilmiştir. Veri madenciliği yöntemleri genel olarak üç ana başlık altında toplanabilir:

1.Sınıflandırma ve kestirim,

2. Kümeleme,

3. Birliktelik analizi.

2.4.1. Sınıflandırma ve Kestirim

Veritabanları iş süreçlerinde karar analizlerinde çok gizli verilere sahip olabilirler. Sınıflandırma ve kestirim veri analiz yöntemlerinden biridir. Fakat sınıflandırma ve kestirim birbirlerinden farklı olarak kendilerine farklı uygulamalarda yer bulacaklardır. Örneğin sınıflandırma bir finansal uygulamada risk veya güvenlik üzerine uygulama alanı bulurken, tahmin yürütme müşterilerin gelir ve meslek dağılımlarına göre potansiyelleri belirleyen bir uygulama alanında kullanılabilir.

Bir çok sınıflandırma ve tahmin yürütme yöntemleri makine öğrenmesine, uzman sistemlere ve istatistiki yöntemlere dayanır.

Veri madenciliğinde sınıflandırma yöntemi mevcut verilerin belirlenen kriterlere göre sınıflandırılmasından ve yeni eklenen her verinin daha önceden oluşturulmuş bu sınıflara dahil edilmesi işlemlerinden meydana gelmektedir.

Kestirim işlemini sınıflandırma işleminden ayıran fark ise “ bekle ve gör ” prensibidir. Kestirim işleminde sınıflandırma gelecek için tahmin edilen sınıfa ya da davranışa göre yapılır. Dolayısıyla sonucun doğru olup olmadığı “ bekle ve gör ” prensibiyle elde edilir.

Sınıflandırma yöntemiyle banka kredisi onaylama işlemleri, kredi kartı sahteciliği tespiti ve sigorta risk analizi gibi işlemler yapılabilirken; kestirim yöntemiyle deprem tahmini seyahat acentesi müşterilerinin önümüzdeki nerede tatil yapmak isteyecekleri gibi işlemler yapılabilir.

Sınıflandırma ve kestirim yöntemlerini karşılaştırmak için kullanabileceğimiz kriterler şunlardır:

a.Doğruluk Kestirimi : Verilerin sınıflandırılmasındaki yeteneği ifade eder.

b.Hız : Hesaplama performansı.

c.Sağlamlık : Verilen gürültülü veriler üzerindeki doğru sınıflandırma yeteneği.

d.Scalability : Ölçülebilirlik.

e.Açıklanabilirlik : Anlaşılırlık [2].

Sınıflandırma ve kestirim yöntemlerinin kendi içlerinde alt yöntemleri bulunmaktadır. Bunlar: Karar ağaçları, Bayesian sınıflandırması, Geri besleme (Backpropagation) ile sınıflandırma, K - en yakın komşu sınıflandırıcısı (K-nearest neighbor classifiers) Duruma sonuçlandırma (Case based reasoning), Genetik algoritmalar, Kaba küme yaklaşımı (rough set approach), Bulanık küme yaklaşımı (Fuzzy set approaches) [2].

2.4.2. Kümeleme

Kümeleme, fiziksel ya da soyut nesnelerin benzerliklerine göre gruplanmasıdır. Küme benzer nesnelerin oluşturduğu bir gruptur. Kümeleme işleminde temel prensip, küme içi benzerliği maksimum, kümeler arası benzerliği minimum yapmaktır. Bir kümeleme yönteminin kalitesi bu prensibi sağlaması ile doğru orantılıdır. Kümelemenin sınıflandırmadan farkı sınıflandırmadaki gibi önceden tanımlı sınıf etiketlerinin olmamasıdır. Bu sebeple kümelemede, sınıflandırmadaki gibi örnekleyerek öğrenme yerine gözlemleyerek öğrenme kavramı geçerlidir [2].
Veri kümeleme çok hızlı bir gelişim içindedir. Uygulama alanları hızlı bir şekilde genişlemektedir. Yıllar içinde analiz edilecek veri miktarı da sürekli arttığı için çok kullanılacak bir yöntemdir. Kümeleme yöntemiyle, pazarlamacıların kendi müşterileri arasındaki farklı grupları karakterize etmesi, yeryüzü incelemelerinde belli toprak parçalarının tanımlanması gibi konularda sonçlar elde edilebilir.

Veri madenciliği alanında kümeleme yapabilmek için bazı gereksinimlerin sağlanmış olması gerekir [2].

1.Ölçeklendirilebilme: Kümelendirme algoritması küçük çaplı nesneler üzerinde çalışabilmesine rağmen büyük veriler üzerinde çok performanslı olmayabilir. Bu durumlarda ölçeklendirme algoritmalarına ihtiyaç vardır.

2.Değişik Nesne Tiplerine Göre Çalışabilme: Günümüzde birçok kümelendirme algoritması sayısal veriler üzerinde çalışması için geliştirilmiştir. Ancak sayısal olmayan ve ikili veriler üzerinde de çalışacak algoritmalara ihtiyaç gittikçe artmaktadır.

3.Farklı Tipteki Nesneleri Ayırabilme: Birçok kümelendirme algoritması nesneler arasında Euclidean ve Manhattan ölçütlerine göre ayrım yapabilmektedir. Bu tür algoritmalar benzer boyuttaki ve benzer yoğunluktaki nesneleri ayırt edebilmektedir; fakat çok değişik tipte, boyutlarda nesneler olabileceğinden algoritmanın buna uygun olarak çalışması gerekmektedir.

4.En Az Miktarda Alan Bilgisi Gerektirmesi: Birçok kümeleme algoritması kullanıcı girişlerine ihtiyaç duyar. Kümeleme sonucu da bu parametrelere karşı hassastır ve bunlara göre değişiklik gösterir. Algoritma sonucu parametrelere bu kadar bağımlı olmamalı ve sonuç bu derece hassas olmamalıdır. Bu, parametreyi girecek kullanıcılar için büyük bir sıkıntıdır ve analizin sonucunu kontrol etmeyi zorlaştırır.

5.Çöp Veri Ayıklayabilme: Gerçek hayatta kullanılan birçok veritabanı eksik tanımlanmamış, ayrık veriler içerir. Kümelendirme algoritmaları bu çöp verilerden dolayı kötü sonuçlar verebilir. Bu sebeple, algoritma bu çöp verileri ayıklayabilmelidir.

6.Algoritma, Verilen Parametrelerin Sırasına Duyarsız Olmalıdır: Bazı algoritmalarda girilen parametrelerin sırası değiştiğinde algoritma sonucu bundan etkilenir. İstenmeyen bu durumun oluşmaması için, algoritmada girilen parametrelerin sırası önemsiz olmalıdır.

7.Yüksek Boyutluluk: Birçok algoritma 2 ya da 3 boyutlu veriler üzerinde iyi çalışır. İnsan gözü de en çok 3 boyutlu veriyi anlayabilecek yapıdadır. Fakat kümeleme algoritması daha fazla boyutta çalışabilmelidir.

8.Kısıtlama Bazlı Kümeleme: Günümüz ihtiyaçlarına cevap verebilecek bir algoritma çeşitli kısıtlamalarla çalışabilmelidir. Yani sonuca yansıyacak veriler filtrelenebilmelidir.

2.4.3. Birliktelik analizi

Birliktelik analizi ile büyük veri yığınları içerisinden ilginç olabilecek birlikteliklerin yakalanmasını amaçlar. Veri saklama oranındaki artışın gelişen endüstriye bağlı olarak çok hızlanması, bu endüstrilerin saklanan verilerden birliktelik kuralları türetmeye yönelmelerini sağlamıştır. Veri tabanlarındaki bu tür ilginç örüntülerin

yakalanması karar verme işlemlerinin daha etkin yapılmaya başlamalarını sağlamıştır. Örnek olarak katalog tasarımları, satış işlemleri, kar-zarar analizleri gibi [2].

Birliktelik analizinin en yaygın kullanıldığı alan sepet analizi uygulamalarıdır. Bu uygulamalarda genellikle müşteri davranışları, market stratejilerinin belirlenmesi, beraber satılan ürünlerin tespiti gibi örüntüler yakalanmaya çalışılır.

Birliktelik analizlerinde kullanılan iki parametre vardır: Destek (support) ve güven (confidence). Bu parametre aşağıdaki örnekte açıklanmıştır.

Bilgisayar ———— Finansal Yazılım (2.1)
(destek = %2, güven = %60)

Bu örneğe göre bilgisayar alanlar finansal yazılımda almaktadır. Kuralın destek ve güven değeri kuralın ilginçliğini ifade eder. Kuralın destek değerinin %2 olması analiz edilen tüm işlemlerin %2 'si bu iki ürünün birlikte alındığını gösterir. Güven değerinin %60 olması ise bilgisayar alan müşterilerin %60'ının yazılımda aldığını gösterir. Her iki değer içinde eşik değerleri vardır. Kural eşik değerini aşabilirse ilginç kabul edilir.

Birliktelik kuralları madenciliği iki aşamalıdır:

1.Tüm sık geçen nesne kümelerini bul: Tanıma göre her nesne kümesinin sık geçenler kümesinde yer alabilmesi için, her nesnesinin destek (support) değerinin önceden tanımlanmış olan min_sup değerinden büyük olması gerekir.

2.Sık geçen nesne kümelerinden güçlü ilişki kuralları yarat: Tanıma göre, bu kurallar min_sup ve min_conf durumunu sağlamalıdır [2].

Birliktelik analizinde kullanılan en bilinen algoritma Apriori algoritmasıdır. Apriori algoritması sık geçen birlikteliklerin yakalanmasında kullanılan temel bir algoritmadır.

2.5. Veri Madenciliğinin Uygulama Alanları

Veri madenciliği Bölüm 1.1'de de anlatıldığı gibi günümüzde birçok farklı disiplin içerisinde yer almakta ve karşılaşılan problemlere getirdiği çözümlerle daha çok kullanılır hale gelmektedir. Dolayısıyla veri madenciliği birçok alanın içine girmiş bulunmaktadır.

Bu bölümde veri madenciliğinin hangi alanlarda ve nasıl kullanıldığı anlatılacaktır. Veri madenciliği pazarlama, bankacılık ve finans, tıp, biyoloji, genetik, telekomünikasyon, endüstri ve mühendislik, kimya, yüzey analizi ve coğrafi bilgi sistemleri, görüntü tanıma ve robot görüş sistemleri, uzay bilimleri ve teknolojileri, meteoroloji ve atmosfer bilimleri, eğitim, sosyal ve davranış bilimleri, metin ve internet madenciliği gibi alanlarda kullanılmaktadır. Veri madenciliğinin kullanım alanları aşağıda açıklanmıştır.

Pazarlama: Bu alanda veri madenciliği; müşteri memnuniyetinin sağlanması için yeni pazarlama yöntemlerinin oluşturulmasında, satın alma hareketlerinin analizi, düzenlenen kampanyalara müşterilerin cevap verme oranlarının belirlenmesi, çapraz satış analizlerinde, tedarik ve mağaza yerleşim yöntemlerinin belirlenmesinde, satış tahmini ve analizlerinde kullanılmaktadır.

Bankacılık ve Finans: Finans alanında farklı göstergeler arsındaki ilişkiler veri madenciliği yöntemiyle incelenmektedir. Yine bu alanda risk analizleri, kredi kartı sahteciliğinin tespiti, kredi kartı kullanımlarına göre müşteri grupların belirlenmesi gibi konularda çalışmalar bulunmaktadır. Ayrıca sigortacılık alanında müşteri analizleri ve bu alandaki sahteciliklerin tespiti alım-satım analizleri gibi değerlendirmeler veri madenciliği yöntemleriyle yapılabilmektedir.

Bu alanda yapılan bir çalışmada müşterilerin kredi kartı verileri ele alınarak, bu müşterilerden hayat sigortası yaptırma potansiyeli olanların tespitine çalışılmıştır [16]. Burada müşterinin hangi yaş aralığında olduğu, kredi kartının olup olmadığı maaşının hangi aralıklar içerisinde olduğu, cinsiyeti, daha önceden hayat sigortasının olup olmadığı gibi verileri ele alınmıştır.

Tıp ve İlaç Sektörü,Biyoloji, Genetik: Veri madenciliği bu alanda hastaların kişisel ve laboratuvar verilerinin kullanılmasıyla, hastaya konulacak teşhis ve tedavi yöntemleri, ürünlerin geliştirilmesi ve tahlil sonuçlarının tahmini, bitki türlerinin ıslahı, gen haritalarının analizi gibi konularda kullanılmaktadır.

Bu alanda yapılan çalışmalardan birinde kadınların göğüs kanseri olma riskleri incelenmiştir. Bu çalışma veri madenciliğinde genetik algoritma tekniği kullanılarak yapılmıştır. Çalışmanın sonucunda bu tür kanserde erken teşhis olanağı sağlanmış ve doğru sonuçlar ürettiği görülmüştür [14].

Tıp alanında yapılan başka bir çalışmada da genetik özelliklerin ve çevresel faktörlerin obezite ve diabet hastalığı gibi birden fazla faktöre bağlı olan hastalıklar üzerindeki etkisi incelenmiştir. Deneyler Lille Biyolojik Enstitüsünün verileriyle yapılmıştır. Çok büyük sayıda veri olduğundan, keşifsel (heuristic) yaklaşım seçilmiştir.

İlk aşamada nitelik seçimi için genetik algoritma kullanılmıştır. Bu çok spesifik problemi bölmek için genetik algoritmada tanımlanmış bazı mekanizmalar kullanılmıştır; paylaşım, rastgele yerleştirme(göç), genetik operatörler gibi. İkinci aşamada, bir önceki aşamada seçilen niteliklerin sınıflandırılmasına çalışılmıştır. Bunun içinde en populer sınıflandırma algoritması olan k-means kullanılmıştır [21].

Telekomünikasyon: Telekomünikasyon alanında veri madenciliği kalite ve iyileştirme analizlerinde, hatların yoğunluk tahminlerinde ve ağ kurulumlarında kullanılmaktadır.

Endüstri ve Mühendislik: Kalite kontrol analizleri, lojistik, üretim süreçlerinin eniyilenmesinde, veriler üzerinde modeller kurarak bilimsel ve teknik problemlerin çözümünde veri madenciliği teknikleri kullanılmaktadır.

Kimya: Yeni moleküllerin keşfi ve sınıflandırılması, ilaçların yapımında veri madenciliği kullanılmaktadır.

Yüzey Analizi ve Coğrafi Bilgi Sistemleri: Para makinelerinin dağılımı, yüzey şekillerinin belirlenmesi, yerleşim alanların tespiti, suç oranı, köken tespiti, posta ve otobüs duraklarının yerleştirilmesi gibi konularda veri madenciliği kullanılmaktadır.

Görüntü Tanıma ve Robot Görüş Sistemleri: Bu alanda robotların karşılaştıkları engelleri aşabilmeleri sağlanmaya çalışılmış; yüz, parmak izi ve yol tanıma gibi konularda veri madenciliğine başvurulmuştur.

Robotlarla ilgili yapılan uygulamalarda, başarılı çözüm teknikleri gerektiren birçok problem meydana gelir. Evrimsel Hesaplama, bu durumda çıkabilecek problemlerin bir kısmında başarı getiren bir yöntemdir.

Evrimsel metotlar ceşitli akıllı robot mimarilerinde uygulanmışlardır. Örneğin evrimsel algoritmalar kural tabanlı otonom ajanların kural kümelerinin öğrenilmesinde, robot kontrolü için kullanılan sinir ağlarının ağırlıklarının ve topolojisinin öğrenilmesinde, bulanık mantık kontrol sistemlerinde ve davranış tabanlı robotların kurallarında kullanılmaktadır [23].

Uzay, Meteoroloji ve Atmosfer Bilimleri: Gezegen yüzey şeklilerinin tespiti, gezegenlerin yerleşimi ve keşfi, yıldızların gruplandırılması, bölgesel iklim ve yağış haritalarının çıkartılması, hava tahminleri, ozon tabakasında oluşabilecek deliklerin tespiti, okyanus hareketlerinin incelenmesi gibi konularda veri madenciliği kullanılmıştır.

Eğitim: Bu alanda çok çeşitli çalışmalar yapılmıştır. Ders programı hazırlanması, kursların iyileştirilmesi, öğrenci davranışlarının tespiti, derslerin birbiriyle olan etkileşimi ve bu derslerde öğrencilerin başarısı gibi konularda çalışmalar yapılmıştır.

Bu konuda yapılan bir çalışmada, öğrenci bilgileri kullanılarak farklı kestirim kurallarının nasıl bulunabileceği gösterilmeye çalışılmış ve bunlar kullanılarak web üzerinden yapılan kurslarda düzeltilmesi gereken noktalar aranmıştır. Eğitim alanındaki gücünü arttırmak için AHA'da birçok değişiklikler yapılmıştır. Kullanılabilecek veriler arasındaki ilişkiyi bulmak için AHA 'da kütükte tutulmuş bilgiler kullanılmıştır (okuma zamanları, seviye zorlukları ve test sonuçları). En ilgi çekici olan ilişkiler öğretmene gösterilmiştir. Böylece kursun daha verimli olabilmesi için gereken değişikliklerin farkedilebilirliği kolaylaştırılmıştır [17].

Eğitim alanında yapılan bir başka çalışmada da {alındı, alınmadı} ya da {var, yok} şekilde ikili değerler dışında kategorik ve nicel değerler de içeren veri tabanlarında birliktelik kurallarının keşfi için yapay zeka ve zeki hesaplama tekniklerinden genetik algoritma bulanık mantık tabanlı etkili, yeni bir yöntem geliştirilmiştir. Genetik algoritmalarda başlangıç popülasyonunu gelişigüzel üretmek yerine, bunu çözüm uzayına düzgün dağıtan düzenli popülasyon yöntemi kullanılmıştır. Genelde kullanılan yöntemlerin aksine yüksek destek ve güven değerlerine sahip birliktelik kuralları yoğun nesne kümeleri üretilmeden doğrudan ve her veri tabanı için belirlenmesi güç olan minimum güven ve minimum destek eşiklerine ihtiyaç duyulmadan keşfedilmiştir. İlginç birliktelik kurallarını bulmak için uyarlamalı mutasyon ve elitzim stratejisi uygulanmıştır. Bu şekilde genetik algoritmanın son popülasyonu ilginç birliktelik kurallarını temsil etmiştir. Önerilen yöntem hem yapay bir veri tabanında hem de Fırat Üniversitesi Elektrik-Elektronik Mühendisliği lisans öğrencilerinin ders not kayıtlarında denenmiş kullanışlı ve ilginç kurallar etkili şekilde bulunmuştur [13].

Sosyal ve Davranış Bilimleri: Seçimlerde öngörülerde bulunabilmesi için, kamuoyu yoklamaları ve genel eğilimlerin belirlenmesi gibi istatistiki bilgilerin elde edilebilmesi için veri madenciliği kullanılmaktadır.

Metin ve İnternet Madenciliği: Çok büyük ve anlamsız metin yığınları arasından anlamlı ilişkilerin tespiti metin madenciliğinin konusudur. İnternet madenciliğinde ise internetin belli kategorilere ayrılarak veriye ulaşım süresinin azaltılması amaçlanmaktadır.

3. GENETİK ALGORİTMALAR

3.1. Genetik Algoritmalar Nasıl Ortaya Çıkmıştır?

1950 ve 1960' larda bilgisayar bilimcilerin bir kısmı, mühendislik problemlerinde bir eniyileme aracı olarak kullanılabileceğini düşündükleri evrimsel sistemler üzerinde çalışmalar yapmışlardır. Bu tarihler arasında birçok araştırmacı evrimsel algoritmaları eniyileme ve makine öğrenmesi konusunda geliştirmişlerdir (Box-1957, Friedman-1959 Bledsoe-1961 Bremermann-1962 ve Redd, Toombs, Baricelli-1967) [3].

Bununla birlikte genetik algoritmaları ilk ortaya çıkaran Michigan Üniversitesi'nden psikoloji ve bilgisayar bilimi uzmanı John Holland'dır. Holland'ın geliştirdiği genetik algoritma mantığı, Darwin'nin evrim teorisine dayanmaktadır. Evrim teorisine göre doğada en iyi uyumu yakalayan birey yaşam hakkı kazanır. John Holland' da, 'en iyi uyumu sağlayan' birey prensibini ele alarak; oluşturduğu popülasyonda evrim kurallarını işletip genetik algoritma çalışmalarına başlamıştır. Holland'ın çalışmasındaki en önemli nokta ise daha önce geliştirilen genetik algoritmalar gibi algoritmayı problemlere özel tanımlamamış olmasıdır [3]. Mekanik öğrenme (Machine Learning) üzerine çalışan John Holland, bilgisayar ortamına aktararak uyguladığı evrim teorisinin prensiblerini; 1975 yılında kitap haline getirmiştir. 'Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence' adlı kitabında geliştirdiği genetik algoritmaları ve genetik operatörlerini ele almıştır. 1975' te John Holland'ın yaptığı çalışmaların ardında genetik algoritmaların önemi anlaşılamamış olup; farklı alanlarda geliştirilmek üzere adımlar atılmamıştır. 1985 yılında David E. Goldberg'in doktora tezinde gaz boru hatlarının denetimiyle ilgili yaptığı çalışmayla birlikte genetik algoritmaların farklı konulardaki kullanılabilirliği ve ürettiği etkin çözümler görülmüştür. Bu çalışmasıyla David E. Goldberg 1985 National Science Foundation Genç Araştırmacı ödülünü kazanmıştır. David E. Goldberg'in 1989

yılında yayınlanan ‘ Genetic Algorithms in Search, Optimization and Machine Learning ’ adlı kitabı, genetik algoritmalarla ilgili ön yargıyı ortadan kaldırmış kitabında birçok uygulamaya yer vererek genetik algoritmaların bir çok konuda kullanılabildiğini göstermiştir.

Son yıllarda yapılan çalışmalarda kullanılan genetik algoritmalar, Holland'ın tanımladığı algoritmadan birçok konuda farklılık göstermektedir [3]. Günümüzde de, popülaritesini sürdürmeye devam eden genetik algoritmalar; optimizasyon, makine öğrenmesi, otomatik programlama ve bilgi sistemleri, ekonomik ve sosyal sistem modelleri gibi bir çok alanda kullanılmakta olup; genetik algoritmaların bir çok algoritmaya göre etkin ve verimli çözümler ürettiği ortaya konmuştur.

Yapılan bir çalışmada sınıflandırma tekniği olan k-means ile iki farklı genetik algoritma yaklaşımı örnek bir veri kümesi kullanılarak karşılaştırılmıştır. İki teknik arasında yapılan karşılaştırmada genetik algoritmayla elde edilen sonuçların daha iyi olduğu görülmüştür [6].

Genetik algoritma yöntemleri son dönemlerde oldukça yaygınlaşmış eniyileme yöntemlerindendir. Bu çalışmada genetik algoritma metotları kullanılarak ÇDH (Çoklu Dizi Hizalama) probleminin çözülmesi ve şimdiye kadar yapılmış olan çalışmalardaki güçlüklerin aşılmasını sağlayacak ve uzmanlara üzerinde yorum yapılmak üzere alternatif iyi hizalamalar sunacak bir yöntemin geliştirilmesi amaçlanmıştır [7].

Genetik algoritmalarla yapılan bir başka çalışmada da, bulanık birliktelik kurallarının madenciliği için otomatik metot geliştirilmeye çalışılmıştır. Bu amaçla, ilk önce sınıflandırma temelli olan genetik algoritmalar; ikinci olarak da literatürde çok bahsedilen ve etkili bir yöntem olan örneklemeyle sınıflandırma yöntemi kullanılmıştır. Genetik algoritma tabanlı yaklaşımla, literatürde geçen diğer yaklaşımlar karşılaştırılmıştır. Deneylerde, genetik algoritmayla bulunan ilginç kuralların sayısının diğer metotlara göre daha fazla olduğu görülmüştür [8].

3.2. Genetik Algoritmaların Tanımı ve Çalışma Şekli

3.2.1. Genetik algoritmaların temel yapısı

Genetik algoritmalar, bir önceki bölümde de bahsedildiği gibi Darwin' in evrim teorisi temel alınarak yapılandırılmıştır. Amaç oluşturulan popülasyonda en iyi uyumu yakalayan bireyi elde edebilmektir.

Bu bölümde, 'en iyi uyumu sağlamış birey' ile neyin tanımlandığı, başlangıç popülasyonunun oluşturulması, genetik algoritmaların çalışma mantığı ve operatörleri, hangi parametrelerin esas alındığı gibi algoritmanın işleyiş aşamalarından bahsedilecektir.
Genetik algoritmaların üzerinde çalıştığı popülasyon kromozomlar halinde tanımlanmış dizi şeklindeki bireylerden oluşturulmaktadır. Genetik algoritmaların kullanıldığı problemlere göre popülasyonu oluşturan bireylerin yapısı değişmektedir. Popülasyondaki bireyler ne kadar iyi tanımlanırsa problemin sonuçları o kadar verimli olmaktadır.

Başlangıç popülasyonunun oluşturulması bilinen iki yöntem vardır :

1.Bir sayı üretici tarafından rasgele üretilmiş olası çözümler.

2.Problem için belirli koşul ve kısıtları karşılayan olası çözümler [9].

Bu aşamada bireylerin üzerinde çalışılan problem için, ne kadar uygun çözüm oldukları, genetik algoritmaların yapısında önemli bir işlevi olan uygunluk fonksiyonu sayesinde anlaşılmaktadır. Uygunluk fonksiyonu, oluşturulan popülasyonun içerisindeki her bir bireyinin problem için uygun bir çözüm olup olamayacağı değerlendirilmesi aşamasında her bir birey için hesaplanan değerdir. Uygunluk fonksiyonunun seçimi probleme özel olarak hesaplanır. Her problem için aynı fonksiyonun kullanılması, bireylerin değerlendirilmesinde iyi sonuçlar üretemeyebilir. Algoritmanın bu aşamasına kadar algoritmanın üzerinde

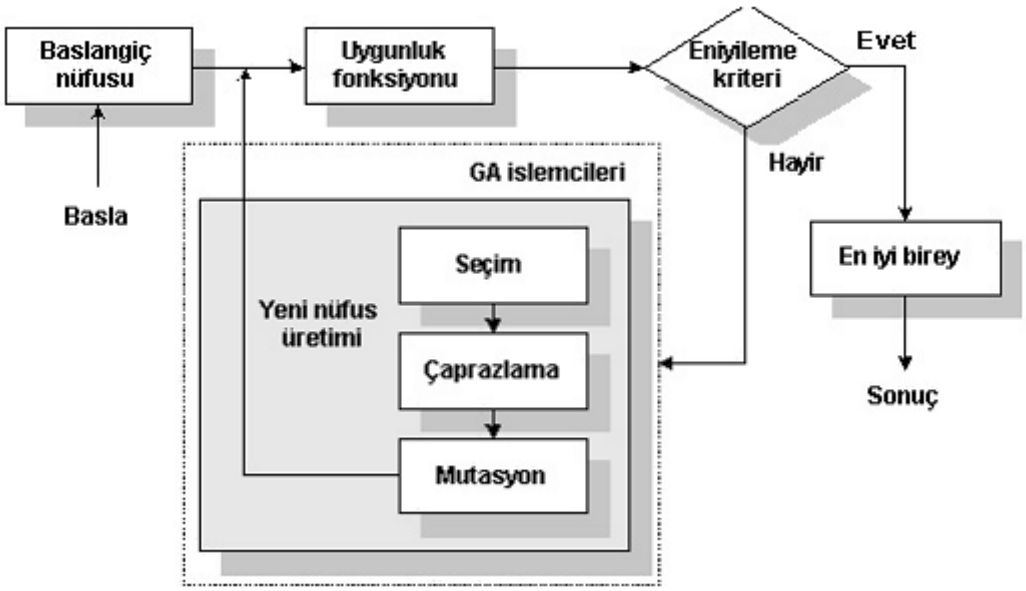
çalıştırılacağı popülasyon bireyleri oluşturulmuş ve bu bireylerin çözüm için uygunluğuna bakılmıştır.

Bireylerin uygunluk fonksiyon değerleri bu bireylere genetik işlemlerin uygulanıp uygulanmayacağı konusunda yardımcı olur. Belirlenen uygunluk fonksiyonu değerini geçemeyen bireyler genetik operatörler uygulanmak üzere tutulur. Genetik algoritmaların operatörleri işlem sırasıyla seçim yöntemleri, çaprazlama ve mutasyondan oluşmaktadır.

Genetik algoritmaların temelini oluşturan çaprazlama ve mutasyon operatörleri ile esas olan uygunluk fonksiyon değerleri düşük olan bireylerin uyumluluğunu arttırabilmektir. Genetik operatörlerin uygulanması için seçilen bireylere, öncelikle çaprazlama işlemi yapılır. Çaprazlama işleminde oluşturulacak yeni nesil için uygun görülen bir seçim yöntemiyle ebeveynler seçilir. Çaprazlama seçilen ebeveynlerin kromozom yapılarının karşılıklı olarak kırılan noktalardan değiştirilmesidir. Böylece yeni nesil elde edilir. Genetik algoritmalarda çaprazlama işlemi popülasyondaki her bireye uygulanmaz. Belirlenen çaprazlama oranına – popülasyonda kaç bireye çaprazlama işleminin uygulanacağını gösteren parametredir - göre popülasyondaki belli sayıdaki bireylere uygulanır. Algoritmanın çaprazlama adımı geçildikten sonra, popülasyondaki tüm bireylere mutasyon işlemi uygulanır. Mutasyon işleminde bireylerin kromozom yapılarındaki belli sayıdaki – kaç genin mutasyona uğratılacağı seçilmiş olan mutasyon oranına göre belirlenir - genlerin yapıları kodlama türüne bağlı olarak değiştirilir.

Bu aşamaları geçen bireyler, uygunluk fonksiyon değerini ilk aşamada geçen bireylerle beraber yeni popülasyonu oluştururlar. Algoritmanın sonraki adımında sonlandırma koşulunun sağlanıp sağlanmadığı kontrol edilir. Sonlandırma koşulu olarak genelde tercih edilen iki yöntem vardır. İlk yöntemde algoritma çalıştırılmadan önce bir maksimum iterasyon sayısı seçilir, algoritma bu sayıya ulaşıldığından sonlandırılır. İkinci yöntemde de kullanıcı tarafından belirlenen koşul ya da koşullar yakalandığında algoritma sonlandırılır.

Genetik algoritmaların genel olarak kullanılan yapısı bu adımlardan oluşmaktadır. İşlemlerin sıralaması Şekil 3.1‘de gösterilmiştir.



Şekil 3.1 Genetik algoritmanın adımları [9]

Genetik algoritmaların şemasında da görüldüğü gibi algoritmanın gidişatını etkileyen birçok unsur bulunmaktadır. Dolayısıyla genetik algoritmaların yapılandırılmasında şunlara dikkat edilmesi gerekir:

a.Bireylerin gösterimi doğru bir şekilde yapılmalı,

b.Uygunluk fonksiyonu etkin bir şekilde oluşturulmalı,

c.Doğru genetik işlemciler seçilmeli [9].

Doğru çözümlere ulaşabilmek için genetik algoritmaların oluşturulma yöntemlerine ve probleme uygun operatör seçimine dikkat edilmelidir.

Genetik algoritmalar hızlı bir arama yöntemi olmanın yanı sıra, iyi bir eniyileme yöntemidirler. Peki bu özellikler genetik algoritmaların hangi özelliklerinden kaynaklanır?

Genetik algoritmaların geleneksel arama ve eniyileme yöntemlerinden dört temel farkı vardır:

1.Genetik algoritmalar problemlerin çözümünü parametrelerin değerleriyle değil kodlarıyla arar. Parametreler kodlanabildiği sürece çözüm üretebilirler. Bu sebeple genetik algoritmalar ne yaptığı konusunda bilgi içermez, nasıl yaptığını bilir.

2.Genetik algoritmalar aramaya tek bir noktadan değil, noktalar kümesinden başlar. Bu nedenle çoğunlukla yerel en iyi çözümde sıkışıp kalmazlar.

3.Genetik algoritmalar türev yerine uygunluk fonksiyonunun değerini kullanır. Bu değerin kullanılması ayrıca yardımcı bir bilginin kullanılmasını gerektirmez.

4.Genetik algoritmalar gerekirci kuralları değil olasılıksal kuralları kullanır [9].

3.2.2. Genetik kodlama yöntemleri ve başlangıç popülasyonu

3.2.2.1. Genetik kodlama yöntemleri

Bir problemin çözümü için genetik algoritma geliştirmenin ilk adımı, tüm çözümlerin aynı boyutlara sahip bitler dizisi biçiminde gösterilmesidir. Dizilerden her biri, problemin olası çözümler uzayındaki rastsal bir noktayı simgeler. Parametrelerin kodlanması, probleme özgü bilgilerin genetik algoritmanın kullanacağı şekle çevrilmesine olanak tanır [10]. Bu dizilerin her biri kromozom denilen parçalardan oluşmaktadır. Kromozomlar temsil ettiği çözümle ilgili bilgiler içermektedir.

Genetik algoritmalardan birey kodlama yöntemi olarak en çok ikili kodlama tercih edilirken genetik algoritmaların birçok probleme uygulanmaya başlanması ve farklı verilerin kullanılması gibi nedenler birey kodlamalarında farklı yöntemlerin geliştirilmesi ihtiyacını ortaya çıkarmıştır.

Genetik algoritmalarda kullanılan kodlama teknikleri şunlardır: İkili (binary) kodlama permütasyon kodlama, değer kodlama, ağaç kodlama.

1.İkili (binary) Kodlama: Bu yöntemde bireyler ikilik düzende temsil edilirler. Kromozomlar bütün olarak ya da parça parça gruplar halinde bir çözümü ya da değeri temsil edebilir. İkili kodlama, genellikle eniyileme problemleri için kullanılır. İkili kodlama basit, hızlı ve daha kolay bir biçimde işlenebilmesine rağmen; gerçel sayılara dönüştürme işlemi sırasında yapılan hesaplamalardan dolayı zaman kaybına neden olmaktadır.

Birey 1	00101010100011111001111001010101 0

Şekil 3.2 İkili kodlanmış birey

İkili sayıları kullanılarak yapılan Gray kodlama literatürdeki alternatif ikili kodlama yöntemlerinden biridir. Yukarıdaki kodlama yönteminden tek farkı bir sayı arttırılırken ya da azaltılırken tek bir bit değiştirilir. Gray kodlama bu şekilde en düşük ağırlıklı bitin değişimiyle devam eder. Bu gösterimde problemi bireyler içinde kodlamaktan daha çok gerçel sayıya dönüştürme işlemi sırasında yapılan hesaplamalar önemlidir. Bu hesaplamaları yapmak için ayrıca bir algoritma kullanılmalıdır. Bu tür kodlamanın dezavantajı olarak bu gösterilmektedir [9].

İkili kodlama kullanılırken parametrelerin kodlanmasında kullanılacak bit sayısının belirlenmesi önemli bir noktadır. Her bir parametre çözüm uzayındaki olası tüm çözümleri karşılayacak en iyi bit sayısıyla kodlanmalıdır. Çok az veya çok fazla kullanılan bit sayısı genetik algoritmanın performansını ters yönde etkileyebilir [9].

2.Permütasyon Kodlama: Tam sayı kodlama ilk defa bu tür veriler üzerine çalışanlar tarafından ortaya atılmıştır. Bu tür kodlamada kromozom değerleri sabittir. Permütasyon kodlama ya da diğer adıyla tam sayı kodlama yöntemi genellikle sıralama problemlerinde kullanılır. Permütasyon kodlama daha çok çizelgeleme, kesme, paketleme ve gezgin satıcı problemleri için kullanılır. Permütasyon kodlamada kodlama ve kod çözme işlemlerine gerek kalmadığı için kazandırdığı zaman açısından avantaj sağlar.

Birey 1	32844903821110986547219121123645 3

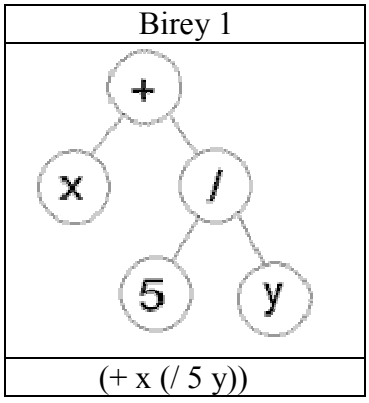
Şekil 3.3 Permütasyon kodlamayla oluşturulmuş birey

3.Değer Kodlama: Değer kodlama yöntemi, karmaşık değerler içeren problemlerde ikili kodlama yöntemi zor olduğu için tercih edilir. Bu tip kodlama yönteminde kromozomlar reel sayılar olabileceği gibi, harfler ya da farklı objelerde olabilir. Değer kodlama yöntemi de çizelgeleme, kesme, paketleme ve gezgin satıcı problemleri için kullanılabilen bir kodlama türüdür.

Birey 1	1.45654 4.234 9.45435 0.34324 3.5609
Birey 2	DFEKKLTOBNSAEENBKFŞDOKRL
Birey 3	(back), (right), (forward), (left), (left)

Şekil 3.4 Değer kodlamayla oluşturulmuş birey

4.Ağaç Kodlama: Ağaç kodlama yöntemi daha çok değişen programlar ya da ifadeler içeren problemlerin çözümlerinde kullanılır. Eğer değerler ağaç yapısıyla ifade edilebiliyorsa, değişen problemler ve yapılar için ağaç kodlama yöntemi oldukça kullanışlıdır.



Şekil 3.5 Ağaç kodlamayla oluşturulmuş birey

3.2.2.2. Başlangıç popülasyonu

Genetik algoritmalarda başlangıç popülasyonu, ele alınan problemin çözümü olabilecek kuralların belirlenen kodlama yöntemiyle oluşturulan bireylerden elde edilen çözüm grubudur.

Genetik algoritmalarda başlangıç popülasyonunun büyüklüğü çözümün elde edilmesinde önemli parametrelerdendir. Başlangıç popülasyonunun büyüklüğünün algoritmanın verimliliğine etkileri üzerine yapılan çalışmalarda şu sonuçlar elde edilmiştir:

1975 yılında De Jong tarafından yapılan çalışmada popülasyon büyüklüğünün 50-100 birey arasında, çaprazlama oranının yaklaşık 0.6, mutasyon oranının da 0.001 seçilmesi halinde genetik algoritmaların en ideal sonuçları verdiği görülmüştür.

Yine 1986 yılında Grefenstette tarafından genetik algoritmalarda parametrelerin etkisi incelendiğinde ise popülasyon büyüklüğünün 30, çaprazlama oranının 0.95, mutasyon oranının da 0.01 seçilmesinin daha verimli sonuçlar ortaya çıkardığı belirtilmektedir.

1989'da Schaffer, Caruana, Eshelman ve Das tarafından yapılan çalışmalarda ise Grefenstette'nin 1986' da elde ettiği sonuçlara yakın olmakla beraber şu sonuçlar elde edilmiştir. Popülasyon büyüklüğü 20-30, çaprazlama oranı 0.75 – 0.95, mutasyon oranı 0.005 – 0.01 arasındaki değerlerden seçilmesinin uygun olacağı ortaya atılmıştır [3].

3.2.3 Genetik algoritma operatörleri ve parametreleri

3.2.3.1 Genetik operatörler

Genetik algoritmaların operatörleri çaprazlama, mutasyon ve seçimdir. Bu operatörler kendi içlerinde en iyi çözüme ulaşılabilmesi için çeşitlendirilmiştir.
1.Çaprazlama: Çaprazlama operatörü, iki ebeveyn arasında kromozom değişimi olarak adlandırılabilir. Kromozom değişimleri belli yerinden kırılan ebeveynlerde bu kromozom parçacıklarının yer değiştirmesiyle olur.

Çaprazlama operatörü temelinde basit bir işlem gibi gözükse de bu işleme başlanmadan önce belirlenmesi gereken bazı kurallar vardır. Bu kurallar çaprazlama

operatörünün nasıl yapılacağı, ne kadar bireye uygulanacağı ve uygulanacak olan ebeveynlerin nasıl seçileceğinin belirlenmesidir.

Öncelikle popülasyondaki kaç bireye çaprazlama işlemi uygulanacağının belirlenmesi gerekir. Bu sayı çaprazlama oranı dediğimiz, kullanıcı tarafından belirlenen bir parametre ile belirlenir. Kullanıcı çaprazlama oranıyla, popülasyondaki bireylerin kaçta kaçına çaprazlama işlemi uygulanacağını belirler ve popülasyondan rasgele seçilen bu bireyler çaprazlama işlemi uygulanmak üzere ayrılır. Bireylerin seçimi tamamlandıktan sonra ebeveyn seçimi yapılması gerekir. Sonuçta çaprazlama işlemi iki birey arasında olmaktadır ve ebeveyn seçimi içinde farklı seçim yöntemleri bulunmaktadır. Ebeveyn seçim yöntemleri problemlere uygun olarak seçilmelidir.

Çok nüfuslu kararlı hal genetik algoritması kullanılarak yapılan ders programı hazırlamayla ilgili bir çalışmada, ebeveyn seçimi sıralama stratejisi kullanılmıştır. En iyi uygunluk değerine sahip olan bireyin seçilme ihtimali, en kötü olana göre n kare daha fazla olarak ayarlanmıştır [11].

Yine bu çalışmada ebeveyn seçim yöntemi olarak ‘ Rulet Tekerleği Yöntemi ’ belirlenmiş olup; bu yöntemde de ebeveynler uygunluk fonksiyon değeri yüksek olan bireylerin seçilme olasılığının fazla olması ihtimaliyle göz önüne alınarak işlem yapılmaktadır.

Sonuç olarak genetik operatörlere gereksinim duyulmasının nedeni uygunluk fonksiyon değeri daha iyi olan bireyler elde etme çabasıdır. Bu da yine genetik işlemlerin uygulanacağı kötü bireyler arasından en iyilerinin seçilmesi gerekliliğini ortaya çıkarmaktadır.

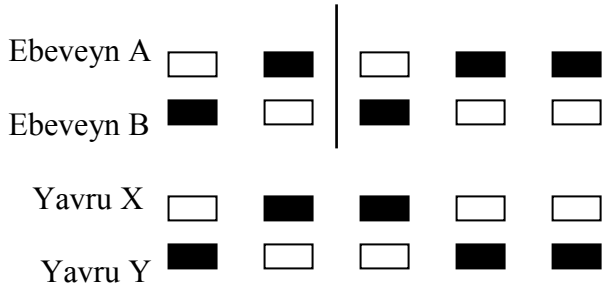
Çaprazlama işleminin iki amacı vardır:

1.Varolan bireylerin genetik özelliklerinin birleşimiyle arama uzayında yeni bireyler, başka bir deyişle yeni çözümler elde etmek.

2.Düşük uygunluk fonksiyonu değerine sahip bireylerin “en iyi olanın hayatta kalma” ilkesine dayanarak popülasyon içerisinden elenmesi [9].

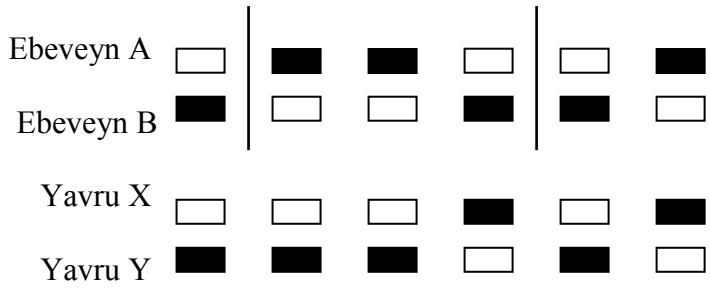
Çaprazlama işlemi kendi içinde de bazı yöntem farklılıkları içermektedir. Çaprazlama işleminin çeşitleri şunlardır:

a.Tek Noktalı Çaprazlama: Bu çaprazlama yöntemi en temel ve en basit çaprazlama yöntemidir. Bu yöntemde ebeveynlerin kromozomları rasgele seçilmiş tek bir noktadan kırılır. Bu kırılma noktasından sonraki kısımları ebeveynler arasında değiş tokuş edilmektedir.



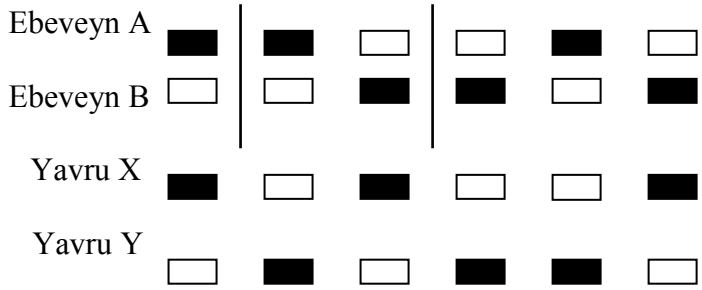
Şekil 3.6 Tek noktalı çaprazlama

b.İki Noktalı Çaprazlama: İki noktalı çaprazlama yönteminde ebeveynler iki noktadan kırılırlar. İki noktalı çaprazlama yönteminin, tek noktalı çaprazlama yöntemine göre daha etkin bir çözüm olduğu söylenebilir. Bu çaprazlama yönteminde iki kesim noktası arasında kalan bölümler ebeveynler arasında karşılıklı olarak değiştirilerek yeni bireyler oluşturulur.



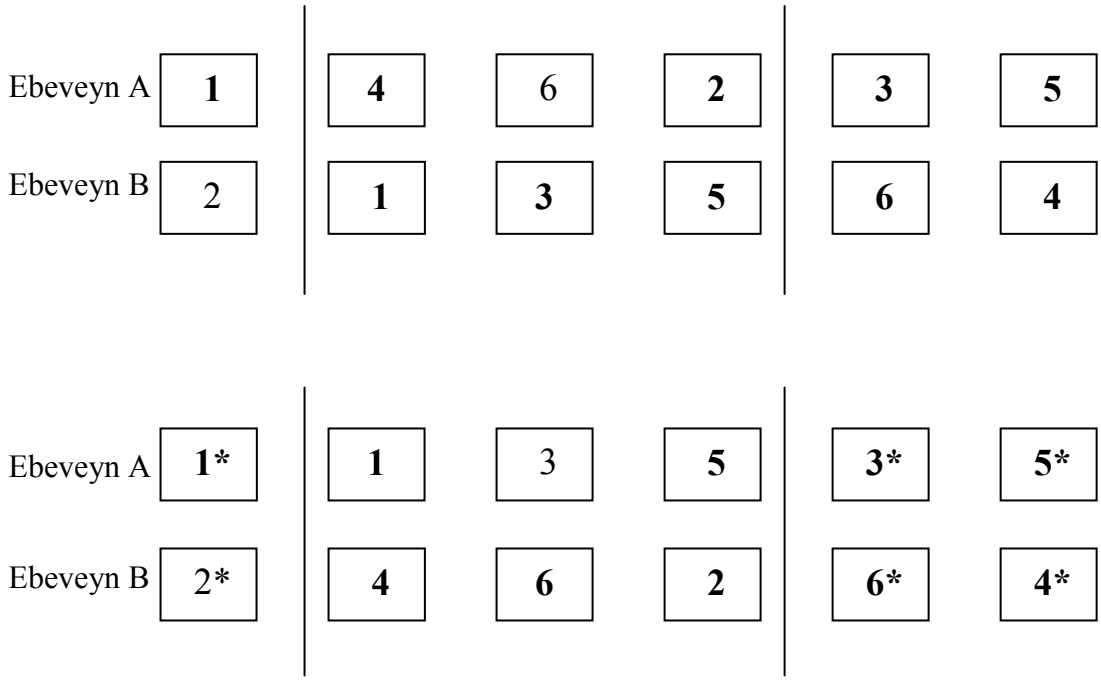
Şekil 3.7 İki noktalı çaprazlama

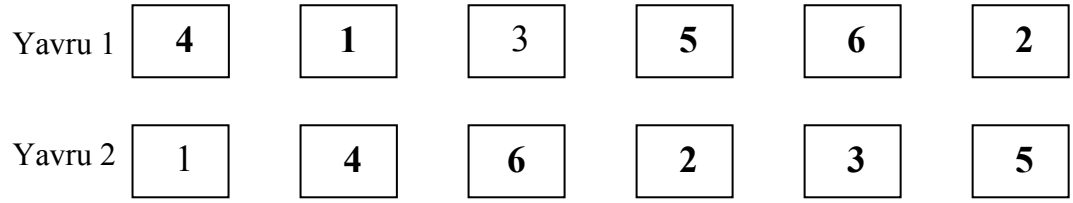
c.Nokta Sayısına Göre Çaprazlama: Bu tip çaprazlama yönteminin tek ya da iki noktalı çaprazlama yöntemlerine uygulanışı benzerdir. Bu tür çaprazlama yönteminde çaprazlama noktası rasgele belirlenmektedir. Kromozomların çaprazlanması işlemi kırılma noktalarından bölümler birer atlatılarak yapılmaktadır.



Şekil 3.8 Nokta sayısına göre çaprazlama

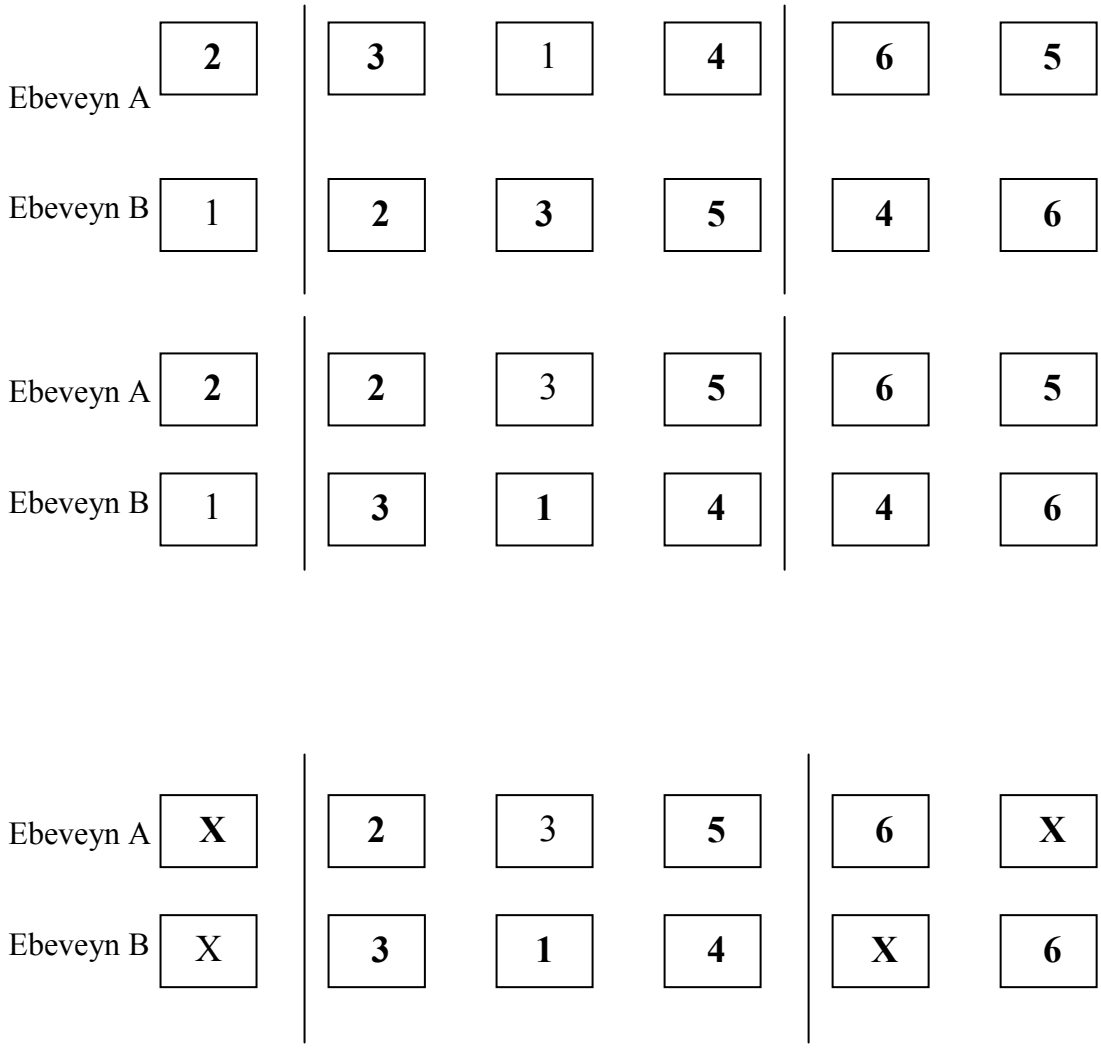
d.Kısmi Planlı Çok Noktalı Çaprazlama: Bu çaprazlama yöntemi Goldberg tarafından geliştirilmiştir. Çaprazlama işleminde kırılan noktalardan kromozom parçacıkları karşılıklı olarak değiştirilmektedir. Şekil 3.9 'da da görüldüğü gibi çaprazlama işlemi farklı ilerlemektedir.

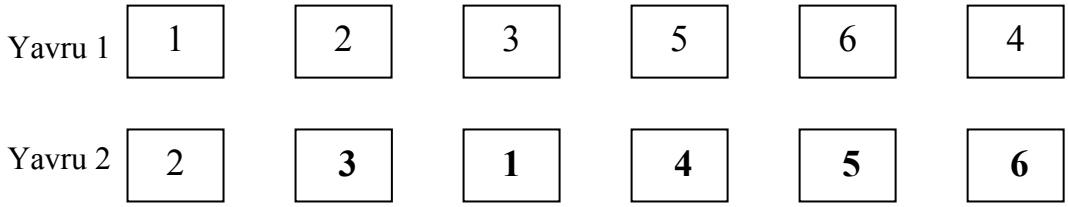




Şekil 3.9 Kısmi planlı çok noktalı çaprazlama

e.Sıralı Çok Noktalı Çaprazlama: Sıralı çok noktalı çaprazlama yönteminde bireyler iki noktadan kırılırlar. Kırılma işleminde dikkat edilmesi gereken nokta kırılma bölgelerindeki kromozom sayılarının eşit olması gerektiğidir. Kırılma noktaları arasındaki kromozomlar karşılıklı olarak yer değiştirilir; bu bölgenin dışında kalan bölgelerde tekrarlayan kromozomlar varsa; bu kromozomlar dışındaki kromozomlar yer değiştirilir. Bu yöntem şekil 3.10'da verilen örnekte gösterilmiştir.





Şekil 3.10 Sıralı çok noktalı çaprazlama

Genetik algoritmalarda tam sayı değerler için geliştirilmiş özel çaprazlama yöntemleri de şunlardır:

f.Sıralamaya Dayalı Çaprazlama: Sıralamaya dayalı çaprazlama yönteminde ebeveynlerin kesim noktaları rasgele belirlenir. Bu yöntemde ilk önce babadan kesim noktasına kadar olan kromozomlar yeni bireye kopyalanır, daha sonra anneden sırasıyla kromozomlar yeni bireye kopyalanır. Şekil 3.11'de bu çaprazlama yöntemine örnek olarak verilebilir.

Ebeveyn A	**1 2 3 4** 5 6 7 8 9 10
Yeni Birey	**1 2 3 4** 9 10 6 8 7 5
Ebeveyn B	**9 10** 1 3 2 4 **6 8 7 5**

Şekil 3.11 Sıralamaya dayalı çaprazlama

g.Devirli Çaprazlama: Devir çaprazlama yöntemi diğer yöntemlere göre daha karmaşık bir yöntemdir. Çaprazlama işlemi önce babadan başlar. Babanın ilk kromozomu yeni bireye kopyalanır, daha sonra bu gene annede karşılık gelen kromozoma bakılır bu gen babanın kromozomlarında bulunur ve yeni bireye babanın sıralamasından kopyalanır. Bu bir devirdir ve ilk kopyalanan gene annede karşılaşılıncaya kadar devam edilir. Çaprazlama işlemi bu sırada devam eder.

Ebeveyn A	**1** 2 3 4 5 6 **7** 8 **9** 10
Yeni Birey	**1** 10 3 4 2 6 7 8 9 5
Ebeveyn B	9 10 1 3 2 4 6 8 **7** 5

Şekil 3.12 Devirli çaprazlama

h.Sıralı Çaprazlama: Bu çaprazlama yöntemi daha önceki sıralı çaprazlama yöntemleriyle benzerlikler göstermektedir. Çaprazlama işlemi yapılırken rasgele iki kesim noktası seçilir, babada kesim noktaları arasında kalan kromozomlar yeni bireye aynı konumda yerleştirilir. Kalan boşluklar sırasıyla annenden kopyalanır. Burada önemli olan anneden alınan kromozomların babadan kopyalananlarla aynı olmamasına dikkat edilmelidir.

Ebeveyn A	1 2 3 **4 5 6** 7 8 9 10
Yeni Birey	9 10 1 **4 5 6** 3 2 8 7
Ebeveyn B	9 10 1 3 2 4 6 8 7 5

Şekil 3.13 Sıralı çaprazlama

i.Düzenli Çaprazlama: Düzenli çaprazlama yönteminde öncelikle babanın tüm kromozomları yeni bireye kopyalanır. Boyutu birey boyutunun bir eksiği kadar, değerleri 0 ve 1 olmak üzere rasgele değişen bir kromozom oluşturulur. Oluşturulan bu vektörün 1 değerine sahip elemanlarının indisleri babanın bir kopyası olan yeni bireydeki değiştirilecek kromozomları göstermek için kullanılır. Babadaki bu indislere sahip kromozomların annedeki karşılık kromozomlarının indis değerleri alınır ve sıralanır. Daha sonra sıralanan bu indis değerlerindeki kromozomlar anneden alınarak yeni bireyde önceden tespit edilmiş değiştirilecek kromozomların yerine yazılır. Çaprazlama işlemi bu sırada gerçekleştirilmiş olur.

Ebeveyn A	1 2 3 4 5 6 7 8 9 10
Yeni Çocuk	9 2 3 1 5 6 4 8 7 10
Ebeveyn B	9 10 1 3 2 4 6 8 7 5

Şekil 3.14 Düzenli çaprazlama

j.Stefan Jacobs Çaprazlaması: Stefan Jacobs çaprazlamasında rasgele iki sayı seçilir. Bu sayılar x ve y olsun. Bu noktalardan x çaprazlamaya hangi noktadan başlanacağını, y sayısı da x sayısından itibaren kaç kromozomun alınacağını göstersin. Babanın x. kromozomundan itibaren y adet kromozom alınır; bu kromozomlar yeni bireyin ilk kromozomlarını oluşturur. Yeni bireyin babadan alınmayan kromozomları anneden sırayla alınır. Böylece çaprazlama işlemi

tamamlanmış olur. Şekil 3.15'e bakacak olursak, burada x ve y değerleri sırasıyla 3 ve 4 olsun.

Ebeveyn A	1 2 3 4 5 6 7 8 9 10
Yeni Birey	3 4 5 6 9 10 1 2 8 7
Ebeveyn B	9 10 1 3 2 4 6 8 7 5

Şekil 3.15 Stefan Jacobs çaprazlaması

2.Mutasyon: Mutasyon işlemi, uygunluk fonksiyon değerini geçememiş bütün bireylere uygulanan bir yöntemdir. Dolayısıyla mutasyon işleminde, çaprazlamadan farklı olarak tek bir bireyden yeni birey elde edilmektedir. Mutasyon işlemi kromozomlardan bazılarının tersi alınarak yapılır. Bu işlemde de önemli olan nokta mutasyon oranıdır. Mutasyon oranı, bireydeki kromozomların kaç tanesine mutasyon işleminin uygulanacağını belirler. Mutasyon uygulanacak bireyler rasgele seçilir. Mutasyon oranının çok yüksel seçilmesi, çaprazlama işlemiyle elde edilen iyi bireylerin kaybedilmesine neden olabilir.

Birliktelik kurallarının keşfi için yapılan bir çalışmada mutasyon için uyarlamalı bir mutasyon oranı seçilmiştir. Bilindiği gibi genetik algoritmalar çalıştırılıp sonuç elde edildiğinde son popülasyondaki bireylerin birbirine benzediği görülmektedir. Yani tek bir sonuç elde edilebilmektedir. Burada ise ilginç kuralları bulmak istenmekte yani birden fazla çözüm aranmaktadır. Bu yüzden en iyi birey popülasyonda baskın olduğundan mutasyon oranı da orantılı olarak arttırılmıştır. Kısaca en iyi bireyin popülasyonda ne kadar kopyası varsa o oranda mutasyon oranı da arttırılmıştır [13].

Mutasyon işlemiyle birlikte kaybolan gen değerlerinin yeniden ortaya çıkarılması, genetik yığılmanın önlenmesi, nüfusun çoğunluğuyla çözüme yakınsadığında bütünsel en iyi çözümü bulmak için nüfus çevresinde rasgele adımlarla arama yapılmasını sağlar. Mutasyon rasgele aramanın küçük bir kısmını gerçekleştirmekte ve daha çok arama uzayında araştırılmamış bir nokta kalmaması için diğer genetik algoritma işlemcilerine yardımcı olmaktadır. Ancak geleneksel görüş arama uzayının hızlı bir şekilde araştırılmasında çaprazlama işleminin mutasyon işleminden daha önemli ve etkin olduğu konusunda birleşmektedir [9].

Genetik algoritmalarda bazı işlemlerin ve parametrelerin probleme özel seçilmesi gerekir. Mutasyon işlemi de uygulanacağı probleme göre farklılıklar gösterebilir. Otomatik çizelgeleme problemi için yapılan çalışmada kullanılan mutasyon işlemine baktığımızda farklı mutasyon işlemlerinin denendiği görülmektedir. Uygulanan mutasyon uzmanları şunlardır: Geleneksel mutasyon uzmanı (AI), her geni 1/ (birey uzunluğu) ihtimalle, rasgele yeni bir atama yaparak mutasyona uğratılır; diğer bir uzman (AT), öncelikle müfredat dönemlerinin uygunluk değerine kısmi katkılarını göz önünde bulundurarak, sıralama stratejisi kullanıp rasgele bir müfredat dönemi seçer ve AI uzmanı bu dönem üzerinde uygulanır; sıralama mutasyonu uzmanı (OI), bireydeki her bir genin uygunluk değerine kısmi katkılarını göz önünde bulundurarak, sıralama stratejisi kullanıp rasgele bir gen seçer ve rasgele yeni bir atama yaparak mutasyona uğratır; uygulanan son yöntemde de uzman OT, AT gibi çalışır, son basamakta AI yerine OI uzmanının müfredat dönemi içerisinde kullanır [11].

Mutasyon işlemi kendi içinde de bazı yöntem farklılıkları içermektedir. Mutasyon işleminin çeşitleri şunlardır:

a.İkili Kodlanmış Bireylerde Mutasyon: Mutasyona uğratılacak olan kromozom mutasyon oranıyla rasgele seçilir. İkili kodlanmış olan kromozomlar (0,1) değerlerinden rasgele birini alır.

Orijinal birey	**0**0111**0**101000**1**1
Yeni birey	**1**0111**1**101000**0**1

Şekil 3.16 İkili Kodlanmış bireylerde mutasyon işlemi

b.Permütasyon Kodlamalı Mutasyon: Bu kodlama yöntemiyle kodlanmış olan bireylerde mutasyona uğrayacak olan kromozomun yeri rasgele başka bir kromozomla yer değiştirilerek yapılır.

Orijinal birey	1**2**3687445975**5**99
Yeni birey	1**5**3687445975**2**99

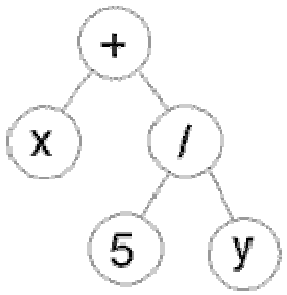
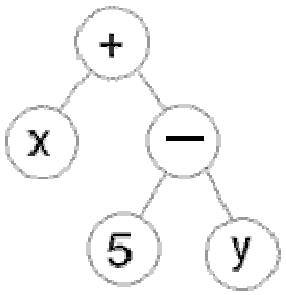
Şekil 3.17 Permütasyon kodlamalı mutasyon

c.Değer Kodlamalı Mutasyon: Bu yöntemde de mutasyon oranıyla belirlenen kromozomlara 0-1 aralığında seçilen rasgele bir sayı eklenir ya da çıkarılır.

Orijinal birey	**2.36** 3.89 8.75 4.62 1.25
Yeni birey	**2.47** 3.89 8.75 4.62 1.25

Şekil 3.18 Değer kodlamalı mutasyon

d.Ağaç Kodlamalı Mutasyon: Ağaç kodlamayla kodlanmış olan bireylerde seçilen düğünlerdeki işlemler ya da sayılar rasgele değiştirilir.

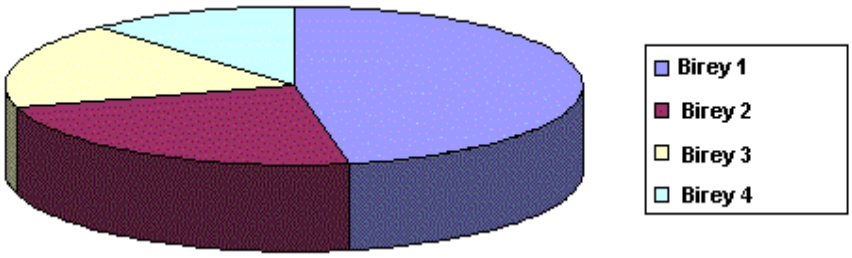
Orijinal birey	Yeni birey
	

Şekil 3.19 Ağaç kodlamalı mutasyon

3.Seçim Yöntemleri: Seçim yöntemi genetik algoritmaların en temel işlemlerindendir. Seçim yöntemleriyle, genetik algoritmaların temelini oluşturan "en iyi uyumluluğa sahip bireyin yaşaması" ilkesinin devamı sağlanmaya çalışılır. İyi bireylerin kaybedilmeden yeni nesillere aktarılması ve genetik işlemciler için seçilecek olan bireylerin uyumluluklarının yüksek olması genetik algoritmalarda çözüme ulaşmayı kolaylaştırır. Seçim yöntemlerinin genetik algoritmalara etkisi popülasyonda arama işleminin yönlendirilmesi ve iyi uyumluluğa sahip bireylerin yani iyi çözümlerin elde edilmesi şeklinde olmaktadır. Bunlarla beraber popülasyon yoğunluğu erken bir yakınsamayı önlemek ve bütünsel en iyi çözüme ulaşmak için korunmalıdır.

Seçim yöntemleri probleme göre seçilebilir. Genetik algoritmalarda kullanılan seçim yöntemleri şunlardır: Rulet tekerleği, SUS (Stochastic Universal Sampling), Elitizm Derece Seçimi, Turnuva Seçimi, Kararlı Hal Seçimi.

a.Rulet Tekerleği Seçim Yöntemi: Rulet seçiminde bireyler uygunluk fonksiyonuna göre bir rulet etrafına toplanır. Uygunluk fonksiyon değeri herhangi bir kritere uyan bireylerin seçilmesi için kullanılır. Bu rulet üzerinden rasgele bir birey seçilir. Daha büyük uygunluk fonksiyon değerine dolayısıyla daha büyük alana sahip bireyin seçilme şansı daha fazla olacaktır. Rulet seçimi eğer uygunluk fonksiyon değeri çok fazla değişiyorsa sorun çıkartabilir.



Şekil 3.20 Rulet tekerleğinde bireylerin sıralanışı

Bu yöntem şöyle uygulanabilir :

1.Popülasyondaki tüm bireylerin uygunluk fonksiyon değerini topla. Bu değer T olsun.

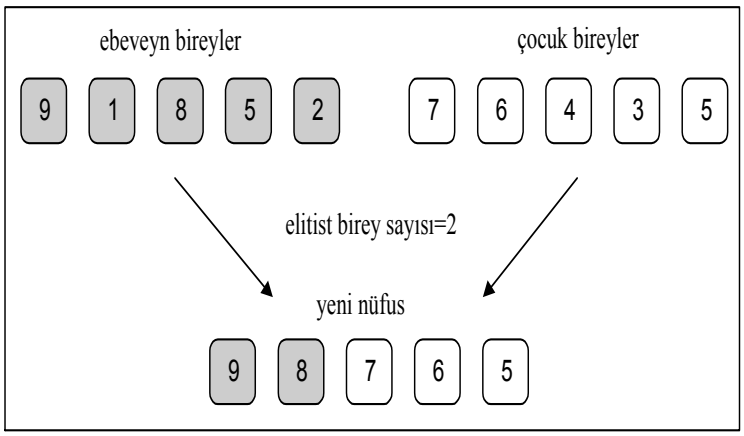
2.Rasgele bir r değeri belirle.

3.Popülasyondaki tüm bireylerin uygunluk fonksiyon değerlerini toplamaya başla; r değeri yakalandığında ya da aşıldığında hangi bireyin uygunluk fonksiyon değeri eklendiyse o bireyi tut [3].

b.Stochastic Universal Sampling (SUS): Bu yöntem James Baker tarafından 1987 yılında ortaya atılmıştır. Temelde rulet tekerleği yöntemine benzemekle beraber bazı yönlerden farklılıklar göstermektedir. Rulet tekerleği yönteminde N adet ebeveyn için N defa rulet tekerleğinin döndürülmesi gerekmekteydi; SUS yönteminde N adet ebeveynin seçimi için N tane birbirine eşit işaretçiler bulunan tekerlek bir defa döndürülmektedir. Bu yöntemin örnek kodu Baker tarafından şöyle verilmiştir [3]:

```
ptr = Rand(); /* [0,1] arasında bir sayı üretilmektedir.*/                    (3.1)
        for (sum = i = 0; i < N; İ++)
                for (sum += ExpVal(i,t); sum > ptr; pt++)
                    select(i);
```

c.Elitizm: Elitizm 1975'te Kenneth De Jong tarafından geliştirilmiştir. Bu yöntemle üreme, çaprazlama ve mutasyon gibi işlemler sonucu yok olabilecek iyi uygunluk değerine sahip bireylerin korunması sağlamaya çalışılmaktadır. Bir çok araştırmada elitizm yönteminin kullanılmasıyla genetik algoritmaların performansının arttığı görülmüştür [3].



Şekil 3.21 Elitizm uygulaması

Örnekte de görüldüğü üzere uygunluk fonksiyon değeri yüksek olan ebeveynler (9 ve 8 değerli olanlar) yeni popülasyonda uygunluk fonksiyon değeri küçük olan (3 ve 4 değerli olanlar) çocukların yerine aktarılmışlardır.

d.Derece Seçimi (Rank Selection): Bu seçim yöntemi 1985 yılında Baker tarafından geliştirilmiştir. Derece seçimi, oldukça basit ve etkili bir seçim yöntemidir. Uygunluk fonksiyon değeri büyük bireyler yeniden üreme seçiminde etkin olabilirler. Bu da genetik bilginin kaybolmasına neden olur. Bütün bu nedenlerden dolayı genetik yoğunluğun azalmasını ve hızlı yakınsamayı önlemek için bu seçim yöntemi kullanılır. Bireyler öncelikle gerçek uygunluk değerine göre sıralanır. Daha sonra yeni elde edilen uygunluk değerleri doğrusal veya üssel olarak derecelerine göre belirlenir. Bir birey mutlak uygunluk değerinden çok derecesine göre orantılı bir

olasılıkla seçilir. Sonuç olarak en büyük uygunluk değeri ve ortalama uygunluk değeri arasındaki oran belirli bir değer için normalize edilir. Böylelikle uç değerlerdeki bireylerin etkileri önemsizleşir [9].

e.Turnuva Seçimi: Turnuva seçim yöntemi, derece seçim yöntemiyle benzerlikler göstermektedir; fakat daha etkili ve paralel uygulamalarda daha iyi sonuçlar vermektedir. Turnuva seçiminde popülasyondan rasgele iki birey seçilir. [0,1] aralığında rasgele bir r sayısı seçilir. Eğer r değeri k (k 0,75 gibi bir parametre olsun) değerinden küçük ise bu iki iyi birey ebeveyn olarak seçilir; değilse bireyler popülasyona geri gönderilirler ve böylece bu bireyler tekrar seçilebilir. Turnuva seçim yöntemi 1991'de Goldberg ve Deb tarafından sunulmuştur [3].

f.Kararlı Hal Seçimi (Steady State Selection): Kararlı hal seçim yönteminde bireylerin büyük bir kısmının bir sonraki popülasyonda hayatta kalma zorunluluğu esas alınır. Bu yöntemle yeni bireylerin oluşturulması için her popülasyonda iyi uygunluk değerine sahip bireyler seçilirler; daha sonra kötü uygunluk değerine sahip bireylerin yerine yeni bireyler yerleştirilirler. Kısaca bu yöntemde önce alt popülasyon oluşturulur; daha sonra daha sonra uygunluk değeri hesaplanır ve kötü bireylerin yerine başlangıç popülasyonundaki en iyi bireyler konulur.

Karalı hal seçim yöntemi kullanılarak yapılan ders programı hazırlanması ile ilgili bir çalışmada bu yöntem şöyle uygulanmıştır: evrim esnasında diğer nesile iletilecek bireylerin seçiminde iki farklı yöntem denenmiştir. Birinci yöntem, eşleşen ebeveynler ile doğan iki çocuk grubundan en iyi ikisini alarak ebeveynleriyle yer değiştirir; diğeri ise oluşan çocukları ile nüfusun en kötü bireyleriyle, onlardan daha iyi olmak koşulu ile yer değiştirir. İki yaklaşımda da yeni nesil oluşturulurken çeşitliliği sağlamak amacıyla, nüfusa eklenecek yeni bireylerin kopyasının nüfusta olmamasına özen gösterilir; aksi takdirde birey yeni nesile giremez [11].

3.2.3.2 Genetik parametreler

Bu bölümde genetik algoritmanın yapılandırılması önemli bir faktör olan genetik parametrelerden bahsedilecektir. Genetik algoritma parametreleri: uygunluk

fonksiyon değeri popülasyon büyüklüğü, iterasyon sayısı, çaprazlama oranı ve mutasyon oranından oluşmaktadır.

Genetik parametreler algoritmanın şekillendirilmesinde önemli rol oynarlar; o yüzden bu parametreler çalışılan probleme uygun seçilmelidir.

Uygunluk Fonksiyon Değeri: Uygunluk fonksiyonu, probleme özel tanımlanan ve problemin çözümünde genetik algoritmaların en temel bileşenidir. Uygunluk fonksiyonu genetik algoritma mantığında her problem için belirlenmesi gereken ve sadece probleme özgü olan tek kısımdır.

Uygunluk fonksiyon değeri, problemin çözümü için kullanılacak olan kromozomların ne derecede iyi bireyler oluşturduklarını ve bu yoldan çıkılarak problemin sonucunun iyileştirilmesinde rol oynar. Genetik algoritmaların başarısı genellikle uygunluk fonksiyon değerinin ne kadar doğru seçildiğiyle orantılıdır.

Uygunluk fonksiyonu oluşturulurken dikkat edilmesi gereken kural, fonksiyonun birey değerini yansıtıp yansıtmadığıdır. Bir uygunluk fonksiyonunun ideal olarak tanımlanması, arama uzayında makul uygunluk değerlerine sahip bireylerin kendilerinden biraz daha iyi uygunluk değerlerine sahip bireylere yakın olmasını sağlar. Bu nedenle fonksiyonun uygun olarak tanımlanması istenir; ancak pek çok problem bu mümkün olamamaktadır [9].

Genetik algoritmalarla, veri tabanlarından bilgi keşfi için yapılan bir çalışmada uygunluk fonksiyon değeri şöyle belirlenmiştir:

$$Fitness(i) = Acc(i) * Surp(i) \quad (3.2)$$

Burada ‘Acc’ değeri kuralın kestirim oranı, ‘Surp’ değeri ise kuralın ilgilik ölçütü olarak tanımlanmıştır [12].

Otomatik çizelgeleme yapılan bir çalışmada da uygunluk değeri, herhangi bir kısıta uymayan bir atamanın cezalandırılması yöntemiyle hesaplanmıştır. Buna göre

Uzaklık(x.,y), y ders buluşması-çizelge hanesi eşleşmesinin, x kısıtını sağlamaktan ne kadar uzak olduğunu, diğer bir deyişle x kısıtını ne kadar ihlal ettiğini belirtsin. Bu durumda, uygunluk değeri aşağıdaki formüldeki gibidir:

$$\text{Uygunluk değeri} = \sum_{k=1}^{N} \sum_{l=1}^{B} \text{Uzaklık}(qk, Pl) \quad [11] \qquad (3.3)$$

Popülasyon Büyüklüğü: Genetik algoritmalarda popülasyon büyüklüğü bir nesilde kaç birey olduğunu ifade eder. Popülasyondaki birey sayısı az olursa genetik algoritmanın çaprazlama işlemi için fazla bir seçeneği olmayacaktır; birey sayısı çok fazla olursa algoritma yavaş çalışacaktır.

Genetik parametreler üzerine yapılan çalışmalarda popülasyon büyüklüğüyle ilgili farklı sonuçlar elde edilmiştir: 1975 yılında De Jong tarafından yapılan çalışmada popülasyon büyüklüğünün 50-100; 1986 yılında Grefenstette tarafından incelendiğinde ise popülasyon büyüklüğünün 30; 1989'da Schaffer, Caruana, Eshelman ve Das tarafından yapılan çalışmalarda ise Grefenstette'nin 1986' da elde ettiği sonuçlara yakın olmakla beraber popülasyon büyüklüğü 20-30 birey arasında olmasının verimli sonuçlar üreteceği söylenmiştir [3].

İterasyon Sayısı: İterasyon sayısı, genetik algoritmaların sonlandırılma koşulu olarak düşünülebilir. Kullanıcı isteğine göre seçilen sonlandırma koşulunda genellikle genetik algoritmalar aranılan kural ya da kuralların bulunmasıyla sonlandırılır. Bazı problemlerde aranılan kuralın bulunamama ihtimali göz önüne alınarak algoritmaya maksimum iterasyon sayısı parametresi de eklenebilir.

Çaprazlama Oranı: Genetik algoritmaların önemli parametrelerinden biri de çaprazlama oranıdır. Uygunluk fonksiyon değerini geçemeyip, genetik operatörler uygulanmak üzere ayrılan bireylere öncelikle çaprazlama işlemi uygulanır. Çaprazlama için seçilecek olan ebeveynlerin sayısı çaprazlama oranına bakılarak hesaplanır. Uygunluk fonksiyon değerini geçememiş olan her bireye çaprazlama işlemi uygulanmaz (çaprazlama oranı %100 seçilmediği sürece) .

Çaprazlama oranının seçimiyle ilgili de birçok çalışma yapılmıştır. Bu çalışmalardan şu sonuçlar elde edilmiştir: 1975 yılında De Jong tarafından yapılan çalışmada çaprazlama oranının yaklaşık 0.6, 1986 yılında Grefenstette tarafından genetik algoritmalarda parametrelerin etkisi incelendiğinde çaprazlama oranının 0.95, 1989'da Schaffer, Caruana, Eshelman ve Das tarafından yapılan çalışmalarda ise çaprazlama oranı 0.75 – 0.95 arasında seçilmesi gerektiği ortaya konmuştur [3].

Mutasyon Oranı: Mutasyon oranı parametresi, mutasyon işleminin bireyin kromozomlarının kaçta kaçına uygulanacağını belirler. Mutasyon oranı değerine göre rasgele seçilen kromozomlara mutasyon işlemi uygulanır. Burada önemli olan mutasyon oranının kaç seçileceğidir. Mutasyon çok büyük seçilerse, çaprazlama işleminden gelen iyi uyumluluğa sahip olabilecek bireyler kaybedilebilir.

Mutasyon oranının kaç olması gerektiği ile ilgili yapılan çalışmalarda şu sonuçlar elde edilmiştir: 1975 yılında De Jong tarafından yapılan çalışmada mutasyon oranının 0.001, 1986 yılında Grefenstette tarafından incelendiğinde 0.01, 1989'da Schaffer Caruana, Eshelman ve Das tarafından yapılan çalışmalarda mutasyon oranı 0.005 – 0.01 arasında değerler bulunmuştur [3].

Genetik algoritmalarla yapılan çalışmalara baktığımızda seçilen genetik parametrelerin probleme göre değiştiğini görmek mümkündür. Göğüs kanseri hastalarının verileri kullanılarak yapılan bir çalışmada mutasyon oranı=0.01, çaprazlama oranı=0.86 popülasyon büyüklüğü=150 seçilmiştir [14].

Diyabet ve obezite hastalarının verileriyle yapılan bir çalışmada da genetik parametrelerden mutasyon oranı kullanılmış ve 0.1 seçilmiştir [15].

3.3. Genetik Algoritmaların Uygulama Alanları

Yaşadığımız yüzyıl içerisinde hızla gelişen ve karmaşık bir hal alan problemler araştırmacıları farklı yollar bulmaya bulmaya yöneltmiştir. Bunun bir sonucu olarak gelişen genetik algoritmalar birçok problemin çözümünde kullanılabilir hale gelmiş ve verimli çözümler ürettiği görülmüştür. Büyük arama uzaylarında geleneksel

yöntemlerle çok uzun süreçlerde elde edilen sonuçlar, genetik algoritmalar sayesinde kısa sürede kabul edilebilir çözümler elde edilmesini sağlamıştır.

Genetik algoritmalar genel olarak kendine özel çözümleri olan problemler için kullanılmazlar; hiçbir çözümü olmayan problemlere çözüm ararlar. Bu bağlamda genetik algoritmaların aşağıda açıklanan tipte problemler için kullanılabileceğini söyleyebiliriz:

1.Arama uzayının büyük ve karmaşık olduğu,

2.Mevcut bilgiyle sınırlı arama uzayında çözümün zor olduğu,

3.Problemin belirli bir matematiksel modelle ifade edilemediği,

4.Geleneksel eniyileme yöntemlerinden istenen sonucun alınamadığı alanlar [9].

Genetik algoritmaların uygulama alanlarını şu başlıklar altında genelleyebiliriz:

Eniyileme (Optimizasyon): Bir arama yöntemi olan genetik algoritmalar, farklı bilim dallarındaki farklı bilim dallarındaki eniyileme problemlerini çözmede kullanılmaktadır. Genetik algoritmaların uygulandığı eniyileme problemleri, fonksiyon eniyilemesi ve birleşi eniyilemesi altında toplanabilir. Genetik algoritma araştırmalarının önemli bir bölümü fonksiyon eniyilemesi ile ilgilidir. Genetik algoritmalar, geleneksel eniyileme yöntemlerine göre zor, süreksiz ve gürültü içeren fonksiyonları çözmede daha etkindirler. Genetik algoritmaların uygulandığı diğer bir eniyileme problem sınıfı olan birleşi eniyileme problemleri ise, istenen amaçlara ulaşmak üzere, sınırlı kaynakların etkin tahsis edilmesiyle ilgilidir. Bu sınırlar genel olarak iş gücü, tedarik veya bütçe ile ilgilidir.

Çeşitli avantajlarına rağmen genetik algoritmaların uygulamalarında bir takım sorunlarla da karşılaşılmaktadır. Bu sorunları aşmak için çeşitli yöntemler geliştirilmiştir. Buna kısıtların ele alınmasındaki soruna karşı ceza fonksiyonu yönteminin kullanılması örnek verilebilir [10].

Otomatik Programlama ve Bilgi Sistemleri: Genetik algoritmalar belirli ve özel problemler içinde kullanılabilir. Bu problemlerin çözümünde kullanılan bilgisayar programlarının arkasında çalışan genetik algoritmalar ders programı hazırlama gibi çizelgeleme problemleri, çip tasarımı, ağların çizelgelendirilmesi ve tıp verileri üzerinde yapılan araştırmalar gibi birçok konuda kullanılabilmektedir.

Mekanik Öğrenme: Mekanik öğrenme iki temel amaçla kullanılır: Birincisi bir veritabanın anlaşılması ve yorumlanabilmesi, ikincisi de yeni objelerin tahmini.

Sınıflama sistemi, genetik algoritmaların mekanik öğrenme alanında bir uygulamasıdır. Basit dizi kurallarını öğrenen bir mekanik öğrenme sistemi olan sınıflama sisteminin kural ve mesaj sistemi, özel bir üretim sistemi olarak adlandırılabilir. Bu üretim sistemi "if-then" kural yapısını kullanır. Bir üretim kuralı "if" yapısından sonra belirtilen durum için "then" yapısından sonra gelen faaliyetin gerçekleştirilmesini içerir.

Genetik algoritmalar, sınıflama sistemlerinde kural bulma mekanizması olarak kullanılmaktadırlar. Genetik algoritmalar ayrıca sinir ağlarında ve proteinin yapısal analizinde de kullanılmaktadır [10].

Ekonomik ve sosyal sistem modelleri: Bu alanda yapılan çalışmalarda genetik algoritmalar verilen girdiler için arzu edilen çıktıları üreten özel bir hesaplama programı ile program uzayında arama yapabilmek için kullanılmaktadır. Genetik algoritmaların kullanıldığı genetik programlama ile bu tip problemlerin çözümlerine daha kolay ulaşılabilmektedir.

Genetik algoritmalar yenilik sürecinin modellenmesi amacıyla da kullanılmaktadır. Ayrıca genetik algoritmaların, fiyat verme stratejilerinin gelişim süreçlerini ve kazanç getiren pazarların ortaya çıkış süreçlerini modelleme alanlarında da kullanımları oldukça yaygındır. Genetik algoritmalar sosyal sistemlerin evrimsel yöntemlerini anlamak amacıyla kullanılmaktadır. Bunlara örnek olarak işbirliğinin evrimi, iletişimin evrimi ve karıncalardaki iz takibi davranışının evrimi verilmektedir [10].

Finans: Genetik algoritmalar, finans modelleme uygulama modelleri için uygun bir algoritma çeşididir. Amaç fonksiyonu odaklı olan genetik algoritmalar, finans problemlerinde amaç fonksiyonları tahmin etme gücüne veya bir kıyaslama sonucuna bağlı getirilerdeki gelişmeleri içerir. Kullanılan araç ve problemler arasında iyi bir eşleşme bulunmaktadır.

Finans problemlerinin çözümünde genetik algoritmalar bulanık ve yapay sinir ağları yaklaşımlarıyla birlikte kullanılmaktadır [10].

Pazarlama: Pazarlama, tüketici hareketlerinin takip edilmesini ve bu hareketlerden tüketici davranışlarının çıkartılıp analiz edilmesini gerektirir. Tüketici verileri oldukça büyük miktarda verilerden oluşmaktadır. Bu verilerin analizi veri madenciliği ile yapılabilmektedir. Veri madenciliğinde genetik algoritmalar kullanılarak bu konuda verimli ve kullanılabilir örüntüler yakalanabilmiştir.

Bunların dışında genetik algoritmalar bazı problemler içinde kullanılmaktadır. Daha çok üretim/işlemler alanındaki problemlerde kullanılan genetik algoritmalar şu problemlerim çözümlerinde etkili sonuçlar vermektedir:

1.Montaj hattı dengeleme problemi,

2.Çizelgeleme problemi,

3.Tesis yerleşim problemi,

4.Atama problemi,

5.Hücresel üretim problemi,

6.Sistem güvenilirliği problemi,

7.Taşıma problemi,

8.Gezgin satıcı problemi,

9.Araç rotalama problemi,

10Minimum yayılan ağaç problemi.

Genetik algoritmaların uygulandığı yukarıda belirtilen problemler başka yöntemlerle de çözülebilmektedir. Ancak genetik algoritmaların başlangıç çözümünden bağımsız olma paralel çözüm arama ve performans hızı gibi özellikleri kullanılan diğer yöntemlere göre daha uygun bir çözüm yöntemi olarak öne çıkarmaktadır.

Genetik algoritmaların bu avantajlarıyla beraber problemin çözümü sırasında karşılaşılabilecek bazı sorunları da içermektedir. Bunların başında parametre seçimi gelmektedir. Her problemin parametre çeşitleri farklı olduğu için bu soruna genellikle deney tasarımı yöntemiyle çözülmektedir. Eniyileme problemlerinde de kullanılan kısıtlardan dolayı sorunlar yaşanabilmektedir.

3.4. Veri Madenciliğinde Genetik Algoritmalar

Veri madenciliği, gelişen bilgisayar teknolojileri sayesinde birçok alanda kullanılır duruma gelmiştir. Gelişen teknolojiler beraberinde yeni problemler getirmişler ve dolayısıyla ortaya çıkan bu ihtiyaçların karşılanması gerekliliği veri madenciliğinde de yeni açılımları gerektirmiştir.

Genetik algoritmalar her ne kadar veri madenciliğinde sıkça kullanılır hale gelse de, beraberinde bazı sorunları da getirmektedir. Veri madenciliğinde, genetik algoritmalar yapılandırılırken şunlara dikkat edilmelidir:

a.Kullanıcıya açıklanması ve anlatılması zor olabilir,

b.Sorunu soyutlamak ve bireyleri temsil etmek için kullanılan modeller zordur,

c.Uygunluk fonksiyonunu belirlemek zordur,

d.Çaprazlama ve mutasyon işlemlerinin nasıl yapıldığına dair sorunun çözümü zordur [5].

Bu bölümde veri madenciliğinde kullanılan birçok algoritmadan biri olan genetik algoritmaların kullanım şekillerinden ve veri madenciliğinde gösterdikleri performanstan bahsedilecektir.

3.4.1. Genetik algoritmaların veri madenciliğinde kullanımı ve performansı

Veri madenciliğinde genetik algoritmalar farklı veri kümeleriyle çeşitli amaçlar için kullanılmıştır. Yapılan çalışmalarda da görüleceği üzere genetik algoritmalar veri madenciliğinde verimli sonuçlar üretilmesini sağlamışlar ve diğer algoritmalarla yapılan performans ölçümlerinde çoğu zaman daha iyi sonuçlar ortaya koymuşlardır. Genetik algoritmalar tıp, öğrenci, müşteri, kredi kartı verileri gibi birbirinden farklı nitelikleri olan veri kümeleriyle çalıştırılmış ve istenen sonuçlar elde edilmiştir. Bu bölümde genetik algoritmalarla yapılan çalışmalar konu alınmıştır.

2006 yılında tıp verileriyle yapılan bir çalışmada göğüs kanseri olma potansiyeli olan kadınların tespiti için genetik algoritma tabanlı bir yaklaşımla göğüs kanseri örnekleri değerlendirilmektedir. Daha önceleri yapılan çalışmalarda göğüs kanseri tanılarında istatistikle desteklenmiş yöntemler kullanılmıştır. Göğüs kanseri tanıları lineer olmayan tipte olmaktadır; istatiksel yaklaşım kullanarak, bağımsız değişkenlerin içinden önemli olanını alıp kapsamlı bir model geliştirmek çok zordur.

Son zamanlarda sinir ağlarıyla yapılan çalışmalar, bilinen istatiksel yaklaşımlardan ve dinamik stres metodundan daha güvenilebilir sonuçlar verdiğini göstermektedir. Sinir ağlarının kullanımının literatürde faydalı olduğu gösterilmiştir; fakat en büyük engelin, kullanılan modelde ya da yapıda, sınıflandırma kurallarının farkedilme zorluğunun olmasıdır.

Bu çalışmada alınan sonuçlar ticari bir veri madenciliği yazılımıyla karşılaştırılmış ve deneysel olarak görülmektedir ki, modelin basitliğini artırmak ve kestirim oranını iyileştirmek için kural çekme yaklaşımı kurulmuştur. Bu çalışmadaki kural çıkarma

sistemi, göğüs kanseri olma potansiyelini tespit edebilen diğer uzman sistemler kadar yetenekli olduğu görülmüştür [14].

Genetik algoritmalar kullanılarak tıp verileri üzerinde yapılan başka bir çalışmada da genetik özelliklerin ve çevresel faktörlerin obezite ve diabet hastalığı gibi birden fazla faktöre bağlı olan hastalıklar üzerindeki etkisi incelenmiştir. Deneyler Lille Biyolojik Enstitüsünün verileriyle yapılmıştır. Çok büyük sayıda veri olduğundan, keşifsel (heuristic) yaklaşım seçilmiştir.

İlk aşamada nitelik seçimi için genetik algoritmalar kullanılmıştır. Bu çok spesifik problemi çözmek için genetik algoritmada tanımlanmış bazı mekanizmalar kullanılmıştır; paylaşım, rasgele yerleştirme (göç) , genetik operatörler gibi. İkinci aşamada bir önceki aşamada seçilen niteliklerin sınıflandırılmasına çalışılmıştır. Bunun içinde en popüler sınıflandırma algoritması olan k-means kullanılmıştır [15].

Öğrenci verileriyle yapılan bir çalışmada da iyi sonuçlar üretilmiştir. Bu makalede {alındı alınmadı} ya da {var, yok} şekilde ikili değerler dışında kategorik ve nicel değerler de içeren veri tabanlarında birliktelik kurallarının keşfi için yapay zeka ve zeki hesaplama tekniklerinden genetik algoritma bulanık mantık tabanlı etkili, yeni bir yöntem geliştirilmiştir.

Genetik algoritmalarda başlangıç popülasyonunu gelişigüzel üretmek yerine, bunu çözüm uzayına düzgün dağıtan düzenli popülasyon yöntemi kullanılmıştır. Genelde kullanılan yöntemlerin aksine yüksek destek ve güven değerlerine sahip birliktelik kuralları yoğun nesne kümeleri üretilmeden direkt olarak ve her veri tabanı için belirlenmesi güç olan minimum güven ve minimum destek eşiklerine ihtiyaç duyulmadan keşfedilmiştir.

İlginç birliktelik kurallarını bulmak için uyarlamalı mutasyon ve elitizm stratejisi uygulanmıştır. Bu şekilde genetik algoritmanın son popülasyonu ilginç birliktelik kurallarını temsil etmiştir.

Önerilen yöntem hem yapay bir veri tabanında hem de Fırat Üniversitesi Elektrik-Elektronik Mühendisliği lisans öğrencilerinin ders not kayıtlarında denenmiş, kullanışlı ve ilginç kurallar etkili şekilde bulunmuştur [13].

Eğitim alanında yapılan bir çalışmada da öğrenci bilgileri kullanılarak farklı kestirim kurallarının nasıl bulunabileceği gösterilmeye çalışılmış ve bunlar kullanılarak web üzerinden yapılan kurslarda düzeltilmesi gereken noktalar aranmıştır. Eğitim alanındaki gücünü arttırmak için AHA' da birçok değişiklikler yapılmıştır. Kullanılabilecek veriler arasındaki ilişkiyi bulmak için AHA'da kütükte tutulmuş bilgiler kullanılmıştır (okuma zamanları seviye zorlukları ve test sonuçları). En ilgi çekici olan ilişkiler öğretmene gösterilmiştir. Böylece kursun daha verimli olabilmesi için gereken değişikliklerin farkedilebilirliği kolaylaştırılmıştır [17].

Genetik algoritmaların başka algoritmalarla beraber kullanıldığı bir çalışmada da veri madenciliğinde en sık kullanılan verilerin gösterim biçimi olan İf-Then kuralları formundaki verilerin keşfi konu alınmıştır. Bu bağlamda, ilginç olan bulanık kestirim kurallarının keşfi için bir genetik algoritma (GA) tasarlanmıştır. GA, kullanıcı için yeni ve sürpriz olabilecek kestirim kurallarını arar.

Ayrıca bulanık mantık, GA ile elde edilmiş kuralların anlaşılabilirliğini kolaylaştırır; çünkü terimler sözlük anlamlarıyla kullanılır. Örnek, gerçek bir bilim ve teknoloji veritabanına uygulanmış olup; GA ve J4.8 algoritmaları bu örnekte karşılaştırmalı olarak verilmiştir. Deneylerde, kestirim oranı ve her iki algoritmada elde edilmiş kuralların ilginçlik derecelerine bakılmıştır. Bu çalışmada kurallar GA ile elde edilmiştir ve en iyi kuralların elde edilmesi için de J4.8 kullanılmıştır [12].

Veri madenciliğinde sınıflandırma için kullanılan birçok kural keşfi teknikleri vardır. Bu çalışmada da bu tekniklerden ikisi denenmiştir. Genellikle, genetik programlama kullanılarak oluşturulmuş sistemlerin öğrenme hızlarının yavaş olduğu görülmüş. Bu nedenle, çevre koşullarına göre ayarlanmış olan yüksek önceliğe sahip bir öğrenme sistemi kurulabilir; çünkü yapı da aynı zamanda oluşturulmaktadır.

Başka bir konuda, büyük veritabanları için kural oluşturulmasında kullanılan Apriori algoritmasıdır. Bu bir birliktelik kuralı algoritmasıdır. Apriori algoritması kural yapısı için iki değer kullanır: destek (support) ve confidence (güven). Her indeksin eşik değerine bağlı olarak, arama uzayı küçültülebilir ya da aday olan birliktelik kurallarının sayısı çoğaltılabilir. Bununla birlikte etkili bir eşik değeri oluşturabilmek için deneyim gereklidir.

Yukarıda bahsedilen her iki teknikte avantaj ve dezavantajlar içermektedir. Bu makalede Apriori algoritmasıyla genetik programlamayı birleştirerek, veritabanları için kural keşfinde kullanılacak teknikler bulunması amaçlanmıştır. Kurallarını böyle oluşturan bir öğrenme metodunun kullanılmasındaki amaç; büyük veritabanlarındaki değişken kuralların aranmasını sisteme oturtmaktır [18].

Sınıflandırmayla ilgili yapılan başka bir çalışmada da şu sonuçlar elde edilmiştir. Bu makalede sınıflandırma tekniği olan k-means ile iki farklı genetik algoritma yaklaşımı örnek bir veri kümesi kullanılarak karşılaştırılmıştır.
İki teknik arasında yapılan karşılaştırmada genetik algoritmayla elde edilen sonuçların daha iyi olduğu görülmüştür [6].

Yapılan başka bir çalışmada da, bulanık birliktelik kurallarının madenciliği için otomatik metot geliştirilmeye çalışılmıştır. Bu amaçla, ilk önce sınıflandırma temelli olan genetik algoritmalar; ikinci olarak da literatürde çok bahsedilen ve etkili bir yöntem olan örneklemeyle sınıflandırma yöntemi kullanılmıştır.

Genetik algoritma tabanlı yaklaşımla, literatürde geçen diğer yaklaşımlar karşılaştırılmıştır. Deneylerde, genetik algoritmayla bulunan ilginç kurallarının sayısının diğer metotlara göre daha fazla olduğu görülmüştür. Deneyler gerçek veri kümesiyle yapılmış olup; önerilen yaklaşımın anlamlı sonuçlar ürettiği, verimli ve etkili olduğu görülmüştür [8].

Veri madenciliği çok büyük boyuttaki verilerden alışılmamış ya da ilginç olan bilgileri keşfedebilme problemiyle ilgilenir. Bu problem genellikle alınmış örneklerin

standart sorgulama mekanizmalarında ya da klasik istatiksel metotlarda kullanım zorlukları ortaya çıktığı zaman sezgisel olarak çözülürdü.

Bu makalede, otomatik kural keşfi süreci insanlar tarafından kolaylıkla anlaşılabilen bir genetik programlama yapısı sunulmuştur. Diğer tekniklerle de karşılaştırıldığında sonuçların başarılı olduğu görülmüştür. Ayrıca, elde edilmiş olan bazı kurallar gösterilmiş ve ayrılmış olan değerler ispatlanmıştır [19].

Bu bölümde bahsedilen çalışmalardan da görüleceği üzere genetik algoritmalar birçok alanda olduğu gibi veri madenciliğinde de kullanılmasıyla verimli sonuçların elde edilmesini sağlamıştır.

4. VERİ MADENCİLİĞİNDE GENETİK ALGORİTMALAR KULLANILARAK ÖĞRENCİ VERİLERİNİN DEĞERLENDİRİLMESİ

4.1. Giriş

Veri madenciliğinin tanımlarında da vurgulandığı üzere bu tür çalışmaların asıl amacı geçmişte toplanmış verilerin kullanılmasıyla gelecek için anlamlı, işe yarar ve ilginç olabilecek kurallar yakalayabilmektir. Eğitim alanında yapılan veri madenciliği çalışmalarında genel amaçlar; eğitim yöntemlerinin doğruluğu, öğrenci başarısının nasıl arttırılabileceği, verilen derslerin birbiriyle ilişkilerinin öğrenci başarısına olan etkisi, öğrenci alışkanlıkları ile ders başarılarının etkileşimi gibi konularda anlamlı kuralların yakalanması yönünde olmuştur.

Bu tez çalışmasında öğrencilerin ÖSYM verileri ve ders başarı ortalamaları dikkate alınarak öğrencilere ait veritabanında bulunan nitelikler, ÖSS sınavındaki başarıları ve ders başarı ortalamaları arasındaki ilginç kuralların yakalanabileceği ve yeni gelecek öğrencilerin başarı durumlarının öngörülebileceği bir öğrenci başarı analiz aracı geliştirilmiştir. Bu çalışmanın amacı genetik algoritmaların incelenmesi ve bu algoritmanın veri madenciliğinde kullanılarak geliştirilen yazılım aracı ile öğrenci verileri arasındaki nitelik incelenmesinin ve değerlendirilmesinin sağlanmasıdır. Bu uygulamayla genetik algoritmaların veri madenciliğinde kullanılması sonucu üretilen kuralların doğruluğunun ve kullanılabilirliklerinin gözlenmesi ve eğitimcilere katkıda bulunulması amaçlanmıştır.

Uygulamada kullanılan veritabanı, ÖSYM'den her sınav dönemi sonunda gönderilen 2003 ve 2004 yıllarına ait öğrenci verileri ile öğrenci bilgi sisteminde bulunan 2003 ve 2004 girişli öğrencilerin ders başarı verileri birleştirilerek oluşturulmuştur. ÖSYM'den gelen verilerden seçilen niteliklerin kullanılma amacı, bu nitelikteki öğrencilerle ders başarılarının etkileşiminin gözlenmesidir.

4.2. Veritabanı Üzerinde Yapılan Çalışmalar

Bu uygulamada kullanılan veritabanı Kocaeli Üniversitesi Bilgi İşlem Daire Başkanlığından alınan 2003 ve 2004 girişli öğrencilere ait verilerden oluşmaktadır.

Uygulama Delphi 7 programlama dili ve Microsoft MS SQL SERVER 2000 veritabanı kullanılarak gerçekleştirilmiştir.

Uygulamanın amacı doğrultusunda hazırlanan veritabanı oluşturulurken 2003 ve 2004 girişli öğrencilerin ders başarıları ile, yine bu yıllarda giriş yapan ÖSYM'den gönderilen öğrenci verileri birleştirilmiştir.

Uygulamaya başlanmadan önce, veri madenciliğinin önemli bir adımı olan veri önişleme adımı gerçekleştirilmiştir. Bu adıma gerek duyulmasının nedenleri şunlardır: ÖSYM'den gelen verilerin metin (text) dosyalarda tutuluyor olması, bu dosyalardaki bütün verilerin bu uygulama için kullanılmayacağından belli bir kısmının içlerinden çekilme ihtiyacının duyulması, ÖSYM verileri ile bilgi işlem bünyesinde tutulan öğrenci verilerinin hazırlanan bu verilerle birleştirilmesi. Bu nedenlerle veri önişleme işlemi aşağıdaki aşamalardan oluşmuştur:

1.Veri madenciliğinde genetik algoritmalar üzerinde yapılan bu uygulama KO.Ü. Mühendislik Fakültesinin beş bölümüne ait 436 öğrencinin verileri kullanılarak hazırlanmıştır. Bu beş bölüm: Bilgisayar Mühendisliği, Elektronik ve Haberleşme Mühendisliği, Endüstri Mühendisliği, Mekatronik Mühendisliği ve Elektrik Mühendisliğidir.

Bu beş bölümün seçilme amacı birbirlerine yakın dallar olduklarının düşünülmesi ve ÖSS puanlarına bakıldığında birbirlerini takip eder durumda olmalarıdır.

2.Öncelikle ÖSYM tarafından bilgi işlem daire başkanlığına metin dosyalar halinde gönderilen verilerin tablolara aktarılması işlemi yapılmıştır. Bu işlem sırasında metin dosyalarda tutulan verilerin nasıl tutulduğunu açıklayan excel dosyasından

faydalanılmış; bu excel dosyasındaki satır aralıkları dikkate alınarak tüm veriler birbirinden ayrılıp tablolara atılmıştır.

3.ÖSYM'den gönderilen veriler genellikle her öğrenci için tutarlı veriler olmakla birlikte, veriler arasında "null" değerli, belirsiz ya da herhangi bir kayıt içermeyen nitelikler de bulunmaktaydı. Bu nitelikler elle düzeltilmiş olup, öğrencinin diğer verilerinden faydalanılarak doldurulmuştur.

Örneğin cinsiyet niteliğinde 1-Kız, 2- Erkek, 0- Belirsiz gibi üç tür kod mevcuttur. Burada sınav formu doldurulurken yapılan yanlışlıklar göz önüne alınarak "kız ya da erkek" işaretlemelerinin doğru kodlanmadığı durumları gösteren "belirsiz" ifadesi öğrencinin adı ya da lise türüne bakılarak gerektiği gibi doldurulmaya çalışılmıştır.

Başka bir düzenleme de lise türleri ayrıştırılırken yapılmıştır. ÖSYM'den gönderilen verilerde öğrencilerin mezun oldukları lise türleri sadece kodlarıyla gönderilmiş durumdadır.

Öğrencilerin hangi tür liseden mezun olduğu uygulamada kullanılacağından ÖSYM'nin resmi sitesinden indirilen kılavuzun lise türleri tablosu bölümünden faydalanılarak bu kodların hangi lise türlerine ait olduğu bulunmuştur. Böylece her öğrencinin lise türü ve buradan elde edilecek sonuçlar netleştirilmiştir.

ÖSYM'den alınan veriler bu işlemlerden geçirilmiş olup, uygulama için hazır hale getirilmiştir. Bu işlemlerin tablolara aktarımı sırasında öğrencilerin özlük haklarının korunmasına özen gösterilmiş ve kimliklerini açık edecek veriler yerine her bir öğrenciye ayrı ayrı tanımlanan numaralandırma sistemi kullanılmıştır.

4.Bilgi işlem daire başkanlığının bünyesinde bulunan Öğrenci Bilgi Sisteminden de faydalanılmıştır. Buradan öğrencilerin 2006-2007 güz yarıyılı itibariyle ders başarı ortalamalarının tutulduğu veriler alınmıştır. Bu veriler üzerinde herhangi bir veri ön işleme işlemi yapılmasına gerek duyulmamıştır.

Öğrenci verilerini içeren veri tabanı yukarıdaki adımlar takip edilerek hazırlanmıştır. Hazırlanan veri tabanı büyük ölçekli olup; 2003 ve 2004 girişli beş bölümün öğrenci sayısı verilerinden oluşmaktadır. Böylece uygulama için gereken veri tabanı yeteri kadar niteliğe ve elemana sahip bir veri tabanından oluşturulmuştur.

Veri tabanı oluşturulurken öğrenci niteliklerinin bazılarında gruplandırılma yapılması gerekmiştir. Bu ihtiyaç kullanılan verilerin birbirinden çok ayrık ve genetik algoritmalarda kullanılabilir duruma getirilmesi gerektiğini ortaya çıkmıştır.

Veri tabanında yapılan gruplandırmalar şunlardır:

1.Öğrencilerin ÖSYM verilerinden alınan adres illeri Batı- Orta- Doğu olmak üç bölgeye ayrılmış; tablolara aktarılırken öğrencilerin geldikleri yerler bölge bazında ifade edilebilecek hale getirilmiştir.

2.Öğrencilerin ÖSS sonucunda elde ettikleri ülke çapında kaçıncı olduklarını gösteren başarı sıralamaları dört gruba ayrılmıştır. Verileri alınan öğrencilerin ülke genelindeki ÖSS başarı sıralamalarına bakıldığında 3048'den 154.879. sıraya kadar bulunan öğrencilerin olduğu görülmüştür. Gruplandırma işleminde aralıklar şöyledir:

a.3048-10.000

b.10.001-20.000

c.20.001-30.00

d.30.001- …..

3.Veriler arasında öğrencilerin kaçıncı tercihlerini kazandıkları da değerlendirmeye alınmıştır. Kaçıncı tercihlerini kazandıklarını gösteren veriler üç grupta incelenmiştir.

a.1.-8. tercihini kazananlar

b.9.- 16. tercihini kazananlar

c.17.-24 tercihini kazananlar

4.Nitelikler arasında lise türleri de değerlendirmeye alınmış ve bunlarda dört grupta toplanmıştır.

a.Düz lise

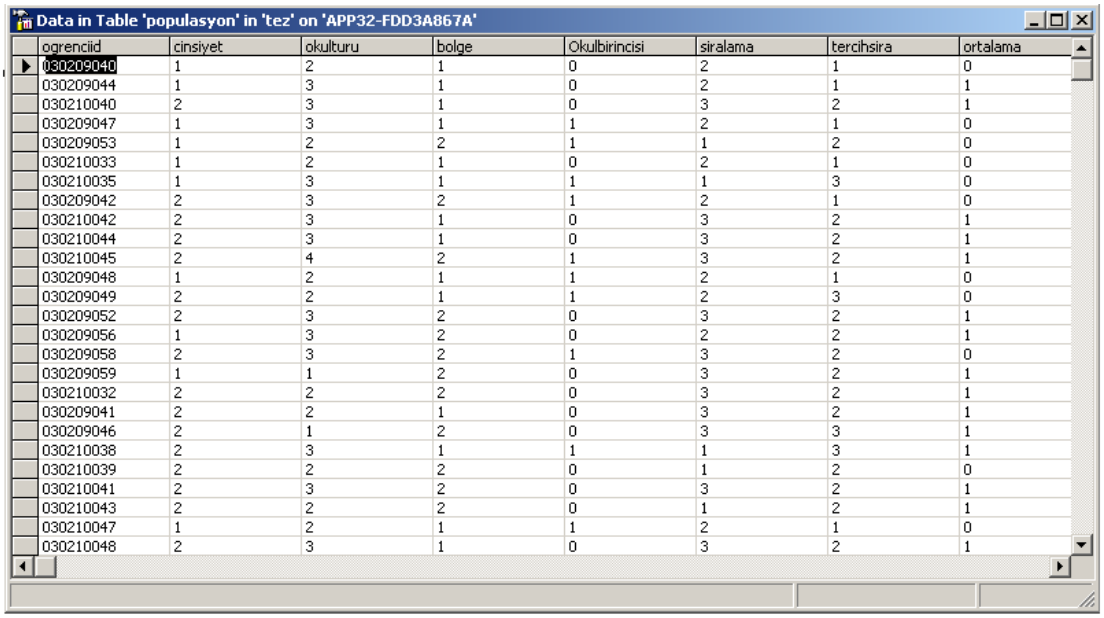
b.Özel lise

c.Anadolu lisesi

d.Fen lisesi

5.Öğrenci bilgi sisteminden alınan, öğrencilerin ders başarı ortalamaları da iki grup şeklinde ifade edilmiştir. Ders başarı ortalaması 2.55' in altında olanlar "başarısız", ders başarı ortalaması 2.55' in üstünde olanlar "başarılı " olarak tabloya aktarılmıştır.

Oluşturulan veri tabanında öğrenci verileri Şekil 4.1' de gösterilen biçimde tutulmaktadır.



Data in Table 'populasyon' in 'tez' on 'APP32-FDD3A867A'

ogrenciid	cinsiyet	okulturu	bolge	Okulbirincisi	siralama	tercihsira	ortalama
030209040	1	2	1	0	2	1	0
030209044	1	3	1	0	2	1	1
030210040	2	3	1	0	3	2	1
030209047	1	3	1	1	2	1	0
030209053	1	2	2	1	1	2	0
030210033	1	2	1	0	2	1	0
030210035	1	3	1	1	1	3	0
030209042	2	3	2	1	2	1	0
030210042	2	3	1	0	3	2	1
030210044	2	3	1	0	3	2	1
030210045	2	4	2	1	3	2	1
030209048	1	2	1	1	2	1	0
030209049	2	2	1	1	2	3	0
030209052	2	3	2	0	3	2	1
030209056	1	3	2	0	2	2	1
030209058	2	3	2	1	3	2	0
030209059	1	1	2	0	3	2	1
030210032	2	2	2	0	3	2	1
030209041	2	2	1	0	3	2	1
030209046	2	1	2	0	3	3	1
030210038	2	3	1	1	1	3	1
030210039	2	2	2	0	1	2	0
030210041	2	3	2	0	3	2	1
030210043	2	2	2	0	1	2	1
030210047	1	2	1	1	2	1	0
030210048	2	3	1	0	3	2	1

Şekil 4.1. Uygulamada kullanılan veritabanı örneği

Veri tabanında kayıtlar veri ön işleme aşamasında yapılan gruplandırma işlemlerine göre tutulmaktadır. Uygulama ara yüzünde gelen kayıtlarda öğrencilerin sadece bölüm bilgileri gösterilmektedir.

4.3. Genetik Algoritmaların Seçim Nedenleri ve Yapısı

4.3.1. Genetik algoritmaların seçim nedenleri

Veri madenciliğinde genetik algoritmalar eniyileme problemlerinde olduğu gibi sınıflandırma yöntemlerinin içerisinde tanımlanmışlardır. Bunun yanı sıra genetik algoritmalar, diğer algoritmaların uygunluğunun değerlendirilmesi içinde kullanılabilir [2].

Bu bölümde genetik algoritmaların bu çalışma için seçilmesinin nedenlerinden, uygulama çalıştırılması ve kodlanması sırasında gözlenen avantaj ve dezavantajlarından, probleme göre farklı tasarlanması gereken genetik algoritmaların bu uygulamadaki yapısından bahsedilecektir.

Bu çalışmada kullanılan genetik algoritmalar, aşağıdaki özellikleri nedeniyle tercih edilmiştir:

a.Genetik algoritmaların hızlı ve iyi sonuçlar üreten iyi bir sınıflandırma ve arama algoritması olması uygulamada amaçlanan veri madenciliği açısından avantaj olarak görülmüştür.

b.Genetik algoritmalar farklı veri tiplerinde iyi sonuçlar üretebilirler; dolayısıyla bireyler doğru tanımlandığı sürece genetik algoritmalarda veri kümesinin tipinden kaynaklanabilecek bir sorun yoktur.

c.Genetik algoritmalar çalıştırılırken verilerin içeriklerinin bir önemi olmadığı için; verilerin birbirleriyle olan farklılıkları da pek önem taşımaz. Örneğin kümeleme yöntemlerinde verilerin birbirine çok yakın ya da çok uzak oluşu algoritmanın başarısını etkilemektedir.

d.Genetik algoritmaların diğer algoritmalarla yapılan karşılaştırmalarında sonuçta elde edilen kurallar bakımından daha verimli olmasının uygulama açısından önemli olduğu düşünülmüştür.
e.Genetik algoritmalar çalışmaya tek bir noktadan değil de, noktalar kümesinden başladıkları için çoğunlukla yerel en iyi çözümde takılıp kalmazlar.

f.Veri madenciliğinde genetik algoritmaların kullanılmasıyla ilginç kuralların keşif sürecinin birçok konuda yapılan çalışmada önemli olduğu görülmüştür.

Genetik algoritmaların yukarıda bahsedilen avantajlarının yanında, veri madenciliğindeki kullanımı sırasında bazı dezavantajları da beraberinde getirmektedir. Bu dezavantajlar şu şekilde sıralanabilir: Genetik algoritmalara göre tanımlanmış olan kural yapılarının kullanıcıya açıklanması ve anlatılması zor olabilir; problemi soyutlamak ve bireylerin temsil edilmesi için kullanılan modeller zor olabilir ve doğru model seçimi önemlidir; uygunluk fonksiyonunun probleme göre doğru seçilmesi ve bireyler bu kritere göre değerlendirilirken dikkatli olunmalıdır; genetik operatörlerin yerinde ve doğru işletilmemesi problemin çözümünde sapmalara neden olabilir.

4.3.2. Uygulamada kullanılan algoritmanın yapısı

Uygulamada kullanılan algoritmanın yapısı iki başlık altında açıklanacaktır:

1.Uygulamada kullanılan genetik algoritmanın temel yapısı ve ilk popülasyonun oluşturulması

2.Algoritma içerisinde kullanılan operatörlerin nasıl çalıştığı ve kullanılan parametreler

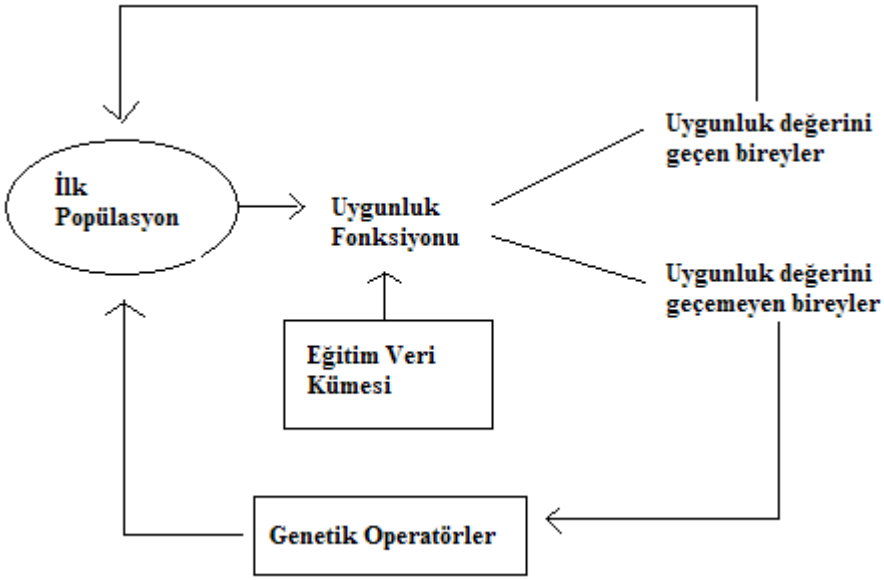
4.3.2.1. Uygulamada kullanılan genetik algoritmanın temel yapısı ve ilk popülasyonun oluşturulması

Bu bölümde uygulamada kullanılan genetik algoritmanın temel yapısı verilecek olup; ara adımlar ikinci başlık altında verilen bölümde ayrıntılarıyla anlatılacaktır. İlk

popülasyonu oluşturan bireylerin kodlanışı, sonlandırma kriteri ve uygunluk fonksiyon değerlerinin hesaplanışı da bu bölümde verilecektir.

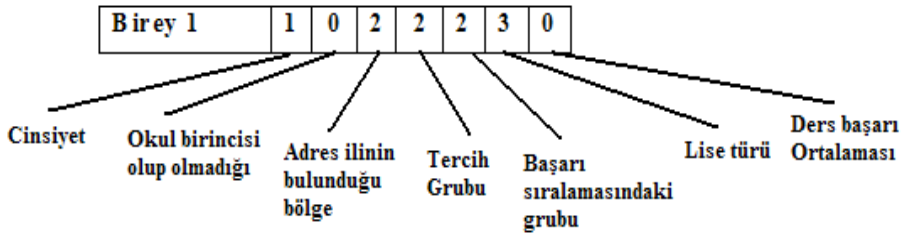
Çalışmada kullanılan genetik algoritma temel genetik algoritma yapısından uzaklaşılmadan uygulanmıştır. Genetik algoritmalarda probleme uygun birey kodlanması, genetik operatörlerin ve parametrelerin doğru seçilme gereği ve anlamlı kuralların çekilebilmesi göz önünde bulundurulmuş; uygulamanın her aşamasında seçilen işlemlerin ve konulan kısıtların uygunluğu gözlenmeye çalışılmıştır.

Uygulamada kullanılan genetik algoritmanın temel yapısı Şekil 4.2'de verilmiştir.



Şekil 4.2. Genetik algoritmanın temel yapısı

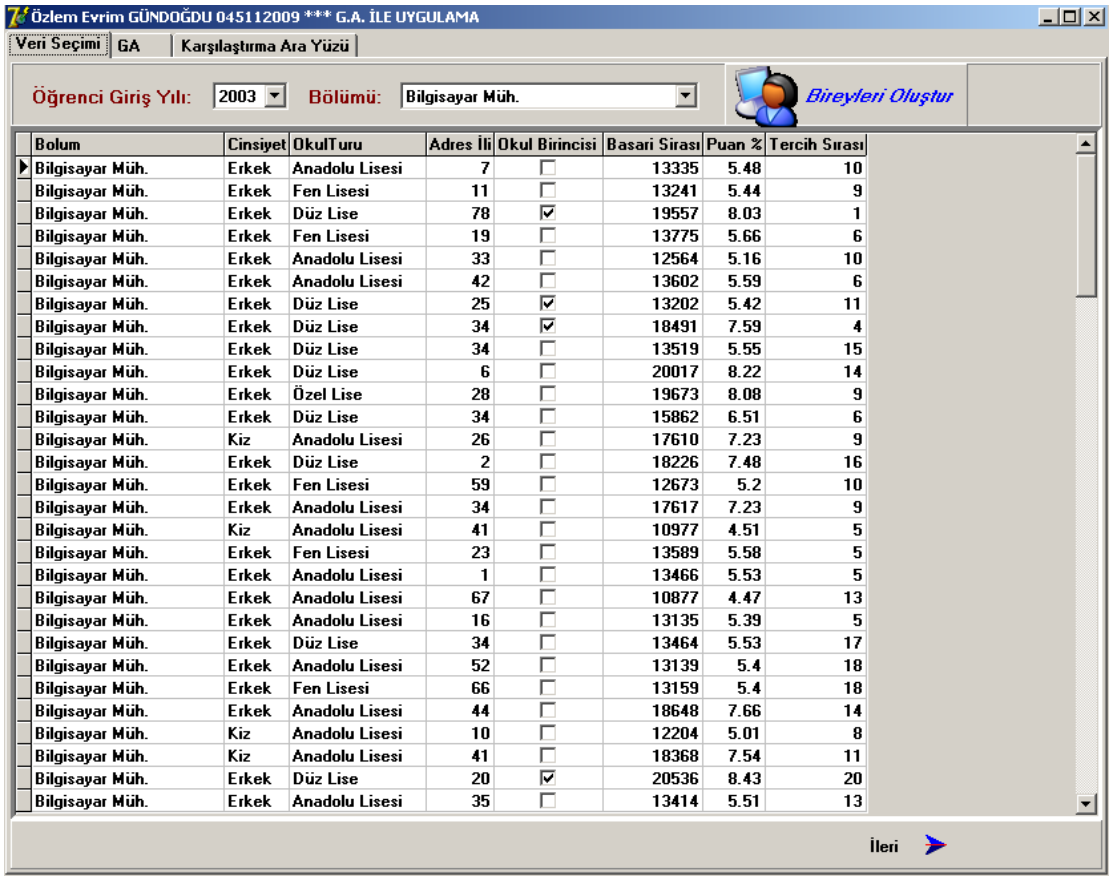
Genetik algoritmalarda ilk yapılması gereken algoritmaya sokulacak olan bireylerin kodlanıp; ilk popülasyonun oluşturulması işlemidir. Bu uygulamada genetik algoritmaya sokulacak olan bireyler oluşturulan tablolardan çekilmektedir. İlk popülasyon bu bireylerden oluşturulmaktadır. Veri tabanındaki tabloların yapısından da anlaşılacağı üzere bireyler tam sayılarla kodlanmıştır.



Şekil 4.3 Tam sayı değerlerle kodlanmış birey

Bireyler Şekil 4.3'te gösterilen kodlama yapısında genetik algoritmaya sokulmuşlardır. İlk popülasyon uygulamanın birinci ara yüzündeki üniversiteye giriş yılı (2003 ya da 2004) ve bölüm (Bilgisayar Müh., Elektronik ve Hab. Müh.,

Endüstri Müh., Mekatronik Müh., Elektrik Müh. Tümü) seçeneklerinden kural keşfinde kullanılacak olan veri kümesinin seçilmesiyle oluşturulmaktadır. Şekil 4.4'te verilen ara yüzde veri tabanından ilk popülasyon oluşturulabilmektedir. Ara yüzde, kullanıcıdan üzerinde çalışacağı öğrenci verilerinin hangi giriş yılı ve hangi bölüme ait olacağı bilgisi alınmaktadır. Burada kullanıcı sadece bir bölüm seçebileceği gibi veri tabanında bulunan beş üzerinde de çalışma yapabilir. Bunun için "Tümü" seçimini yapması yeterlidir. Bu ara yüzde ileri tuşuna bastığında genetik algoritmanın çalıştırılacağı ara yüze geçilir ve seçilmiş olan veri kümesi yani ilk popülasyon üzerinde çalışmaya başlanabilir. Şekil 4.4'te uygulamanın veri seçimi ara yüzü verilmiştir.



Bolum	Cinsiyet	OkulTuru	Adres İli	Okul Birincisi	Basari Sirasi	Puan %	Tercih Sirasi
Bilgisayar Müh.	Erkek	Anadolu Lisesi	7	☐	13335	5.48	10
Bilgisayar Müh.	Erkek	Fen Lisesi	11	☐	13241	5.44	9
Bilgisayar Müh.	Erkek	Düz Lise	78	☑	19557	8.03	1
Bilgisayar Müh.	Erkek	Fen Lisesi	19	☐	13775	5.66	6
Bilgisayar Müh.	Erkek	Anadolu Lisesi	33	☐	12564	5.16	10
Bilgisayar Müh.	Erkek	Anadolu Lisesi	42	☐	13602	5.59	6
Bilgisayar Müh.	Erkek	Düz Lise	25	☑	13202	5.42	11
Bilgisayar Müh.	Erkek	Düz Lise	34	☑	18491	7.59	4
Bilgisayar Müh.	Erkek	Düz Lise	34	☐	13519	5.55	15
Bilgisayar Müh.	Erkek	Düz Lise	6	☐	20017	8.22	14
Bilgisayar Müh.	Erkek	Özel Lise	28	☐	19673	8.08	9
Bilgisayar Müh.	Erkek	Düz Lise	34	☐	15862	6.51	6
Bilgisayar Müh.	Kiz	Anadolu Lisesi	26	☐	17610	7.23	9
Bilgisayar Müh.	Erkek	Düz Lise	2	☐	18226	7.48	16
Bilgisayar Müh.	Erkek	Fen Lisesi	59	☐	12673	5.2	10
Bilgisayar Müh.	Erkek	Anadolu Lisesi	34	☐	17617	7.23	9
Bilgisayar Müh.	Kiz	Anadolu Lisesi	41	☐	10977	4.51	5
Bilgisayar Müh.	Erkek	Fen Lisesi	23	☐	13589	5.58	5
Bilgisayar Müh.	Erkek	Anadolu Lisesi	1	☐	13466	5.53	5
Bilgisayar Müh.	Erkek	Anadolu Lisesi	67	☐	10877	4.47	13
Bilgisayar Müh.	Erkek	Anadolu Lisesi	16	☐	13135	5.39	5
Bilgisayar Müh.	Erkek	Düz Lise	34	☐	13464	5.53	17
Bilgisayar Müh.	Erkek	Anadolu Lisesi	52	☐	13139	5.4	18
Bilgisayar Müh.	Erkek	Fen Lisesi	66	☐	13159	5.4	18
Bilgisayar Müh.	Erkek	Anadolu Lisesi	44	☐	18648	7.66	14
Bilgisayar Müh.	Kiz	Anadolu Lisesi	10	☐	12204	5.01	8
Bilgisayar Müh.	Kiz	Anadolu Lisesi	41	☐	18368	7.54	11
Bilgisayar Müh.	Erkek	Düz Lise	20	☑	20536	8.43	20
Bilgisayar Müh.	Erkek	Anadolu Lisesi	35	☐	13414	5.51	13

Şekil 4.4 Uygulamanın veri seçimi ara yüzü

İlk popülasyon oluşturulduktan sonra genetik algoritmanın adımları işletilmeye başlatılabilir. Bu uygulamada kullanılan genetik algoritmaların adımları şöyledir:

A1. İlk Popülasyonu al.

A2. Popülasyondaki her bireyin uygunluk değerini, uygunluk fonksiyonuna göre hesapla.

A3. Sonlandırma kriterini kontrol et. Kriter sağlanmışsa algoritmadan çık; sağlanmamışsa A4'e git.

A4. Belirlenen uygunluk değerini geçen bireyleri yeni popülasyona ekle; geçemeyen bireyleri genetik operatörler uygulanmak üzere tut.

A4. Uygunluk değerini geçemeyen bireylere genetik operatörleri uygula.

A6. Genetik operatörlerin uygulandığı bireyleri yeni popülasyona ekle.

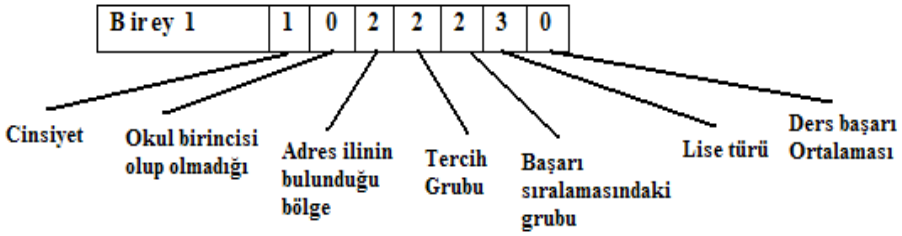
A7. A2'ye git.

Genetik algoritmanın adımlarında da görüldüğü gibi uygunluk fonksiyon değeri algoritmanın temelini oluşturmaktadır. Uygunluk fonksiyonu probleme göre seçilmeli ve popülasyonun iyileştirilmesi için doğru bir eşik değeri alınmalıdır. Bu uygulamada uygunluk fonksiyon değeri şöyle hesaplanmaktadır: Öncelikle ilk popülasyonun belli bir kısmı – bu uygulamada %60'ı alınmıştır – eğitim veri kümesi olarak ayrılır. Eğitim veri kümesi belirlenen kriterde iki sınıfa ayrılır. Uygulama öğrenci verileriyle yapıldığı ve öğrenci başarısının analizi yapılmak istendiği için bu algoritmanın eğitim veri kümesindeki bireyler "başarılı " ve "başarısız " olmak üzere iki sınıfa ayrılmıştır. Bu aşamadan sonra popülasyondaki tüm bireylerin uygunluk değerleri hesaplanabilir. Uygunluk değeri hesaplanırken önce bireyin ders başarı niteliğine bakılır. Birey, iki sınıfa ayırdığımız eğitim veri kümesindeki hangi sınıfa giriyorsa ilk önce eğitim veri kümesinin o sınıfındaki her bir bireyle aynı olan nitelik sayıları bulunur ve toplanır. Bu değer N değeridir. Daha sonra yine eğitim veri kümesinde bireyin ait olmadığı sınıfın bireylerine de aynı işlem uygulanır. Buradan elde edilen toplam değerde M değerimizdir. Bir bireyin uygunluk değeri, N ve M değerleri bulunduktan sonra şöyle hesaplanır:

$$\text{Uygunluk değeri} = N / (M+c) \qquad (4.1)$$

Burada "c" değeri bireylerin uygunluk eşik değerine göre değiştirilebilecek bir sabittir [16].

Uygunluk değerinin hesaplanışı bir örnekle aşağıdaki gibi açıklayarak netleştirilebilir.



Şekil 4.5 Uygunluk değeri hesaplanacak olan birey

Birey 1	0	1	3	2	2	4	1
Birey 2	1	0	1	3	2	1	1
Birey 3	1	0	2	2	3	1	1
Birey 4	1	0	1	2	1	2	0
Birey 5	1	0	3	1	2	3	0
Birey 6	0	0	1	3	2	3	0

Şekil 4.6 Eğitim veri kümesinin elemanları

Görüldüğü gibi eğitim veri kümesinin elemanları “başarılı” niteliğine sahip bireyler başta olmak üzere iki sınıfa ayrılmış durumda verilmiştir. Öncelikle bireyin bulunduğu “başarısız(0)” sınıfından karşılaştırmaya başlayalım.

Cinsiyeti=1 (erkek) , birey 4 ve birey 5 ile eşleşmekte (N=2),
Okul birinciliği=0, birey 4, birey 5 ve birey 6 ile eşleşmekte (N=2+3),
Adres ilinin bulunduğu bölge=2, eşleşen nitelik yok (N=5+0),
Tercih grubu=2, birey 4 ile eşleşmekte (N=5+1),
Başarı sıralamasındaki grubu=2, birey 5 ve birey 6 ile eşleşmekte (N=6+2),
Lise türü=3, birey 5 ve birey 6 ile eşleşmekte (N=8+2),

Buradan N= 10 bulunur.

Bireyi “ başarılı (1)” sınıfıyla karşılaştıralım.
Cinsiyeti=1 (erkek) , birey 2 ve birey 3 ile eşleşmekte (M=2),
Okul birinciliği=0, birey 2 ve birey 3 ile eşleşmekte (M=2+2),
Adres ilinin bulunduğu bölge=2, birey 3 ile eşleşmekte (M=4+1),
Tercih grubu=2, birey 1 ve birey 3 ile eşleşmekte (M=5+2),
Başarı sıralamasındaki grubu=2, birey 1 ve birey 2 ile eşleşmekte (M=7+2),
Lise türü=3, eşleşen nitelik yok (M=9+0),

Buradan M= 9 bulunur. c= 1 olsun.

Birey 1‘ in uygunluk değeri =10/ (9+1) = 1

Bu işlem, ilk popülasyondaki her bireye ve genetik operatörlerin uygulandığı her yeni birey için yapılmaktadır.

Uygulamada kullanılan eşik uygunluk değeri ise şöyledir: İlk popülasyonda uygunluk değeri hesaplanan bütün bireyler en düşükten en yükseğe doğru uygunluk değerlerine göre sıralanırlar. En düşük uygunluk değerini alan popülasyonun %50’sini oluşturan bireyler genetik operatörler uygulanmak üzere seçilir. Algoritmanın son adımına gelindiğinde; yani genetik operatörler uygulanmış olan bireylerinde uygunluk değeri hesaplandıktan sonra diğer bireylerle yeni popülasyon oluşturulur ve aynı işlemler uygulanır.

Uygunluk eşiğinin bu yöntemle uygulanmasının nedeni ilk popülasyondaki her bir bireyin genetik algoritmaya sokulmasını sağlamak ve böylece uygunluk değeri yüksek kurallar elde edebilmektir. Bu yöntemle uygunluk değeri düşük grupta olan bir birey genetik algoritmalar uygulandıktan sonra uygunluk değeri yüksek olan gruba kayabilir ve uygunluk değeri yüksek olan gruptaki başka bireyinde uygunluk değeri düşük olan gruba kayması sağlanabilir. Sonuç olarak genetik algoritmaların geliştirilme amaçlarına uygun olarak popülasyondaki bir çok bireye genetik operatörler uygulanıp uyumluluğu iyi değerde olan bireylerin sayısı arttırılabilir.

4.3.2.2. Algoritmada kullanılan genetik operatörler ve parametreler

Genetik algoritmalarda kullanılan operatörler ve parametrelerden Bölüm 3'de bahsedilmişti. Uygulamada kullanılan genetik operatörler ve parametreler ise şunlardır:

Seçim operatörü: Seçim operatörü daha öncede anlatıldığı gibi çaprazlama işlemi uygulanacak olan ebeveynlerin seçimi için kullanılır. Seçim yöntemleriyle, genetik algoritmaların temelini oluşturan "en iyi uyumluluğa sahip bireyin yaşaması" ilkesinin devamı sağlanmaya çalışılır. İyi bireylerin kaybedilmeden yeni nesillere aktarılması ve genetik işlemciler için seçilecek olan bireylerin uyumluluklarının yüksek olması genetik algoritmalarda çözüme ulaşmayı kolaylaştırır. Bu nedenlerle seçim operatörü yöntemlerinden probleme en uygun olanının seçilmesi gerekir.

Bu uygulamada seçim operatörü olarak Rulet Tekerleği yöntemi seçilmiştir. Bu yöntemde uygunluk değeri yüksek olan bireylerin ebeveyn olarak seçilme olasılıklarının yüksek olduğu görülmüş; böylece daha iyi uygunluk değerine sahip yeni bireylerin üreme olasılıklarının artacağı düşünülmüştür.

Çaprazlama: Çaprazlama işlemi tek noktalı çaprazlama yöntemi kullanılarak yapılmıştır. Bu yöntem uygulanırken kesim noktası 3. ve 4. kromozomların arası olarak belirlenmiştir. Rulet tekerleği yöntemiyle seçilerek ebeveynler popülasyonun rasgele %60' ı alınarak oluşturulmuş bir kümeden belirlenmektedir.

Mutasyon: Mutasyon işlemi, çaprazlama işlemi sonrası gelen her yeni bireye uygulanmakta olup, mutasyon oranına göre rasgele seçilen kromozomlara uygulanmaktadır.

Mutasyon işlemi uygulanırken hangi kromozoma uygulandığı önem taşımaktadır. Mutasyon işlemi uygulanacak olan kromozoma rasgele verilecek olan değer sadece içinde bulunduğu grubun değerleri arasından seçilebilmektedir. Bu uygulamada böyle bir kısıt koyma gereği duyulmuştur. Örneğin mezun olunan lise türü dört gruba ayrılırken, adres ili bölgeleri üç gruba ayrılmış durumdadır. Mutasyon işlemi lise

türünü belirten kromozoma uygulanacaksa (1, 2, 3, 4) değerlerinden biri rasgele seçilirken, adres ili bölgelerinin tutulduğu kromozoma uygulanacaksa (1(batı), 2(orta), 3(doğu)) değerlerinden biri rasgele gelmektedir. Böylece mutasyon işlemi sonrasında anlamız ve karşılığı olmayan niteliklere sahip bireylerin oluşması engellenmiştir.

1.Uygulamada başlangıç popülasyonundaki birey sayısı kullanıcının seçtiği bölüme ve yıla göre değişmektedir. Bu parametredeki değişiklik uygulamanın çalışma hızını etkilemektedir.

2.Çaprazlama ve mutasyon oranı literatür taramalarında karşılaşılan değerlerden yola çıkılarak seçilmiştir. Uygulama kodunda sabit olarak tanımlanmaktadır. Çaprazlama oranı 0.6, mutasyon oranı 0.1 olarak seçilmiştir.

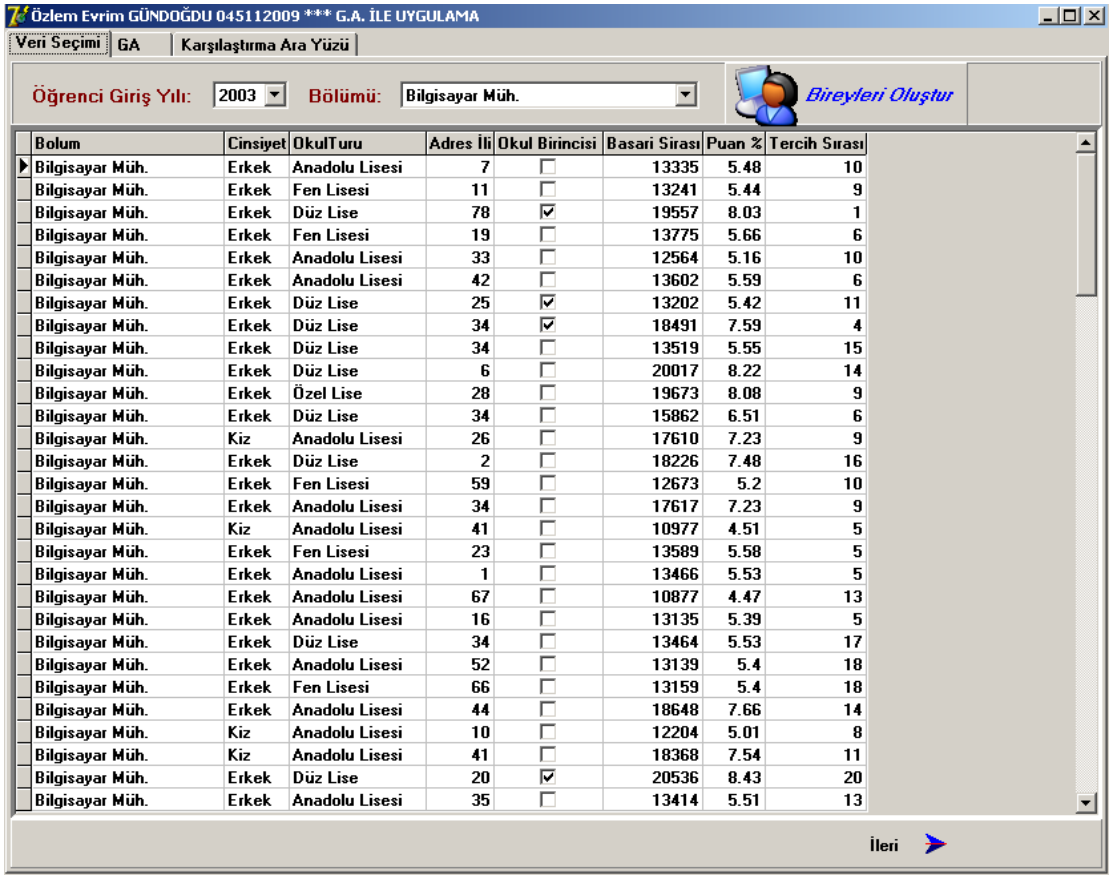
3.Uygulamada iterasyon sayısı seçilebilir bir değer olarak ara yüze koyulmuştur. İterasyon sayısında alınan farklı değerlerin genellikle kural bulmada etkili olduğu görülmüştür.

4.4. Uygulama Ara Yüzlerinin İşlevleri ve Sonuçların Değerlendirilmesi

Uygulamanın ara yüzleri birbirinden farklı işlemler için tasarlanmıştır. Her bir ara yüzden bir sonraki ara yüze ileri-geri tuşları ile geçilebilmekte ve sonraki adım için gerekli görülen değişiklikler yapıldıktan işleme devam edilebilmektedir.

4.4.1 Veri seçimi ara yüzü

Uygulamanın ilk ara yüzünde ilk popülasyonun belirlenebilmesi gereken kriterlerin kullanıcı tarafından seçilmesi gerekmektedir.



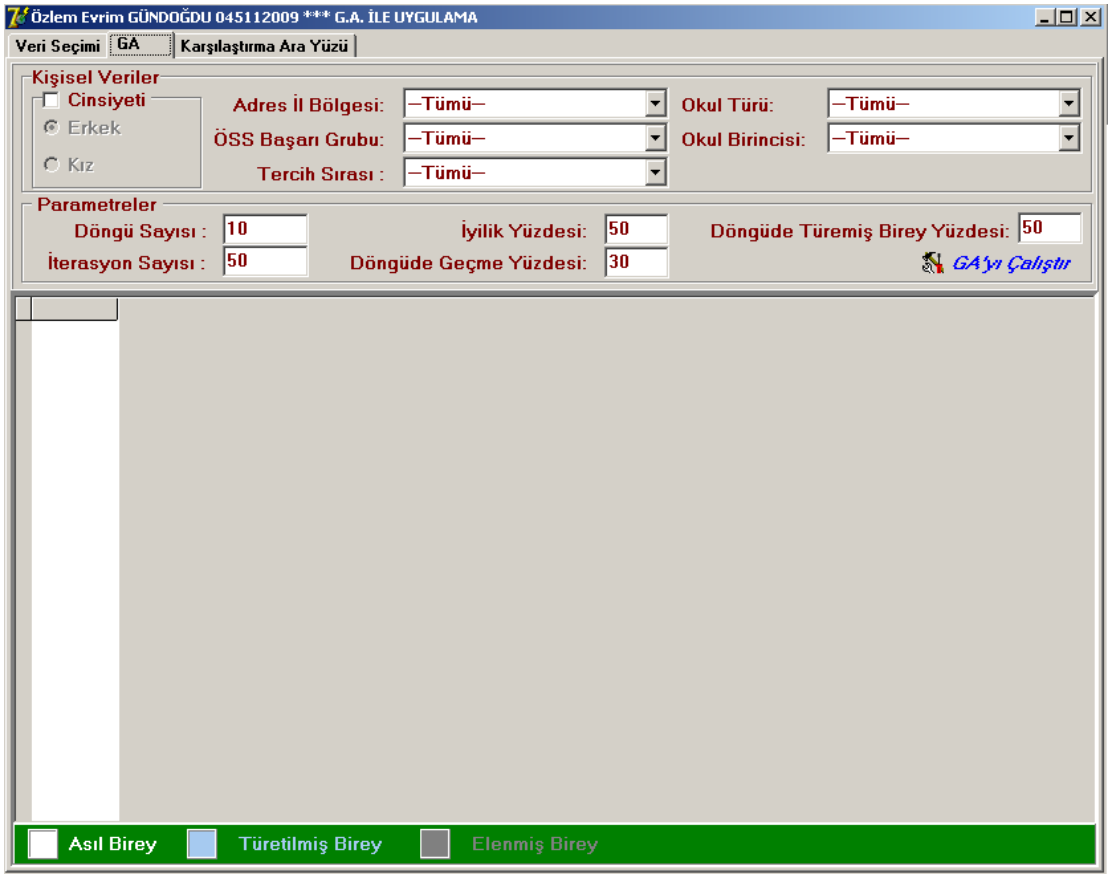
Bolum	Cinsiyet	OkulTuru	Adres İli	Okul Birincisi	Basari Sirasi	Puan %	Tercih Sırası
Bilgisayar Müh.	Erkek	Anadolu Lisesi	7	☐	13335	5.48	10
Bilgisayar Müh.	Erkek	Fen Lisesi	11	☐	13241	5.44	9
Bilgisayar Müh.	Erkek	Düz Lise	78	☑	19557	8.03	1
Bilgisayar Müh.	Erkek	Fen Lisesi	19	☐	13775	5.66	6
Bilgisayar Müh.	Erkek	Anadolu Lisesi	33	☐	12564	5.16	10
Bilgisayar Müh.	Erkek	Anadolu Lisesi	42	☐	13602	5.59	6
Bilgisayar Müh.	Erkek	Düz Lise	25	☑	13202	5.42	11
Bilgisayar Müh.	Erkek	Düz Lise	34	☑	18491	7.59	4
Bilgisayar Müh.	Erkek	Düz Lise	34	☐	13519	5.55	15
Bilgisayar Müh.	Erkek	Düz Lise	6	☐	20017	8.22	14
Bilgisayar Müh.	Erkek	Özel Lise	28	☐	19673	8.08	9
Bilgisayar Müh.	Erkek	Düz Lise	34	☐	15862	6.51	6
Bilgisayar Müh.	Kiz	Anadolu Lisesi	26	☐	17610	7.23	9
Bilgisayar Müh.	Erkek	Düz Lise	2	☐	18226	7.48	16
Bilgisayar Müh.	Erkek	Fen Lisesi	59	☐	12673	5.2	10
Bilgisayar Müh.	Erkek	Anadolu Lisesi	34	☐	17617	7.23	9
Bilgisayar Müh.	Kiz	Anadolu Lisesi	41	☐	10977	4.51	5
Bilgisayar Müh.	Erkek	Fen Lisesi	23	☐	13589	5.58	5
Bilgisayar Müh.	Erkek	Anadolu Lisesi	1	☐	13466	5.53	5
Bilgisayar Müh.	Erkek	Anadolu Lisesi	67	☐	10877	4.47	13
Bilgisayar Müh.	Erkek	Anadolu Lisesi	16	☐	13135	5.39	5
Bilgisayar Müh.	Erkek	Düz Lise	34	☐	13464	5.53	17
Bilgisayar Müh.	Erkek	Anadolu Lisesi	52	☐	13139	5.4	18
Bilgisayar Müh.	Erkek	Fen Lisesi	66	☐	13159	5.4	18
Bilgisayar Müh.	Erkek	Anadolu Lisesi	44	☐	18648	7.66	14
Bilgisayar Müh.	Kiz	Anadolu Lisesi	10	☐	12204	5.01	8
Bilgisayar Müh.	Kiz	Anadolu Lisesi	41	☐	18368	7.54	11
Bilgisayar Müh.	Erkek	Düz Lise	20	☑	20536	8.43	20
Bilgisayar Müh.	Erkek	Anadolu Lisesi	35	☐	13414	5.51	13

Şekil 4.6 Veri seçimi ara yüzü

Kullanıcı bu ara yüzde üzerinde çalışmak istediği öğrenci kümesini seçmektedir. Veri seçimi olarak iki kriter vardır; kullanıcı isterse tüm bölümler üzerinde de çalışabilir.Bunun için "Tümü " seçimini yapması yeterlidir. Veri seçimi ara yüzünden, genetik algoritmanın çalıştırılacağı ara yüze ileri tuşuna basılarak geçilebilir.

4.4.2. Genetik algoritma ara yüzü

Uygulamanın ikinci ara yüzünde olan genetik algoritmanın çalıştırılması sağlamak amacıyla tasarlanmıştır. Genetik algoritma ara yüzünde kullanıcıya aradığı niteliklere sahip öğrenci profilinden oluşan kurallar sunulmaktadır. Bu ara yüzü Şekil 3.7.' te verilmiştir.



Şekil 4.7 Genetik algoritma ara yüzü

Uygulamanın bu ara yüzünden kullanıcıdan aranılan kurallar ve genetik algoritmanın nasıl çalıştırılacağına dair nitelik ve parametreler girmesi istenmektedir.

Öncelikle aranılacak kural nitelikleri seçilir. Kullanıcı, ara yüzde verilen niteliklerin hepsini ya da sadece bir kaçını seçebilir. Genetik algoritma bir önceki ara yüzde seçtiği veri kümesini ilk popülasyon olarak alıp; kullanıcının seçtiği niteliklere uyan kuralları getirecektir.

İkinci aşamada, kullanıcının genetik algoritma için konulması gereken kriterler seçmesi gerekir. Ara yüzde görülen değerler,başlangıç değerleri (default) olarak gelmektedir. Bu ara yüzdeki parametrelerin ne maksatla seçildiği aşağıda açıklanmıştır:

Döngü sayısı: Genetik algoritmanın kaç defa çalışacağını belirler.

İterasyon sayısı: Genetik algoritmada kullanılan sonlandırma kriteridir. Bu sayı yakalandığında, algoritma bir döngüyü bitirmiştir ve eğer girilen döngü sayısına erişilmemişse genetik algoritmayı tekrar çalıştırır.

İyilik yüzdesi: Bu değer uygunluk değerinin hesaplanması için girilen bir değeridir. Uygunluk değerinin hesaplanma şekli bölüm "3.3.2.2. Algoritmada Kullanılan Genetik Operatörler ve Parametreler" ayrıtında örnekleriyle açıklanmıştır.

Döngüde geçme yüzdesi: Bu parametre genetik algoritmada yakalanan kuralların doğruluklarının değerlendirilebilmesi için konmuştur. Buna göre 10 döngü sayısı için çalıştırılan genetik algoritmada, bu döngüler içinde yakalanan kuralların her döngüde kaç defa geçtiği değeri tutulur. Bu değerler toplamı (her kural yakalanan kural için ayrı ayrı değerlendirilir) eğer "döngüde geçme yüzdesi" değerini geçemiyorsa genetik algoritmanın yakaladığı bu kuralın doğru olamayacağı kabul edilir. Böylece yakalanan kurallar arasında yapılan bu değerlendirme yöntemiyle genetik algoritmanın kural yakalamada gösterdiği performans değerlendirilmiş olur.

Döngüde türemiş birey yüzdesi: Döngüde türemiş birey oranı girilen bu değerinde altındaysa kurallar ekrana getirilmez.

Uygulamanın genetik algoritma ara yüzündeki bir ayrıntıda ekrana gelen tüm kuralların uygunluk değeri kriterini geçmiş olan bireyler olduklarıdır. Uygunluk değeri kriterini geçememiş olan bireyler ekranda gösterilmez.

Bu ayrıntı uygulamada kuralların doğruluğunun değerlendirmesini kolaylaştırmak için verilmiştir. Kural değerlendirmesini kolaylaştıran başka bir özellik ise seçilen uygun olarak ara yüze gelen kuralların farklı renkte olmasıyla sağlanmıştır.

Bu uygulamada ara yüzde gösterilen kurallar belirlenen üç sınıftan birine aittir. Sınıflan şunlardır:
Asıl birey: Genetik algoritmanın çalıştırılması sonucunda bazı kurallar, uygunluk değerleri her döngüde uygunluk kriterini geçecek değerde olduğu için genetik operatörlere hiç takılmadan gelebilirler. Bu kurallar "asıl birey" sınıfına dahil edilir.

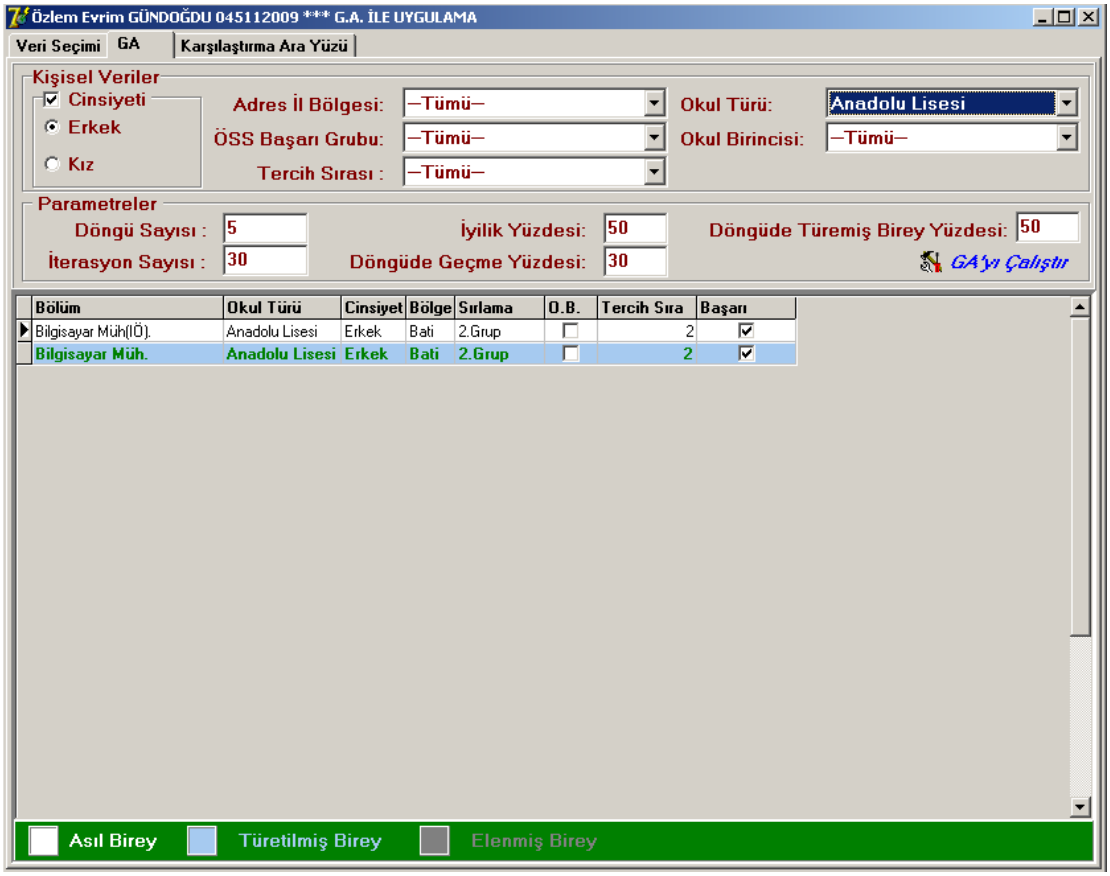
Elde edilen asıl birey sınıfındaki kurallar gerçek verilerdir ve beyazla renklendirilmişlerdir.

Türetilmiş birey: Genetik algoritma çalıştırıldığında bir şekilde uygunluk değerini algoritmanın bir yerinde geçemeyen; sonradan genetik operatörler sayesinde bu kriteri geçen kurallardır. Bu kurallar maviyle renklendirilmişlerdir ve genetik algoritmanın ürettiği doğruluk oranı yüksek kurallardır. Doğruluk oranları yüksektir denebilmesinin nedeni döngüde geçme yüzdesi parametresine uyan kurallar olmalarıdır.

Elenmiş birey: Elenmiş bireyler döngüde geçme yüzdesi parametresini sağlayamayan; dolayısıyla konulan doğru kural olma kriterini yakalayamamışlardır; fakat uygunluk kriterini geçmiş olan bireylerdir. Bu bireyler griyle renklendirilmişlerdir.

Uygulamanın genetik algoritma ara yüzü çalıştırıldığında nasıl sonuçlar elde edilmiştir? Uygulama çalıştırıldığında görülmüştür ki, seçilen ilk popülasyona bağlı olarak bölümler arasında öğrenci profilleri çok farklıdır. Seçilen öğrenci nitelikleri her bölüm için aynı değerleri vermemektedir. Bu da bölümler arasındaki puan, öğrenci sayısı, tercih edilme sıraları gibi niteliklerden kaynaklanmaktadır. Buradan yola çıkılarak her bölüm için genel bir öğrenci profili çıkartılabileceği gibi gelecek olan öğrencilerin de bu profillere uygunluğu değerlendirilebilir.

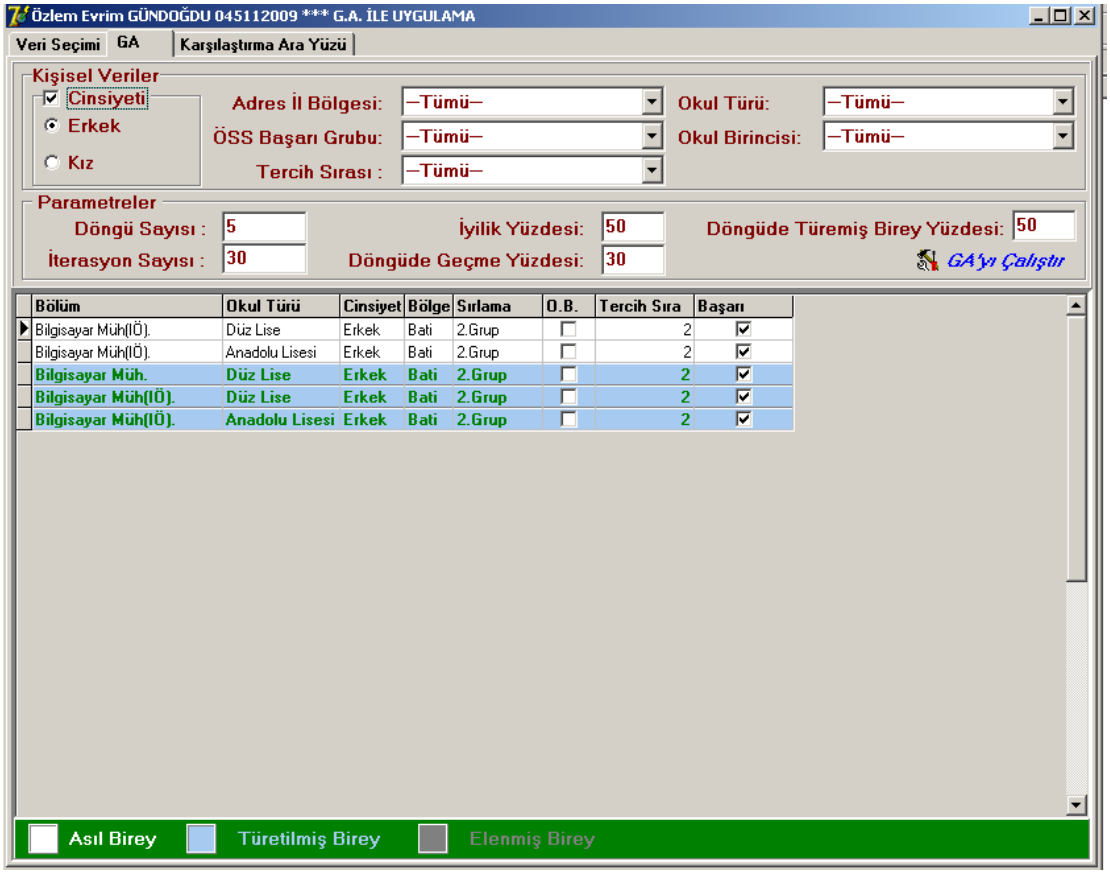
Uygulamada kural doğruluğunun değerlendirilmesi, genetik algoritma tarafından oluşturulmuş bir çok bireyin elenmesine neden olmasına rağmen algoritmanın aynı kuralı birden fazla yakalayabilme performansının değerlendirilebilmesi açısından bu değerlendirmenin yapılaması yerinde olmuştur.



Şekil 4.8 Uygulama ara yüzü-1

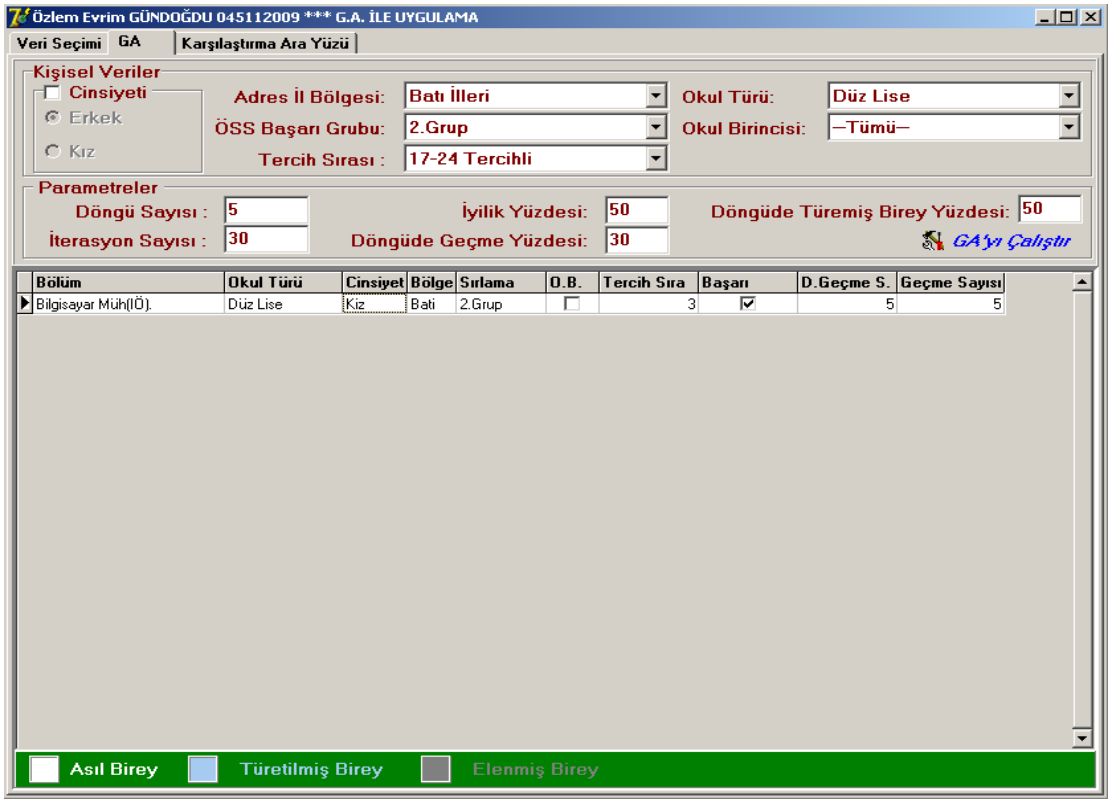
Bu ara yüzde Bilgisayar mühendisliği bölümünde okuyan erkek ve Anadolu lisesi mezunu olan öğrencilerin profilleri görülmektedir. Buna göre bu niteliklere sahip öğrenciler bu bölümde başarılı olabilmekte ve diğer nitelikleri de birbiriyle örtüşmektedir. Burada genetik algoritmanın doğru kural ürettiği görülebilmektedir. Ekrana gelen kurallardan biri gerçek yani hiç genetik operatörlerin işlenmediği uygunluk kriterini beş döngüde de geçen birey; diğeri genetik algoritmanın ürettiği ve doğruluk ve uygunluk kriterini geçen bireydir.

Uygulamada Bilgisayar mühendisliği öğrencilerinden sadece erkek olma niteliği seçildiğinde ise yine birbiriyle tutarlı kurallar elde edilebilmektedir.



Şekil 4.9 Uygulama ara yüzü-2

Uygulamada seçilen niteliklerin sayısı arttırıldığında ise daha farklı sonuçlar elde edilmiştir. Şekil 4.10'da da görüldüğü gibi genetik algoritma bu niteliklere sahip uygunluk kriterini geçmiş olsalar bile doğruluk kriterini geçen bireyler üretemediği için ekrana sadece asıl birey sınıfına giren kural gelmiştir.

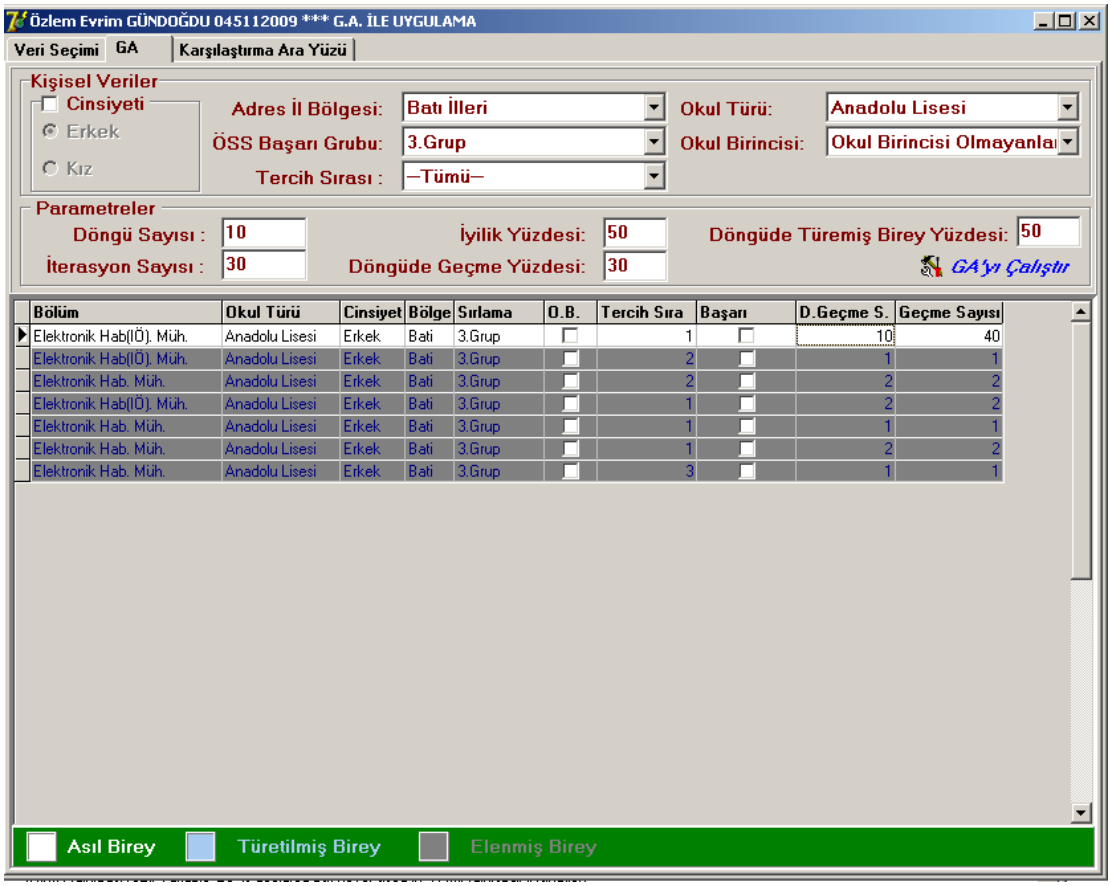


Şekil 4.10 Uygulama ara yüzü-3

Bir başka uygulamada Şekil 4.11' de görülen sonuçlar yakalanmıştır.

Şekil 4.11 Uygulama ara yüzü-4

Uygulama Şekil 4.12'deki gibi çalıştırıldığında ise şu kurallar üretilmiştir.

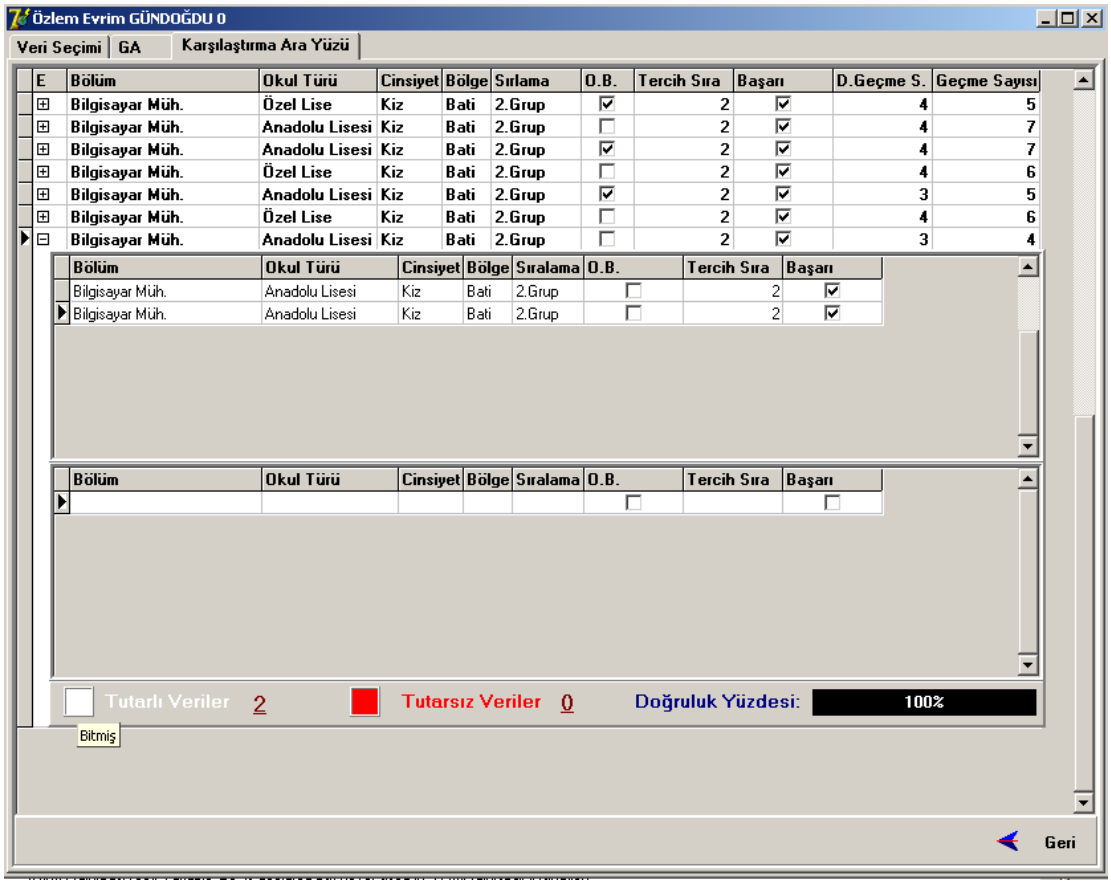


Şekil 4.12 Uygulama ara yüzü-5

Bu ara yüzde de Elektronik ve Haberleşme Mühendisliği bölümü öğrencileri veri kümesiyle çalışılmıştır. Bu uygulamada genetik algoritma uygunluk kriterini geçen ve asıl bireyle örtüşen bir çok kural yakalamıştır. Bu bireyler doğruluk kriterini yakalamadıkları için elenmiş birey sınıfına dahil edilmişlerdir.

4.4.3. Karşılaştırma ara yüzü

Uygulamanın üçüncü ve son ara yüzü olan karşılaştırma ara yüzünde, genetik algoritmalarla üretilmiş olan kurallarla veri tabanında bulunan bireyler karşılaştırılmakta ve genetik algoritmanın ürettiği kuralların doğruluğu ve kabul edilebilir olup olmadıklarının karşılaştırılması yapılmaktadır. Şekil 4.13'de karşılaştırma ara yüzü verilmiştir.



Şekil 4.13 Karşılaştırma ara yüzü-6

Karşılaştırma ara yüzünde amaçlanan genetik algoritmaların ne derece doğru kurallar üretebildiğini ve üretilen bu kuralların gelecek için yapılabilecek tahminlerde kullanılabilirliklerini görmektir. Ara yüzde elde edilen sonuçlar göstermiştir ki genetik algoritmalar probleme uygun, doğru kısıtlar ve parametreler kullanıldığında doğru sonuçlar verebilmekte; hatta Şekil 4.13' te görüldüğü gibi %100 oranında tutarlı performans gösterebilmektedirler. Bu ekran görüntüsünde en üstte gerçek veri tabanı; hemen altında, beyaz zemin üzerinde de genetik algoritmanın ürettiği tutarlı veriler görülmektedir. Ekran sonunda tutarlı ve tutarsız veri sayısıyla beraber genetik algoritmaların performansını gösteren doğruluk yüzdesi de değerlendirme de yardımcı olacağı düşünülerek gösterilmektedir.

5. SONUÇLAR VE ÖNERİLER

Sonuçlar ve öneriler kısmında, tez çalışması ve geliştirilen uygulama hakkında bilgi verilmiş, uygulamanın sonuçları özetlenmiştir. Veri madenciliğinde genetik algoritmaların kullanılmasının avantaj ve dezavantajları açıklanıp, uygulamanın geliştirilebilmesi için ileride yapılabilecek öneriler sıralanacaktır.

Veri madenciliği günümüzde bir çok konuda kullanılmaktadır. Günümüzde geliştirilen çeşitli amaçlı bilgisayar yazılımları, veri madenciliği işlemi yapan araçlar da içermeye başlamıştır. Veri madenciliğiyle eğitim alanında da çalışmalar yapılmış ve verimli sonuçlar elde edilmiştir. Bu tezde de öğrencilerin geçmiş kayıtları değerlendirilerek, öğrenci başarı durumlarının analizini yapan bir araç geliştirilmiştir.

Bu çalışmayla veri madenciliğinde, bir sınıflandırma tekniği olarak geçen genetik algoritmalar incelenerek, öğrenci verileri üzerinde geliştirilen yazılımla verilerin analizinin yapılması amaçlanmıştır. Geliştirilen uygulamayla eğitimcilerin aradıkları niteliklere sahip öğrenci profillerini değerlendirebilmeleri sağlanmıştır. Çalışmanın başında her ne kadar aynı fakültede okuyan öğrenci verilerinin birbirine yakın olabileceği düşünülse de, veritabanı incelendiğinde öğrenci niteliklerinin kendi bölümlerinde birbirine yakın olabildiği; ancak tüm bölümler işleme alındığında birbirinden çok farklı durumların olabileceği gözlenmiştir. Çalışmada kullanılan öğrenci verileri Kocaeli Üniversitesi Bilgi İşlem Daire Başkanlığından alınmıştır.

Uygulama üç ara yüz kullanılarak hazırlanmış olup; her ara yüzde öğrenci başarı analizi için gereken farklı adımlar yapılmaktadır. Birinci ara yüzde kullanılmak istenen öğrenci veri kümesi seçilmekte, ikinci ara yüzde seçilen bu veri kümesi veri madenciliğinde genetik algoritmalar kullanılarak elde edilen kuralların doğruluğu incelenebilmekte ve kuralların farklı renklerle gösterilmesiyle genetik algoritmaların oluşturduğu kural tipleri değerlendirilebilmektedir. Üçüncü ara yüzde ise bulunan kurallar gerçek veritabanı ile karşılaştırılmaktadır. Böylece kullanıcıya Analiz

yapabilmesi için iki farklı yöntem sunulmaktadır. Uygulamanın sonuçlarına bakıldığında genetik algoritmaların döngü sayısı artırıldıkça çıkan yakalanabilen kural sayılarının azaldığı; kural sayısındaki azalmaya rağmen karşılaştırma ara yüzünde gerçek verilerle yapılan karşılaştırmada doğru sonuçlara ulaşılabildiği görülebilmektedir. Ayrıca uygulamada girilen parametre değerleri başlangıç değerleriyle doğru çalışabilmekte, bu da genetik algoritmaların çalışması için seçilen bu değerlerin uygun seçildiğini göstermektedir.

Uygulama en iyi sonuçları döngü sayısı: 5, iterasyon sayısı: 30, iyilik yüzdesi 50, döngüde geçme yüzdesi: 30, döngüde türemiş birey yüzdesi:50 seçildiğinde vermektedir. Bu değerlerle elde edilen türetilmiş bireylerin karşılaştırma ara yüzündeki sonuçları daha yüksektir.

Bu çalışmada veriler genetik algoritmalar yardımıyla değerlendirilmiştir. Genetik algoritmaların çalıştırılması sırasında kullanılan bazı parametreleri kullanıcının seçmesi için ara yüzde parametre girişi sağlanmıştır. Genetik algoritmalar bu çalışmaya aşağıdaki özelliklere uyumu nedeniyle tercih edilmiştir:
1.Genetik algoritmalar büyük veri kümelerinde iyi bir performansla çalışmaktadır. Uygulamada da kullanılan popülasyonun büyük bir veri kümesi olması nedeniyle avantaj sağlanabilmiştir.

2.Algoritma farklı sayıda nitelik seçimlerinde doğru çalışabilmektedir.

3.Genetik algoritmalar farklı veri gruplarına uygulanabilmektedirler. Bu özelliği nedeniyle geliştirilen uygulamanın farklı amaçlarla kullanılabileceği düşünülmektedir. Farklı alanda yapılacak çalışmalarda kullanılabilecek şekilde kodlanmıştır.

4.Bazı parametrelerin kullanıcı tarafından girilebiliyor olması genetik algoritmaların performansının değerlendirilmesini esnek hale getirmiştir.

Uygulama geliştirilirken genetik algoritmaların bazı dezavantajları da gözlenmiştir. Genetik algoritmalarda kullanılan parametrelerin doğru seçilmesi, algoritmanın

performansını doğrudan etkilemektedir. Bu nedenle genetik algoritmalar kullanılırken problem iyi analiz edilmeli ve kullanılacak parametrelerin seçimi doğru yapılmalıdır.

Geliştirilen, öğrenci verileri analiz aracının eğitimcilere öğrencilerin değerlendirilmesi, ders planı hazırlanması ve öğrenci başarı seviyelerinin değerlendirilmesi yapılırken yardımcı olacağı düşünülmektedir. Böylece geçmiş yıllardan elde edilen veriler ile öğrencilerin durumu ve bu niteliklere uyan yeni gelen öğrencilerin başarılarının analizi yapılabilecektir.

Geliştirilen uygulamanın gerçek verilerle çalıştırılması değerlendirme yapılırken bir çok açıdan avantajlar sağlamıştır. Uygulamanın ürettiği karşılaştırma ara yüzü ile sonuçların veritabanıyla karşılaştırılması ve algoritmanın performansı ölçülebilmesi sağlanmıştır. Genetik algoritmalarla elde edilen kuralların doğruluğu değerlendirildiğinde, gerçek veri tabanıyla yapılan karşılaştırmalarda kuralların %100 doğruluk oranıyla elde edilebilir olduğu Şekil 4.13'te de görülebilmektedir.

Sonuçları en çok kullanıcı tarafından girilen döngü sayısının değerinin etkilediği ve bu değerin yükseltilmesiyle türetilmiş bireylerin daha az oluştuğu gözlenmiştir. Uygulamanın bu özelliği nedeniyle genetik algoritmalarda döngü sayısının etkili olduğu söylenebilir.

Bu uygulamada, farklı uygunluk fonksiyonu hesaplamaları gibi yeni eklemeler yapılarak öğrenci verilerinin bir çok konuda daha etkin kullanılmasının sağlanabilir. Ayrıca yazılım farklı veri tabanları kullanılmasına uygun olarak geliştirildiği için farklı amaçlı çalışmalar içinde kullanılabilecek durumdadır.

Geliştirilen uygulamayla görülmüştür ki, genetik algoritmalar iyi bir sınıflandırma algoritmasıdır. Genetik algoritmalarla yakalanan kurallarla, gerçek veritabanı arasında yapılan karşılaştırma çoğu zaman doğru sonuçlar elde edildiğini göstermektedir.

KAYNAKLAR

[1] Alataş, B., ”Veri Madenciliği Konulu Sunumu ”, (2003).

[2] Han, J.,Kamber, M., ”Data Mining:Concepts and Techniques”, ***Morgan Kaufmann Publishers*** (Ağustos 2001).

[3] Mitchell, M., ”An Introduction To Genetic Algorithms “, ***The MIT Press***, (1998).

[4] Vahaplar, A., İnceoğlu, M. M., “Veri Madenciliği ve Elektronik Ticaret “ www.ege.edu.tr, (**Ziyaret tarihi**: **23.01.2005**)

[5] Baykal, N., Veri Tabanı ve Veri Madenciliği konulu sunum, ***Tıp Bilişimi Güz Okulu,*** Ekim, (2003), www.turkmia.org

[6] Werner, J.C., Fogarty, T.C., “Genetic Algorithm Applied in Clustering Datasets” ***Technical Report***, London South Park Univercity, (2001).

[7] Saraç, Ö.S., Alan, Ö., Bahçeci, E., Leblebicioğlu, K., “Genetik Algoritmayla Çoklu Dizi Hizalama”, ***Sinyal İşleme ve Uygulama Kurultayı*** (2003).

[8] Kaya, M., Alhajj, R., “Genetic Algorithm Based Framework for Mining Fuzzy Association Rules”, www.elsevier.com/locate/fss (2004).

[9] Söke, A., “Genetik Algoritma ve Benzetilmiş Tavla ile İki Boyutlu Giyotinsiz Kesme Problemlerine Olasılıksal Yaklaşım”, Yüksek Lisans Tezi, ***Kocaeli Üniversitesi Fen Bilimleri Enstitüsü***, İzmit, (2003).

[10] Emel, G.G., Taşkın, Ç., “Genetik Algoritmalar ve Uygulama Alanları”,Uludağ ***Üniversitesi İktisadi ve İdari Bilimler Dergisi***, 129-152, (2002).

[11] Özcan, E., Alkan, A., “Çok Nüfuslu Kararlı Hal Genetik Algoritması Kullanarak Otomatik Çizelgeleme”,

[12] Romao, W., Freitas, A. A., S. Gimenes, I. M., “Discovering Interesting Knowledge From A Science And Technology Database With A Genetic Algorithm“, ***Science Direct*** 121-137 (2004).

[13] Alataş, B., Arslan, A., , “Birliktelik Kurallarının Madenciliği İçin Genetik Algoritma ve Bulanık Küme Tabanlı Yeni Bir Yaklaşım”, ***F. Ü. Fen ve Mühendislik Bilimleri Dergisi***, 42-51 (2005).

[14] Chen, T., Hsu, T., (2006) , “A GA Based Approach For Mining Breast Canser Pattern”, ***Expert Systems With Applications 30***, 674-681, (2006).

[15] Jourdan, L., Dhaenens, C., Talbi,, E., , "A Genetic Algorithm For Feature Selection in Data-Mining For Genetics", ***4th Metaheuristics International Couference***, 29-33 (2001).

[16] Roiger, R. J., Geatz, M. W., "Data Mining: A Tutorial – Based Primer", ***Addison Wesley***, (2003).

[17] Romero, C., Ventura, S., Bra, P., Castro, C., "Discovering Prediction Rules in AHA! Courses", ***Johnstown, PA***, USA, 25-34, (2003).

[18] Niimi, A., Tazaki, E., , "Rule Discovery Technique Using Genetic Programming Combined with Apriori Algorithm", ***Springer Verlag Berlin***, 273-278, (2000).

[19] Defalce, I., Cioppa, A.D.,Tarantino, E., , "Discovering İnteresting Classification Rules with Genetic Programming", ***Applied Soft Computing***, V-1,N-4, 257-269, www.elsevier.com, www.sciencedirect.com , (2001).

[20] Kurubaş, Ö., "OLAP Küpleri ve Bir Etkileşimli Sorgulama Aracı", Yüksek Lisans Tezi, ***Kocaeli Üniversitesi Fen Bilimleri Enstitüsü***, İzmit, (2005).

[21] Dinçer, E.,"Veri Madenciliğinde K-Means Algoritması ve Tıp Alanında Uygulanması", Yüksek Lisans Tezi, ***Kocaeli Üniversitesi Fen Bilimleri Enstitüsü***, İzmit, (2006).

[22] Cattral, R., Oppacher, F., Deugo, D., , "Supervised and Unsupervised Data Mining with an Evolutionary Algorithm", ***IEEE***, (2001).

[23] www.robot.cmpe.boun.edu.tr/593/evrim.pdf (**Ziyaret tarihi**: **05.02.2007**)

[24] Kaya, M., Alhajj, R., , "Genetic Algorithm Based Framework for Mining Fuzzy Association Rules", ***Fuzzy Sets and Systems***, www.sciencedirect.com, www.elsevier.com, 587-601, (2005).

[25] Lin, W., Kuo, I., , "A Genetic Selection Algorithm for OLAP Data Cubes", ***Knowledge and Information Systems – Springer***, 6: 83 – 102, (2004).

[26] Toprak, Ş., Ganiz, M., Toprak, Ş., Arslan, A., ,"Genetik Algoritmalarla Makina Öğrenmesi İçin Tıbbi Verilerden Hipotez Uzayı Oluşturulması", ***Akademik Bilişim*** Adana, www.ab.org.tr, (2003).

[27] Berry, M. J. A., Linoff, G.S., "Data Mining Techniques For Marketing, Sales and Customer Relationship Management ", ***Wiley Publishing Inc.,***2004.

[28] www.osym.gov.tr (**Ziyaret tarihi**: **10.03.2007**)

[29] Akpınar, H.: "Veri Tabanlarında Bilgi Keşfi ve Veri madenciliği", ***İ.Ü. İşletme Fakültesi Dergisi*****,** Sayı :1, 1 – 22. (Nisan 2000)

[30] Duru, N., “An Application of Apriori Algorithm on a Diabetic Database”, ***Springer-Verlag*** Berlin Heidelberg, LNAI 3681, pp. 398.404, (2005).

[32] www.eecs.mit.edu.tr (**Ziyaret tarihi**: **21.04.2006**)

[33] http://en.wikipedia.org (**Ziyaret tarihi**: **14.06.2006**)

ÖZGEÇMİŞ

1980 yılında İstanbul, Üsküdar'da doğdu. İlk ve orta öğrenimini İzmir'in Menemen ilçesinde tamamladı. 1998 yılında Menemen Lisesi (Yabancı Dil Ağırlıklı)' nden mezun olarak lise öğrenimini tamamladı. 1999 yılında girdiği Kocaeli Üniversitesi Bilgisayar Mühendisliği Bölümü' nden 2003 yılında mezun oldu. 2004 yılında Kocaeli Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı'nda yüksek lisans öğrenimine başladı.

2005 yılında, Kocaeli Üniversitesi Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü'nde araştırma görevlisi olarak çalışmaya başlamıştır.