İstatistik Araştırma Dergisi
Journal of Statistical Research

ISSN: 2791-7614

(2026), 16 (1)
Başvuru / Received :26.02.2026
Kabul / Accepted :06.07.2026

https://dergipark.org.tr/tr/pub/jsstr

ARAŞTIRMA MAKALESİ **RESEARCH ARTICLE**

Veri Madenciliğinde Genetik Algoritma ile Kümeleme

Erkut TEKELİ
Adana Alparslan Türkeş Bilim ve Teknoloji Üniversitesi, Bilgisayar ve Bilişim Fakültesi / Dr. Öğr. Üyesi
etekeli@gmail.com
Orcid No: 0000-0001-9468-5378

Özlem AKAY
Gaziantep İslam Bilim ve Teknoloji Üniversitesi, Tıp Fakültesi / Doç. Dr.
ozlem.akay@gibtu.edu.tr
Orcid No: 0000-0002-9539-7252

Güzin YÜKSEL
Çukurova Üniversitesi, Fen Edebiyat Fakültesi, Bilgisayar Bilimleri Bölümü / Prof. Dr.
guzinigs@gmail.com
Orcid No: 0000-0002-1644-3696

Özet

Veri madenciliğinde yaygın şekilde kullanılan kümeleme analizi, *m* adet özelliğe sahip *n* adet birimin çeşitli algoritmalar aracılığıyla k adet alt kümeye ayrılması işlemidir. Farklı kümeleme algoritmaları aynı veri seti için birbirinden farklı sonuçlar üretebilmekte; bu durum, birimlerin doğru biçimde kümelendiği en uygun yöntemin belirlenmesini zorunlu kılmaktadır. Bu çalışmada, küme içi ve kümeler arası uzaklıklar ile silhouette genişliğini birleştiren yeni bir uygunluk fonksiyonu tanımlanmış; bu fonksiyon aracılığıyla genetik algoritma (GA) tabanlı kümeleme yaklaşımı geliştirilmiştir. Çalışma, karşılaştırmalı bir Monte Carlo simülasyonu ve gerçek veri uygulamasından oluşmaktadır. Simülasyon aşamasında R programlama dili kullanılarak birim sayısı, küme sayısı, küme çakışıklığı, gürültü ve sapan değer parametrelerinin farklı kombinasyonlarını içeren 144 koşul için 100'er tekrarlı yapay veri setleri üretilmiş; GA yöntemi k-ortalama ve Ward hiyerarşik kümeleme yöntemleriyle karşılaştırılmıştır. Gerçek veri uygulamasında Seeds ve Iris benchmark veri setleri kullanılmış, sonuçlar 10 katlı çapraz doğrulama ile doğrulanmıştır. Yöntemlerin performansı Fowlkes-Mallows (F), Mirkin (M) ve bilgi varyansı (VI) küme geçerlilik indeksleri ile değerlendirilmiştir. Elde edilen bulgular, GA yönteminin her iki veri türünde de k-ortalama ve Ward yöntemlerine kıyasla daha yüksek kümeleme doğruluğu sağladığını ortaya koymaktadır.

Anahtar Sözcükler: Geçerlilik İndeksleri, Genetik Algoritma, Kümeleme, Veri Madenciliği

Sorumlu Yazar / Corresponding Author: Özlem AKAY, Gaziantep İslam Bilim ve Teknoloji Üniversitesi, Tıp Fakültesi, Biyoistatistik ABD.

Atıf / Citation: TEKELİ E., AKAY Ö., YÜKSEL G. (2026). Veri Madenciliğinde Genetik Algoritma ile Kümeleme. *İstatistik Araştırma Dergisi*, 16 (1), 43-58.

Atıf / Citation: TEKELİ E., AKAY Ö., YÜKSEL G. (2026). Veri Madenciliğinde Genetik Algoritma ile Kümeleme. *İstatistik Araştırma Dergisi*, 16 (1), 43-58.

Clustering with Genetic Algorithm in Data Mining

Abstract

Cluster analysis, widely used in data mining, is the process of partitioning n units with *m* features into *k* sub-clusters using various algorithms. Different clustering algorithms may yield divergent results for the same dataset, making it essential to identify the most appropriate method for accurate cluster assignment. In this study, a new fitness function combining within-cluster distances, between-cluster distances, and silhouette width was defined, and a genetic algorithm (GA)-based clustering approach was developed accordingly. The study comprises a comparative Monte Carlo simulation and a real-data application. In the simulation phase, 144 experimental conditions were generated using the R programming language by varying sample size, number of clusters, cluster overlap, noise, and outlier parameters, with 100 replications each; GA was compared against k-means and Ward's hierarchical clustering. In the real-data application, the Seeds and Iris benchmark datasets were used, and results were validated via 10-fold cross-validation. Method performance was evaluated using the Fowlkes-Mallows (F), Mirkin (M), and Variance of Information (VI) cluster validity indices. The findings demonstrate that the GA-based approach achieves higher clustering accuracy than both k-means and Ward methods across all data types examined.

Keywords: Validity Indexes, Genetic Algorithm, Clustering, Data Mining

1. Giriş

Veri madenciliğinin üstlendiği görevler geniş bir bakış açısıyla sınıflama, tahmin, kümeleme ve özetleme olarak incelenebilir. Görevin ana hedefine bağlı olarak, veri madenciliği sürecinin çıktısı ön kestirimsel modeller veya betimleyici bilgiler olacaktır. Veri madenciliği süreci tarafından türetilen ön kestirimsel modeller sınıflama veya tahmin görevleri söz konusu olduğunda kullanılırken betimleyici bilgiler, kümeleme ve özetleme türünden problemler için uygun olacaktır.

Kümeleme analizi kavramı ilk kez Tryon tarafından 1939 tarihinde literatüre kazandırılmıştır (Lorr, 1993). Kümeleme analizi, nesneleri birbirlerine benzerliklerine göre sınıflandırmaya yarayan bir tekniktir. Bu teknikte nesnelere önceden belirlenen kriterlere göre kümelendiğinde aynı kümedeki nesnelerin özellikleri birbirine çok benzerken kümeler arası ölçülen özellikler ise birbirine benzememektedir.

Genetik algoritmanın optimizasyon problemini çözmek için güçlü bir araç olduğu iyi bilinmektedir. Farklı alanlardaki birçok uzman, genetik algoritmalara dayanan problem çözme metodolojilerini etkili problem çözme araçları olarak kabul etmiştir.

Günümüzde bazı araştırmacılar veri madenciliğinde kullanılan kümeleme analizi ve sınıflama için genetik algoritmayı kullanmaya başlamışlardır. Tiwari vd. (2010), alt uzaylardaki yüksek boyutlu nesneleri kümelemek için genetik bir k-ortalama algoritması sunmuşlardır. Mor vd. (2014) kümelenme için genetik algoritma kullanmış ve elde edilen sonuçları k-ortalama kümeleme yöntemiyle karşılaştırmışlardır. Cömert vd. (2016) çalışmalarında, atölye tipi yerleşime sahip bir imalat işletmesi ele alarak öncelikli olarak üretimi devam eden ürünlerin üretildiği makinalara k-ortalama algoritması ve genetik algoritma uygulayıp uygun kümeler oluşturmuşlardır. Çalışmanın sonucunda mevcut ve önerilen iki algoritmanın elde ettiği toplam taşıma maliyetleri karşılaştırılmış ve önerilen sistemin sağladığı üstünlükler bir örnek uygulama üzerinde gösterilmiştir. Haznedar vd. (2017), karaciğer mikrodizi kanser veri setinin sınıflandırılması için Uyarlamalı Ağ Tabanlı Bulanık Mantık Çıkarım Sistemi (ANFIS) ve Genetik Algoritmaya (GA) dayalı hibrit bir yaklaşım önermişlerdir. Simülasyon sonuçları, diğer bazı yöntemlere ait sonuçlarla karşılaştırılmıştır. Elde edilen sonuçlardan, önerilen yöntemin diğer yöntemlere göre daha başarılı olduğu görülmüştür. Ayık vd. (2007), Atatürk Üniversitesi öğrencilerinin mezun oldukları lise türleri ve lise mezuniyet dereceleri ile kazandıkları fakülteler arasındaki ilişki, veri madenciliği teknikleri kullanarak incelenmişlerdir. Adak vd. (2016), ikili gaz karışımlarının başarılı bir şekilde sınıflandırılması için yapay sinir ağının eğitim aşamasını genetik algoritma yardımıyla optimize etmişlerdir. Yedi farklı ağ tasarımında yapılan birçok test, yapay sinir ağının genetik algoritma ile eğitilmesinin geleneksel yöntem olan geri yaylım (BP) ile eğitilmesine göre çok daha başarılı sonuçlar verdiği görülmüştür. Göğebakan ve Servi (2020), çok değişkenli homojen ve heterojen büyük verilerin kümelemesi için yeni bir kümeleme algoritması geliştirmişlerdir. Genetik algoritmalar, heterojen verideki parçalanmalara karşılık gelen kümelenme merkezlerinin yerini ve yapısını belirlemede kullanılmıştır. Şengür vd. (2005), klasik k-ortalama kümeleme algoritmasının belirtilen

yetersizliklerini, genetik tabanlı bir kümeleme algoritması ile gidermişlerdir. Genetik algoritmaların arama yetenekleri *k* küme merkezlerinin bulunması için kullanılmıştır. Sriwindono vd. (2022), genetik algoritmaların performansını artırmak amacıyla arama uzayını daraltmaya yönelik olarak kümeleme temelli bir ön işleme yaklaşımı önermiştir. Araştırmada, öğretmen yerleştirme probleminin çözümünde genetik algoritma uygulanmadan önce, orijinal veri seti K-ortalama kümeleme yöntemi kullanılarak alt gruplara ayrılmıştır. Genetik algoritma aşamasında ise çözüm kalitesini artırmak amacıyla Sıralı Çaprazlama ve Kısmi Karıştırma Mutasyonu operatörleri kullanılmıştır. Bu yaklaşımın, arama uzayının etkin biçimde sınırlandırılması yoluyla algoritmanın hesaplama performansını ve çözüm etkinliğini iyileştirdiği görülmüştür. Khan vd. (2022), yoğunluk tabanlı veri kümeleme problemleri için daha önce önerilmiş olan ağaç tabanlı hibrit genetik algoritma üzerinde iç geçerlilik indeksi kullanarak gerçek bilgiler kullanmadan minimum komşuluk mesafesini tahmin etmişlerdir. YiFei vd. (2023), yapısal hasar tespiti amacıyla yeni bir hibrit optimizasyon yöntemi önermiştir. Çalışmada, K-ortalama kümeleme optimizasyon algoritması ile genetik algoritmanın birleşimine dayanan bir yaklaşımı geliştirilmiştir. Wang vd. (2023), zaman pencereleri ve üç boyutlu yükleme kısıtları içeren Çoklu Depo Araç Rotalama Problemi'ni ele alarak, toplam işletme maliyetlerini en aza indirirken araç yükleme oranını maksimize eden iki amaçlı bir karma tamsayılı programlama modeli önermiştir. Çalışmada, müşteri kümelenmesine dayalı kaynak paylaşım şeması geliştirilmiş ve hesaplama karmaşıklığını azaltmak amacıyla üç boyutlu k-harmonik ortalama kümeleme algoritması kullanılmıştır. Modelin çözümünde ise kümeleme yaklaşımı, Genişletilmiş Baskın Olmayan Sıralama Genetik Algoritması-II ile bütünleştirilerek iki aşamalı hibrit bir kümeleme ve genetik algoritma yapısı oluşturulmuştur. Akopov vd. (2025), kentsel trafik yönetiminde artan kapasite ve esneklik gereksinimlerine yanıt olarak yeniden yapılandırılabilir çok katmanlı yol ağlarının) en uygun topolojik tasarımını ele almıştır. Çalışmada, trafik akışını maksimize ederken ağ karmaşıklığını ve maliyetleri minimize etmeyi amaçlayan çok amaçlı bir optimizasyon çerçevesi önerilmiştir. Bu amaç doğrultusunda, genetik algoritma, biyo-amaçlı ayrık parçacık sürü optimizasyonu ve bulanık kümeleme tekniklerini bütünleştiren Çoklu Ajan Hibrit Kümeleme Destekli Genetik Algoritma geliştirilmiştir. Gaon vd. (2025), sürdürülebilir ulaşım kapsamında elektrikli araçlar (EV) için rota planlama problemini ele alarak K-ortalama kümeleme ve genetik algoritma (GA) temelli hibrit bir optimizasyon çerçevesi önermiştir. Çalışmada, şarj istasyonları coğrafi yakınlığa göre K-ortalama kümeleme yöntemiyle gruplandırılarak gereksiz sapmalar azaltılmış ve optimum durak seçimi desteklenmiştir. Kümeleme aşamasının ardından genetik algoritma, seyahat mesafesi, enerji tüketimi, zaman ve şarj istasyonu kuyruk uzunlukları gibi dinamik parametreleri dikkate alarak en az şarj durağı içeren ve verimliliği koruyan rotaları belirlemek üzere çözümü iyileştirmiştir.

Bu çalışmada genetik algoritma kullanılarak birimlerin doğru kümelerde yer alması amaçlanmıştır. Bu amaç doğrultusunda genetik algoritma için yeni bir uygunluk fonksiyonu tanımlanarak simülasyon verisi ve gerçek veri üzerinde uygulanmıştır. Her iki veri türü için bazı küme geçerlilik indeksleri hesaplanmıştır. Çalışma altı bölümden oluşmaktadır. İlk bölümde genel bir giriş ve konu ile ilgili yapılan çalışmalardan bahsedilmiştir. İkinci bölümde materyal ve yöntemler detaylı bir şekilde açıklanmıştır. Üçüncü bölümde kümeleme analizi ve genetik algoritmaya değinilmiş ve genetik algoritmanın aşamaları ayrıntılı olarak anlatılmıştır. Ayrıca bazı küme geçerlilik indekslerinden söz edilmiştir. Dördüncü bölümde simülasyon, beşinci bölümde ise gerçek veri setleri üzerinde genetik algoritma, k-ortalama ve Ward yöntemi ile kümeleme analizi yapılmış ve yöntemlere ait geçerlilik indeksleri hesaplanmıştır. Son bölümde elde edilen sonuçlar değerlendirilmiştir.

Bu çalışmanın temel özgün katkısı, literatürde genellikle tek bir kümeleme ölçütünü esas alan uygunluk fonksiyonlarından farklı olarak, küme içi uzaklık, kümeler arası uzaklık ve silhouette genişliğini birlikte dikkate alan yeni bir uygunluk fonksiyonunun önerilmesidir. Böylece genetik algoritmanın arama sürecinde kümelerin hem kendi içindeki benzerliğinin artırılması hem de kümeler arasındaki ayrımın güçlendirilmesi amaçlanmıştır.

2. Materyal ve Yöntem

Bu çalışma, veri madenciliğinde kümeleme problemi için önerilen genetik algoritma (GA) tabanlı yaklaşımın etkinliğini değerlendirmeye yönelik karşılaştırmalı bir simülasyon çalışması ve gerçek veri uygulamasından oluşmaktadır. Çalışmada GA yöntemi, kümelemede yaygın olarak kullanılan k-ortalama ve Ward hiyerarşik kümeleme yöntemleriyle karşılaştırılmıştır. Yöntemler arası karşılaştırma, kontrollü simülasyon koşullarında ve kamuya açık gerçek veri setleri üzerinde, önceden tanımlanmış küme geçerlilik indeksleri aracılığıyla nesnel biçimde gerçekleştirilmiştir.

Bu çalışmada simülasyon raporlaması, Morris vd. (2019) tarafından önerilen STRESS (Structured Reporting of Simulation Studies) kılavuzu esas alınarak yapılandırılmıştır. Söz konusu kılavuz; hedeflerin, veri üretim mekanizmasının, analiz yöntemlerinin, performans ölçütlerinin ve sonuçların sistematik biçimde raporlanmasını önermektedir.

2.1. Veri Setleri

Çalışmada iki tür veri kullanılmıştır: Monte Carlo simülasyonu ile üretilen yapay veri setleri ve kamuya açık gerçek veri setleri.

Simülasyon verisi, R programlama dilinde (sürüm 4.3.2; R Core Team, 2023) mixsim paketi (sürüm 1.1.7; Melnykov vd., 2012) kullanılarak çok değişkenli normal dağılımdan üretilmiştir. Yapay veri setleri; birim sayısı (n = 50, 100), küme sayısı (k = 2, 3, 4, 5), küme çakışıklık düzeyi (BarOmega: ω = 0,001 ve ω = 0,05), gürültü değişken sayısı (noise = 0, 1, 2) ve sapan değer sayısı (outlier = 0, 1, 2) parametrelerinin farklı kombinasyonlarına göre oluşturulmuş; toplam 144 farklı koşul için her biri 100 tekrarlı olmak üzere simülasyon yürütülmüştür. Yeniden üretilebilirlik için sabit rastgele tohum değeri kullanılmıştır (set.seed(1234)). Simülasyon çalışmasında örneklem büyüklükleri n = 50 ve n = 100 olarak belirlenmiştir. Bu değerler, kümeleme simülasyon literatüründe küçük ve orta ölçekli veri setlerini temsil eden yaygın benchmark değerleri olup benzer karşılaştırmalı çalışmalarda da kullanılmıştır (Melnykov vd., 2012; Mor vd., 2014). Söz konusu örneklem büyüklükleri, GA'nın hesaplama yükü ile istatistiksel güvenilirlik arasındaki dengeyi sağlamak amacıyla tercih edilmiştir. Simülasyon veri setleri, çok değişkenli normal dağılımdan olasılığa dayalı rastgele örnekleme yoluyla üretilmiştir. Bu sayede her koşul için birbirinden bağımsız ve temsil edici örnekler elde edilmiştir.

Gerçek veri uygulamasında ise makine öğrenmesi ve kümeleme literatüründe sıkça kullanılan iki benchmark veri seti tercih edilmiştir: Seeds veri seti (n = 210, p = 7, k = 3) ve Iris veri seti (n = 150, p = 4, k = 3). Her iki veri seti de UCI Makine Öğrenmesi Deposu'ndan (Dua ve Graff, 2019) elde edilmiştir. Veri setlerinin seçiminde, gerçek küme yapısının bilinmesi (etiketli veri), boyutlarının farklılaşması ve kümeleme araştırmalarındaki yaygın kullanımları belirleyici ölçütler olarak alınmıştır. Gerçek veri setlerinin seçiminde olasılığa dayanmayan amaçsal örnekleme (purposive sampling) yaklaşımı benimsenmiştir. Seeds ve Iris veri setleri; gerçek küme yapısının bilinmesi, farklı boyut ve karmaşıklık düzeylerine sahip olmaları ve kümeleme araştırmalarındaki yaygın kullanımları nedeniyle tercih edilmiştir.

2.2. Ön İşleme

Simülasyon verileri kontrollü koşullarda üretildiğinden herhangi bir ön işleme adımı uygulanmamıştır. Gerçek veri setleri için ise Öklid uzaklığına dayanan GA ve k-ortalama algoritmalarının ölçek farklılıklarından etkilenmemesi amacıyla tüm değişkenlere z-puanı standardizasyonu uygulanmıştır:

$$z = (x - \bar{x}) / s$$

Burada $\bar{x}$ örneklem ortalamasını, s ise örneklem standart sapmasını göstermektedir. Her iki veri setinde aykırı değer tespiti için kutu grafiği (boxplot) incelemesi yapılmış; Seeds ve Iris veri setlerinde analizi etkileyecek düzeyde uç değer gözlemlenmemiştir.

2.3. Karşılaştırılan Yöntemler

Bu çalışmada üç kümeleme yöntemi karşılaştırılmıştır. Yöntemler, birbirinden farklı algoritmik paradigmaları temsil etmeleri ve kümeleme literatüründeki yaygın kullanımları nedeniyle seçilmiştir:

- Genetik Algoritma (GA): Evrimsel hesaplamaya dayanan, sezgisel bir arama ve optimizasyon yöntemidir. Bu çalışmada önerilen yeni uygunluk fonksiyonu uygulanmıştır (bkz. Bölüm 3).
- K-Ortalama: Bölümleyici kümeleme yöntemlerinin en yaygın temsilcisidir; küme merkezlerini yinelemeli biçimde güncelleyerek küme içi uzaklığı minimize eder.
- Ward Yöntemi: Hiyerarşik kümeleme yöntemleri arasında en sık kullanılan yaklaşımlardan biridir; her birleşme adımında küme içi varyans artışını minimize eder.

Bu çalışma değişkenler arası ilişkileri modellemeyi değil, gözlem birimlerini benzerliklerine göre gruplara ayırmayı amaçlayan denetimsiz (unsupervised) bir öğrenme çalışmasıdır. Bu nedenle bağımlı-bağımsız değişken ayrımı yapılmamış; yöntemler yalnızca kümeleme performansları açısından karşılaştırılmıştır.

2.4. GA Hiperparametreleri

Genetik algoritma uygulamasında kullanılan hiperparametreler Tablo 1'de verilmiştir. Parametre değerleri, GA literatüründeki yaygın öneriler doğrultusunda belirlenmiş (Holland, 1975; Goldberg, 1989) ve tüm deney koşullarında sabit tutulmuştur.

Tablo 1. GA hiperparametre ayarları

Hiperparametre	Değer	Kaynak
Popülasyon büyüklüğü	50	Goldberg (1989)
Nesil sayısı (maksimum iterasyon)	100	Holland (1975)
Çaprazlama oranı	0,80	Goldberg (1989)
Mutasyon oranı	0,01	Holland (1975)
Seçim yöntemi	Turnuva seçimi (boyut = 3)	Goldberg (1989)
Çaprazlama operatörü	Tek noktalı çaprazlama	Holland (1975)

2.5. Birincil ve İkincil Çıktı Değişkenleri

Yöntemlerin kümeleme performansını değerlendirmek amacıyla üç dış geçerlilik indeksi kullanılmıştır. Birincil performans ölçütü olarak Fowlkes-Mallows (F) indeksi benimsenmiştir; çünkü bu indeks, elde edilen küme yapısı ile gerçek küme etiketleri arasındaki uyumu doğrudan ölçmekte ve değeri 1'e ne kadar yakınsa kümeleme o kadar başarılı kabul edilmektedir. İkincil performans ölçütleri olarak ise Mirkin (M) indeksi ve Bilgi Varyansı (VI) indeksi hesaplanmıştır; her iki indeks için de 0'a yakın değer daha iyi kümeleme performansını ifade etmektedir. Bu indekslerin formülleri Bölüm 3.1'de ayrıntılı olarak sunulmaktadır.

2.6. İstatistiksel Analiz

Bu çalışmada yöntemler arası karşılaştırma, istatistiksel hipotez testi yerine küme geçerlilik indekslerinin (F, M ve VI) simülasyon koşulları genelinde sistematik olarak incelenmesi yoluyla gerçekleştirilmiştir. Kümeleme simülasyon literatüründe bu yaklaşım yaygın biçimde benimsenmektedir (Melnykov vd., 2012); zira birincil amaç anlamlılık sınaması değil, farklı koşullar altında algoritmaların göreli performansını ortaya koymaktır. Her simülasyon koşulu için 100 tekrar yapılmış ve indeks değerleri ortalama olarak raporlanmıştır.

Simülasyon çalışmasında her koşul için 100 tekrar üzerinden hesaplanan F, M ve VI indeks değerleri; aritmetik ortalama ve standart sapma (Ort. ± SS) ile özetlenmiştir. Gerçek veri setleri için ise 10 katlı çapraz doğrulama aşamalarına ait indeks değerlerinin ortalaması raporlanmıştır.

Tüm analizler R sürüm 4.3.2 (R Core Team, 2023) ile gerçekleştirilmiş; mixsim (sürüm 1.1.7), cluster (sürüm 2.1.4) paketleri ve yazarlar tarafından geliştirilen özel GA kodu kullanılmıştır.

3. Kümeleme için Genetik Algoritma

Genetik algoritma Holland (1975) tarafından geliştirilmiştir. Genetik algoritma, gerçek hayattaki doğal seçim mekanizması ve gen yapılarını örnek alan çaprazlama ve mutasyon içeren sezgisel bir arama algoritmasıdır. Bu algoritma kalıtım, mutasyon, seçim ve çaprazlama gibi evrimsel biyolojideki tekniklerden esinlenerek geliştirilmiştir. Genetik algoritmayı anlayabilmek için öncelikle bazı kavramlar açıklanmalıdır. Genetik algoritmada çözüme ait genetik bilgi taşıyan her bir elemana gen adı verilmektedir. Bir veya daha fazla gen bir araya gelerek problemin çözümüne ait tüm bilgiyi içeren dizileri oluşturur. Bu diziler kromozom olarak adlandırılmaktadır. Her bir kromozom bir çözümü temsil etmektedir. Kromozomlar bir araya gelerek ise popülasyonları oluşturmaktadır.

Genetik algoritmalar problemlere tek bir çözüm üretmek yerine farklı çözümlerden oluşan bir çözüm kümesi üretir. Böylelikle, arama uzayında aynı anda birçok nokta değerlendirilmekte ve sonuçta bütünsel çözüme ulaşma olasılığı yükselmektedir. Veri kümelerinin genetik algoritma yardımıyla çok amaçlı bir biçimde kümelenmesi geleneksel genetik algoritmanın çalışma yapısına oldukça benzemektedir. Kümeleme işleminde temel prensip, sınıf içi benzerliği maksimum, sınıflar arası benzerliği minimum yapmaktır. Bir kümeleme yönteminin kalitesi bu prensibi sağlamasıyla doğru orantılıdır. Bu nedenle genetik algoritma kümeleme analizi için uygun bir yöntemdir. Çalışmada kullanılan başlangıç popülasyonu ve uygunluk fonksiyonu aşağıdaki gibi tanımlanmıştır.

Popülasyonda her kromozom *k*p* uzunluğundaki gerçek vektör değeridir, burada *k* oluşturulacak küme sayısı iken *p* veri setindeki değişken sayısıdır. n bireyden oluşan bir popülasyonda bireyleri temsil etmek için **k** tanesi rastgele olarak seçilir. Her k_i ($i=1, 2, \ldots, p$), $kromozom_j$ ($j=1, 2, \ldots, n$)'nin merkezlerinden birini gösterir. Başlangıç popülasyonu çözüm uzayından seçilen rastgele kromozomlardan oluşur, yani n*k küme merkezi başlangıç popülasyonu için rastgele seçilir.

Genetik algoritmalar bir sonucun başarımını ölçerken, kullanıcı tarafından tanımlanmış olan fonksiyonlar doğrultusunda bu işlemi yapmaktadırlar. Bu fonksiyonlara uygunluk fonksiyonu (fitness function) ismi verilmektedir. Genetik algoritma bulduğu sonucun uygunluk fonksiyonundan aldığı değer doğrultusunda bir sonraki nesilde o sonucun kalmasının ya da silinmesinin gerekliliğine karar vermektedir. Bu nedenle uygunluk fonksiyonunun seçilmesi işlemi genetik algoritmanın iyi sonuçlar vermesi açısından en önemli noktadır.

Çalışmada uygunluk fonksiyonunu hesaplamak için küme içi uzaklık, kümeler arası uzaklık ve silhouette genişliği değerleri kullanılmıştır. Küme içi uzaklık, küme içindeki elemanlar arasındaki uzaklıktır; küme arası uzaklık ise iki küme elemanları arasındaki uzaklıktır. *q'ıncı* ve *r'inci* (*q, r=1, 2, …, k*) kümeler arasındaki uzaklık (1)'de gösterildiği gibi hesaplanır.

$$D_{inter}^{q,r}(X_i, X_j) = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}\left(X_i - X_j\right)^2/(m*n)} \quad (1)$$

Burada *m* ve *n* sırasıyla *q'ıncı* ve *r'inci* kümelerdeki eleman sayılarıdır. Eğer $r = q$ ise küme arası uzaklık yoktur ve r, q ve q, r aynıdır.

q'ıncı (*q = 1, 2, …, k) kümenin* küme içi uzaklığı (2)'de gösterildiği gibi hesaplanır.

$$D_{\text{İ}ntra}^{q}(X_i, X_j) = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m}\left(X_i - X_j\right)^2/(m*m)} \quad (2)$$

Burada *m*, *q*'ıncı kümenin eleman sayısıdır.

Küme içi ve kümeler arası uzaklık toplamları sırasıyla (3) ve (4)'te gösterildiği gibi hesaplanır.

$$Toplam(D_{inter}) = \sum_{q=1}^{k-1}\sum_{r=q+1}^{k} D_{inter}^{q,r} \quad (3)$$

$$Toplam(D_{intra}) = \sum_{q=1}^{k} D_{intra}^{q} \quad (4)$$

Silhouette genişliği ise her bir gözlemin silhouette değerinin ortalamasıdır. Silhouette değeri, belirli bir gözlemin atandığı kümedeki yoğunlaşma derecesini ölçer. Silhouette değeri iyi kümelenmiş gözlemlerde 1'e yakın değer alırken, kötü kümelenen gözlemlerde ise -1'e yakın değer almaktadır. *i*'inci gözlem için silhouette değeri (5)'te gösterildiği gibi tanımlanır.

$$S(i) = \frac{b_i - a_i}{\max(b_i, a_i)} \quad (5)$$

Burada a_i, i ve i ile aynı kümedeki diğer gözlemler arasındaki ortalama uzaklıktır, b_i, i ile en yakın komşu kümedeki elemanlar arasındaki ortalama uzaklıktır. a_i ve b_i , (6)'da gösterildiği gibi hesaplanır.

$$a_i = \frac{1}{n(C(i))}\sum_{j\epsilon C(i)} dist(i,j), \quad b_i = \min_{C_k \epsilon \mathcal{C} \setminus C(i)} \sum_{j\epsilon C_k} \frac{dist(i,j)}{n(C_k)} \quad (6)$$

Burada *C(i), i.* gözlemi içeren kümedir. *dist(i, j), i ve j'inci* gözlemler arasındaki uzaklıktır (Öklid, Manhattan gibi). *n(C),* C kümesinin eleman sayısıdır. Silhouette genişliği [-1; 1] arasında değer alır ve maksimum olmalıdır. Ortalama silhouette değeri (7)'de gösterildiği gibi hesaplanır.

$$Ortalama\big(S(i)\big) = \frac{\sum_{i=1}^{n} S(i)}{n} \quad (7)$$

Burada *n*, örneklem boyutudur.

Kümeler arası toplam uzaklık ve toplam silhouette değeri maksimum değer almalıdır ve küme içi uzaklık ise minimum değer almalıdır. Bu nedenle GA yönteminde uygunluk fonksiyonu maksimum değer almalıdır. Uygunluk fonksiyonu (8)'de gösterildiği gibi tanımlanır.

$$FF = \frac{Toplam(D_{inter})}{Toplam(D_{intra})} + Ortalama(S(i)) \quad (8)$$

Literatürde genetik algoritma tabanlı kümeleme çalışmalarında uygunluk fonksiyonları çoğunlukla yalnızca küme içi uzaklıkların en küçüklenmesi veya kümeler arası uzaklıkların en büyüklenmesi gibi tek bir ölçüte dayanmaktadır. Bu çalışmada önerilen uygunluk fonksiyonu ise küme içi uzaklık, kümeler arası uzaklık ve silhouette genişliğini birlikte değerlendirerek küme bütünlüğü ile kümeler arasındaki ayrımı eş zamanlı olarak optimize etmeyi amaçlamaktadır. Bu yönüyle önerilen yaklaşım, kümeleme kalitesini daha kapsamlı biçimde değerlendirmeyi amaçlayan bütünleşik bir uygunluk fonksiyonu sunmaktadır.

3.1. Küme Geçerlilik İndeksleri

Kümeleme performansını değerlendirebilmek için bazı geçerlilik indeksleri kullanılır. Çalışmada Tablo 2'de verilen Mirkin (M), Fawkles ve Mallows (F) ve Variance of Information (VI) geçerlilik indeksleri ele alınmıştır.

Tablo 2. Küme geçerlilik indeksleri, eşitlikleri ve kaynakları

	Eşitlikler	**Kaynaklar**
Mirkin (M)	$M(c_1, c_2) = n(n-1)\big(1 - R(c_1, c_2)\big)$	Mirkin, 1996
Fawkles ve Mallows (F)	$F(c_1, c_2) = \sqrt{W_1(c_1, c_2) W_2(c_1, c_2)}$ ve $W_i(c_1, c_2) = \frac{2N_{11}}{\sum_{k=1}^{K^{(i)}} n_k^{(i)} \left(n_k^{(i)} - 1\right)}$	Fawkles ve Mallows, 1983
Variance of Information (VI)	$VI(c_1, c_2) = H(c_1) + H(c_2) - 2I(c_1, c_2)$ ve $H(c_i) = -\sum_{k=1}^{K^{(i)}} \frac{n_k^{(i)}}{n} log \frac{n_k^{(i)}}{n}$ $I(c_1, c_2) = \sum_{k=1}^{K^{(1)}} \sum_{k=1}^{K^{(2)}} \frac{n_{kr}}{n} log \frac{n_{kr} n}{n_k^{(1)} n_r^{(2)}}$	Meila, 2006

Not: F: Fowlkes-Mallows indeksi (birincil ölçüt; iyi kümeleme için 1'e yakın olmalı). M: Mirkin indeksi (ikincil ölçüt; iyi kümeleme için 0'a yakın olmalı). VI: Bilgi Varyansı indeksi (ikincil ölçüt; iyi kümeleme için 0'a yakın olmalı). c_1 ve c_2: karşılaştırılan iki bölüm vektörü; N_{11}: her iki bölümde aynı kümede yer alan çift sayısı.

Tablo 2'de bazı küme geçerlilik indekslerinin eşitlikleri verilmiştir. Buna göre c_1 ve c_2 sırasıyla birinci ve ikinci bölüm vektörleridir. N_{11}, c_1 ve c_2, altında aynı kümede bulunan nokta çiftlerinin sayısıdır. N_{00}, c_1 ve c_2 altında farklı kümelerde bulunan nokta çiftlerinin sayısıdır. N, bölüm vektöründeki nokta sayısıdır. $n_k^{(i)}$ c_i bölümüne göre k kümesinin boyutudur. n_{kr} sırasıyla c_1 ve c_2 bölümleri altında eşzamanlı olarak *k'ıncı* ve *r'inci* kümelere atanan gözlemlerin sayısıdır. Bu çalışmada birincil performans ölçütü olarak Fowlkes-Mallows (F) indeksi belirlenmiştir; zira bu indeks, elde edilen küme yapısının gerçek küme etiketleriyle uyumunu doğrudan ölçmekte ve kümeleme literatüründe yaygın biçimde kullanılmaktadır. Mirkin (M) ve Bilgi Varyansı (VI) indeksleri ise ikincil performans ölçütleri olarak ele alınmıştır. İyi bir kümeleme performansı için F değeri 1'e yakın olmalıdır, M ve VI ise 0'a yakın olmalıdır.

4. Monte Carlo Simülasyonu

Bu çalışmada birimlerin daha doğru kümelere atanması amacıyla genetik algoritma ile kümeleme analizi yapılmıştır. Genetik algoritmanın etkinliğinin araştırılması için bir simülasyon çalışması yapılmış ve elde edilen sonuçlar, kümelemede en çok kullanılan k-ortalama ve Ward yöntemleri ile karşılaştırılmıştır.

Simülasyon çalışması için R programında "mixsim" paketi (Melnykov vd., 2012) kullanılarak birim sayısı (n = 50, 100), küme sayısı (k = 2, 3, 4, 5), BarOmega ($\bar{\omega}$ = 0.001, 0.05) ve gürültü (noise = 0, 1, 2) ve sapan değer (outlier = 0, 1, 2) durumlarına göre yapay veri setleri üretilmiştir. Her durum için 100 farklı deneme yapılmış ve ortalamaları alınmıştır. Böylece toplam 144 farklı durum için simülasyon çalışması yapılmıştır. BarOmega parametresi kümelerin çakışıklığını gösterir ve simülasyon veri seti üretirken $\bar{\omega}$ = 0.001 olması, kümelerin birbirinden iyi bir şekilde ayrılmış veri setlerinin $\bar{\omega}$ = 0.05 olması ise kümelerin kısmen çakışık olduğunu veri setlerinin üretilmesini sağlar. Ayrıca gürültü parametresi kümelemeyi etkilemeyen değişken sayısını, sapan değer parametresi ise veri setinde herhangi bir kümeye girmeyen gözlem sayısını verir.

Elde edilen veri setlerine genetik algoritma, k-ortalama ve Ward yöntemleri kullanılarak kümeleme analizi yapılmıştır. Elde edilen kümeleme sonuçlarına göre, üç yöntem için F, M ve VI geçerlilik indeks değerleri hesaplanmıştır. İyi bir kümeleme için F indeksi değerleri 1'e, M ve VI indeks değerleri de 0'a yakın olmalıdır. F, M ve VI geçerlilik indeks değerleri EK-1'de verilmiştir. EK-1'e bakıldığında, F indeksi için 144 durum içinde 4 durumda k-ortalama yöntemi, 6 durumda Ward yöntemi en iyi performansı göstermiştir. 1 durumda ise k-ortalama ve Ward yöntemleri beraber en iyi performansı göstermiştir. Geri kalan 133 durumda ise GA yöntemi ile en iyi performans elde edilmiştir. M geçerlilik indeksine göre ise sadece 2 durumda Ward yöntemi, geri kalan 142 durumda ise GA yöntemi başarılı olmuştur. VI geçerlilik indeksi kullanıldığında ise 36 durumda Ward, 3 durumda k-ortalama, kalan 105 durumda ise GA yöntemi en iyi performansı göstermiştir.

Yine EK-1'e göre F geçerlilik indeksini baz alarak, kümelerde çakışma olduğu durumda, yani $\bar{\omega} = 0.05$ olan 72 durumun 64'ünde GA yöntemi başarılı olmuştur; kümelerin ayrık olduğu durumda, yani $\bar{\omega} = 0.001$ olan 72 durumun 69'unda GA yöntemi başarılı olmuştur. Küme sayısına göre EK-1'i gözden geçirdiğimizde ise 2 kümenin simüle edildiği durumların hepsinde GA yöntemi başarılı olurken 3, 4 ve 5 kümenin simüle edildiği durumlarda GA dışındaki yöntemler sırasıyla 3, 2 ve 6 defa daha başarılı olmuşlardır. Gözlem birimi sayısına göre ise GA dışındaki yöntemler 50 birimlik vakalarda 7 defa, 100 birimlik vakalarda ise 4 defa en iyi performansı göstermişlerdir. Veride gürültü olup olmadığına bakıldığında, gürültü olmayan vakaların sadece 1'inde GA dışındaki yöntemler başarılı olurken, 1 değişkenin kümelemeyi etkilemediği durumların içinde 2 vaka, 2 değişkenin kümelemeyi etkilemediği durumlarda ise 8 vaka GA dışındaki k-ortalama ve Ward yöntemleri başarılı olmuştur. Sapan değere göre incelediğimizde, sapan değerin 0, 1 ve 2 olduğu durumlarda sırasıyla sadece 4, 4 ve 3 durumda GA dışındaki yöntemlerin başarılı olduğu gözlemlenmiştir.

5. Gerçek Veri Uygulaması

Simülasyon çalışmasının yanı sıra uygulamalarda en çok kullanılan Seeds ve Iris gerçek veri setleri de GA, k-ortalama ve Ward yöntemleri ile kümelenmiştir. Kümeleme işleminin veri setindeki tüm veriler kullanılarak yapılması durumunda aşırı uyumluluk problemi oluşabilir; bu da modelin daha önce görmediği farklı verilerle karşılaştığında kestirim gücünün azalmasına yol açar (Hawkins, 2004). Aşırı uyumluluk problemini giderebilmek için 10 katlı çapraz doğrulama uygulanmıştır. Veri seti 10 parçaya bölünüp her aşamasında farklı bir parça test verisi olarak seçilmiş, geri kalan 9 parça ise eğitim seti olarak kullanılmıştır. Eğitim seti ile küme merkezleri tespit edildikten sonra bulunan küme merkezlerinin kestirim performansı test verisi kullanılarak ölçülmüştür.

Her iki veri seti için de F, M ve VI küme geçerlilik indeks değerleri hesaplanarak elde edilen sonuçlar Tablo 3-6'da verilmiştir.

Tablo 3. Seeds veri seti için GA yönteminin performansı (Değerler; eğitim ve test aşamalarına ait 10 katlı çapraz doğrulama sonuçlarını göstermektedir. ORT.: 10 katın ortalaması)

		Çapraz doğrulama aşamaları										**ORT.**
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	
F	**Eğitim**	0,849	0,899	0,868	0,823	0,867	0,889	0,850	0,860	0,859	0,840	0,860
	Test	1,000	0,631	0,898	0,887	0,724	0,725	0,891	0,707	0,762	1,000	0,822
M	**Eğitim**	3552	2360	3100	4172	3148	2612	3514	3282	3312	3784	3283,6
	Test	0	104	26	30	70	76	30	78	70	0	48,4
VI	**Eğitim**	0,584	0,420	0,526	0,662	0,490	0,453	0,563	0,558	0,539	0,592	0,539
	Test	0,000	0,871	0,280	0,293	0,811	0,706	0,292	0,721	0,674	0,000	0,465

Not: F: Fowlkes-Mallows indeksi (1'e yakın = daha iyi). M: Mirkin indeksi (0'a yakın = daha iyi). VI: Bilgi Varyansı indeksi (0'a yakın = daha iyi). ORT.: 10 katın aritmetik ortalaması. Eğitim: modelin eğitildiği 9 kat; Test: doğrulama katı.

Tablo 4. Seeds veri seti için k-ortalama yönteminin performansı (Değerler; eğitim ve test aşamalarına ait 10 katlı çapraz doğrulama sonuçlarını göstermektedir. ORT.: 10 katın ortalaması)

		Çapraz doğrulama aşamaları										**ORT.**
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	
F	**Eğitim**	0,641	0,719	0,713	0,648	0,576	0,717	0,698	0,712	0,711	0,579	0,671
	Test	0,596	0,629	0,580	0,566	0,714	0,761	0,673	0,748	0,702	0,656	0,662
M	**Eğitim**	10876	9134	9466	10876	12634	9302	10320	8858	9736	12634	10383,6
	Test	160	148	138	136	128	88	124	94	144	124	128,4
VI	**Eğitim**	1,049	0,780	0,789	1,014	1,287	0,773	0,828	0,792	0,793	1,289	0,939
	Test	1,037	0,996	1,098	1,037	0,721	0,626	0,742	0,645	0,560	0,826	0,829

Not: F: Fowlkes-Mallows indeksi (1'e yakın = daha iyi). M: Mirkin indeksi (0'a yakın = daha iyi). VI: Bilgi Varyansı indeksi (0'a yakın = daha iyi). ORT.: 10 katın aritmetik ortalaması. Eğitim: modelin eğitildiği 9 kat; Test: doğrulama katı.

Tablo 3 ve 4'te görüldüğü gibi, genetik algoritma ve k-ortalama yöntemi ile Seeds verisine kümeleme analizi yapılmıştır. Elde edilen sonuçların geçerliliği için çapraz doğrulama yapılarak küme geçerlilik indeksleri hesaplanmıştır. İndeks değerlerinin hem "Eğitim" hem de "Test" aşamasında GA'nın k-ortalama yönteminden daha iyi sonuçlar verdiği görülmüştür.

Tablo 5. Iris veri seti için GA yönteminin performansı (Değerler; eğitim ve test aşamalarına ait 10 katlı çapraz doğrulama sonuçlarını göstermektedir. ORT.: 10 katın ortalaması)

		Çapraz doğrulama aşamaları										**ORT.**
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	
F	**Eğitim**	0,722	0,755	0,700	0,716	0,719	0,695	0,707	0,731	0,726	0,729	0,720
	Test	0,730	0,529	0,719	0,625	0,751	1,000	0,623	0,658	0,626	0,660	0,692
M	**Eğitim**	3308	2990	3596	3406	3358	3650	3528	3200	3262	3268	3356,6
	Test	40	64	36	48	32	0	46	48	48	54	41,6
VI	**Eğitim**	0,833	0,714	0,834	0,840	0,836	0,890	0,817	0,778	0,829	0,812	0,818
	Test	0,558	1,007	0,734	1,028	0,504	0,000	0,981	0,803	0,837	0,831	0,728

Not: F: Fowlkes-Mallows indeksi (1'e yakın = daha iyi). M: Mirkin indeksi (0'a yakın = daha iyi). VI: Bilgi Varyansı indeksi (0'a yakın = daha iyi). ORT.: 10 katın aritmetik ortalaması. Eğitim: modelin eğitildiği 9 kat; Test: doğrulama katı.

Tablo 6. Iris veri seti için k-ortalama yönteminin performansı (Değerler; eğitim ve test aşamalarına ait 10 katlı çapraz doğrulama sonuçlarını göstermektedir. ORT.: 10 katın ortalaması)

		Çapraz doğrulama aşamaları										ORT.
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	
F	**Eğitim**	0,454	0,632	0,699	0,712	0,587	0,599	0,487	0,673	0,523	0,560	0,593
	Test	0,325	0,552	0,455	0,704	0,698	0,887	0,423	0,842	0,446	0,771	0,610
M	**Eğitim**	7094	4674	3592	4682	5070	5020	7854	4336	7350	5774	5544,6
	Test	92	72	72	52	46	16	96	24	126	40	63,6
VI	**Eğitim**	1,651	1,224	0,892	0,757	1,247	1,286	1,572	0,990	1,459	1,217	1,229
	Test	1,652	1,001	1,109	0,712	0,778	0,319	1,382	0,594	1,214	0,619	0,938

Not: F: Fowlkes-Mallows indeksi (1'e yakın = daha iyi). M: Mirkin indeksi (0'a yakın = daha iyi). VI: Bilgi Varyansı indeksi (0'a yakın = daha iyi). ORT.: 10 katın aritmetik ortalaması. Eğitim: modelin eğitildiği 9 kat; Test: doğrulama katı.

Tablo 5 ve 6'da görüldüğü gibi GA ve k-ortalama yöntemleri ile Iris verisine kümeleme analizi yapılmıştır. Elde edilen sonuçların geçerliliği için 10 katlı çapraz doğrulama yapılarak küme geçerlilik indeksleri hesaplanmıştır. İndeks değerlerinin hem "Eğitim" hem de "Test" durumunda genetik algoritmanın k-ortalama yönteminden daha iyi sonuçlar verdiği görülmüştür.

Simülasyon çalışmasında yapay veri setine Ward yöntemiyle de kümeleme analizi uygulanmış ve karşılaştırma yapılmıştır. Ancak gerçek veri setleri için Ward yöntemi ile karşılaştırma yapılamamıştır. Bunun nedeni, simülasyon çalışmasında çapraz doğrulama üzerinden geçerlilik indeksleri hesaplanmamış olmasıdır. Yani, simülasyon çalışmasında, deneme parametreleri değiştirilerek kümeleme analizi uygulanmış ve buna göre küme geçerlilik indeksleri hesaplanarak karşılaştırma yapılmıştır. Gerçek veri setleri içinse parametreler bilindiğinden çapraz doğrulama üzerinden geçerlilik indeksleri hesaplanmıştır. Ancak Ward yöntemi hiyerarşik kümeleme yaptığından Eğitim kısmının geçerlilik indeksleri hesaplanamamıştır. Bu nedenle önerilen genetik algoritmanın gerçek veri setleri için Ward yöntemi ile karşılaştırılması yapılamamıştır.

Gerçek veri setlerinin gözlem sayılarının (210 ve 150) simülasyon çalışmasında kullanılan veri setlerinden (50 ve 100) daha yüksek olması çalışmanın değerlendirilmesi açısından bir sınırlılık oluşturmamaktadır. Nitekim simülasyon bulgularında önerilen genetik algoritmanın gözlem sayısı arttıkça kümeleme performansını koruduğu ve bazı koşullarda daha başarılı sonuçlar verdiği görülmüştür. Bu nedenle Seeds ve Iris veri setlerinin kullanılması, yöntemin daha büyük veri yapıları üzerindeki performansının değerlendirilmesine de olanak sağlamıştır.

6. Tartışma ve sonuçlar

Bu çalışmada, birimlerin doğru kümelerde bulunmasını sağlamak amacıyla GA kullanılarak kümeleme analizi gerçekleştirilmiştir. Genetik algoritmanın etkinliğini değerlendirmek amacıyla bir simülasyon çalışması yapılmıştır. Bu kapsamda, R programında "mixsim" paketi (Melnykov vd., 2012) kullanılarak birim sayısı (n=50, 100), küme sayısı (k=2, 3, 4, 5) ve barOmega ($\bar{\omega}$=0.001 ve 0.05), sapan değer (outlier=0,1,2) ve gürültü (noise= 0,1,2 parametrelerinin farklı kombinasyonları için veri setleri üretilmiştir. Elde edilen veri setleri GA ile kümelenmiştir ve sonuçların karşılaştırılabilmesi amacıyla aynı veri setlerine k-ortalama ve Ward yöntemleri de uygulanmıştır. Kümeleme performansları, F, M ve VI geçerlilik indeksleri kullanılarak değerlendirilmiştir. Bulgular genel olarak genetik algoritmanın diğer yöntemlere kıyasla daha yüksek performans gösterdiğini ortaya koymuştur. Bununla birlikte, kümelerde çakışma olması, küme sayısının fazlalaşması ve veri setinde gürültü değişkenlerinin bulunması durumlarında k-ortalama ve Ward yöntemlerinin performansının GA yöntemine yaklaştığı gözlemlenmiştir. Öte yandan, birim sayısı ve sapan değerlerin varlığı performans farkı üzerinde belirgin bir etki oluşturmamıştır.

Simülasyon çalışmasına ek olarak, GA gerçek veri setleri olan Seeds ve Iris veri setlerine de uygulanmıştır. Yöntemin performansını karşılaştırmak amacıyla bu veri setlerine k-ortalama algoritması da uygulanmıştır. Eğitim aşamasında aşırı uyum (overfitting) problemini önlemek için 10 katlı çapraz doğrulama kullanılmıştır. Elde edilen sonuçlar, hem simülasyon hem de gerçek veri setleri üzerinde genetik algoritmanın daha doğru ve tutarlı sonuçlar verdiğini göstermiştir.

Elde edilen bulgular, önerilen uygunluk fonksiyonunun küme içi benzerlik, kümeler arası ayrışma ve silhouette genişliğini birlikte değerlendirmesinin, genetik algoritmanın daha doğru küme çözümlerine ulaşmasına katkı sağladığını göstermektedir. Bu yönüyle çalışma, yalnızca genetik algoritmanın kümeleme amacıyla kullanılmasını değil, aynı zamanda kümeleme başarısını artırmaya yönelik bütünleşik bir uygunluk fonksiyonu önermesi bakımından literatüre metodolojik katkı sunmaktadır. Bu bulgular doğrultusunda, bu çalışmanın literatüre katkı sağlayacak bazı geliştirme alanları bulunmaktadır. Gelecek çalışmalarda, önerilen yaklaşım çok amaçlı optimizasyon çerçevesine genişletilerek farklı küme geçerlilik indekslerinin eş zamanlı optimize edilmesi sağlanabilir. Ayrıca, küme sayısının önceden belirlenmesi yerine genetik algoritma içerisinde otomatik olarak belirlenmesine yönelik yöntemlerin geliştirilmesi, modelin gerçek hayat uygulamalarındaki kullanımını artıracaktır. Bunun yanı sıra, yöntemin yüksek boyutlu, büyük ölçekli ve karma veri setleri üzerinde test edilmesi genellenebilirlik açısından önem taşımaktadır. Özellikle biyomedikal, genomik ve sosyal bilimler alanlarında yapılacak uygulamalar literatüre önemli katkılar sunabilir. Ayrıca, genetik algoritmaların diğer yöntemlerle birlikte kullanıldığı hibrit yaklaşımlar da performansı artırabilecek önemli bir araştırma alanı olarak değerlendirilmektedir. Son olarak, yöntemin hesaplama maliyeti ve ölçeklenebilirliği daha ayrıntılı olarak incelenmeli ve parametre duyarlılığı analizleri ile elde edilen bulgular daha güçlü şekilde desteklenmelidir.

Bu çalışma önemli katkılar sunmakla birlikte bazı sınırlılıklara da sahiptir. Öncelikle, küme sayısının (k) önceden belirlenmiş olması yöntemin gerçek hayat uygulamalarındaki esnekliğini sınırlamaktadır. Pek çok uygulamada küme sayısı bilinmediğinden, bu durum yöntemin esnekliğini azaltmaktadır. Çalışma kapsamında kullanılan gerçek veri setleri sınırlı sayıdadır. Bu veri setleri literatürde yaygın olarak kullanılmasına rağmen, gerçek hayat verilerinin karmaşıklığını tam olarak yansıtmayabilir. Bu nedenle elde edilen sonuçların genellenebilirliği kısıtlı olabilir. Bunun yanında, genetik algoritmanın hesaplama maliyeti detaylı olarak analiz edilmemiştir. Geleneksel yöntemlere kıyasla daha yüksek hesaplama maliyeti gerektirmesi, özellikle büyük veri setlerinde yöntemin uygulanabilirliğini sınırlandırabilecek önemli bir faktördür.

Kaynaklar

Adak, M.F., Canpolat, K., ve Yumuşak, N. (2016). Genetik algoritma kullanan yapay sinir ağları ile ikili gaz karışımlarının sınıflandırılması. *4th International Symposium on Innovative Technologies in Engineering and Science (ISITES2016) 3-5 Nov 2016 Alanya/Antalya-Turkey*.

Akopov, A. S., & Beklaryan, L. A. (2025). Evolutionary synthesis of high-capacity reconfigurable multilayer road networks using a multiagent hybrid clustering-assisted genetic algorithm. *IEEE Access*.

Ayık, Y.Z., Özdemir, A., ve Yavuz, U. (2007). Lise türü ve lise mezuniyet başarısının, kazanılan fakülte ile ilişkisinin veri madenciliği tekniği ile analizi. *Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, *10*(2), 441-454.

Cömert, S.E., Gökler, S.H., ve Yazgan, H.R. (2016). Hücresel imalat sistemlerinin k-means algoritması ve genetik algoritma ile tasarlanması: bir uygulama. *Akademik Platform Mühendislik ve Fen Bilimleri Dergisi*, *4*(3).

Dua, D. ve Graff, C. (2019). UCI Machine Learning Repository. University of California, Irvine. https://archive.ics.uci.edu/ml

Fowlkes, E. ve Mallows, C. (1983). A method for comparing two hierarchical clusterings, *Journal of American Statistical Association*, *78*, 553-569.

Gaon, T., Gabay, Y., & Weiss Cohen, M. (2025). Optimizing electric vehicle routing efficiency using k-means clustering and genetic algorithms. *Future Internet*, *17*(3), 97.

Goldberg, D.E. (1989). Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley.

Gögebakan, M. ve Servi, T. (2020). Genetik algoritma kullanılarak verilerin karma normal modele dayalı kümelenmesi. *Erciyes Üniversitesi Fen Bilimleri Enstitüsü Fen Bilimleri Dergisi*, *35*(3), 12-23.

Hawkins, D.M. (2004). The problem of overfitting. *Journal of Chemical Information and Computer Sciences 44* (1), 1-12

Haznedar, B., Arslan, M., Kalınlı, A. (2017). Training ANFIS structure using genetic algorithm for liver cancer classification based on microarray gene expression data. *Sakarya University Journal of Science, 21* (1), 54-62

Holland, J.H. (1975). *Adaptation in natural and artificial systems*, Cambridge, MA: University of Michigan Press.

Khan, M., Chowdhury, M. A. M., Meem, N. A., & Khan, M. H. (2022, December). Neighborhood Distance Estimation for Tree-Based Hybrid Genetic Algorithm for Density Based Data Clustering. In *2022 4th International Conference on Electrical, Computer & Telecommunication Engineering (ICECTE)* (pp. 52-55). IEEE.

Lorr, M. (1993). *Cluster Analysis for Social Scientists*, San Francisco: Jossey-Bass.

Meila, M. (2006). Comparing clusters - an information based distance. *Journal of Multivariate Analysis, 98*, 873-895.

Melnykov, V., Chen, W.C. ve Maitra, R. (2012). MixSim: an R package for simulating data to study performance of clustering algorithms. *Journal of Statistical Software*, *51*(12), 1-25.

Mirkin, B. (1996). Mathematical classification and clustering. Dordrecht: Kluwer Academic Press.

Mor, M., Gupta, P., ve Sharma, P. (2014). A genetic algorithm approach for clustering—*International Journal of Engineering and Computer Science*, *3*(06).

Morris, T.P., White, I.R. ve Crowther, M.J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2102.

R Core Team (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/

Sriwindono, H., Rosa, P. P., & Pinaryanto, K. (2022, January). The Teacher Placement using K-Means Clustering and Genetic Algorithm. In *Conference Series* (Vol. 4, pp. 43-51).

Şengür, A., Karabatak, M., Türkoğlu, İ., ve İnce, M.C. (2005). Genetik kümeleme ile görüntü bölütleme. *Fırat Üniversitesi Doğu Araştırmaları Dergisi*, *4*(1), 63-67.

Tiwari, A.K., Sharma, L.K., ve Krishna, G.R. (2010). Entropy weighting genetic k-means algorithm for subspace clustering. *Entropy*, *7*(7).

Wang, Y., Wei, Y., Wang, X., Wang, Z., & Wang, H. (2023). A clustering-based extended genetic algorithm for the multidepot vehicle routing problem with time windows and three-dimensional loading constraints. *Applied Soft Computing*, *133*, 109922.

YiFei, L., Minh, H. L., Khatir, S., Sang-To, T., Cuong-Le, T., MaoSen, C., & Wahab, M. A. (2023). Structure damage identification in dams using sparse polynomial chaos expansion combined with hybrid K-means clustering optimizer and genetic algorithm. *Engineering Structures*, *283*, 115891.

Ek 1. Simülasyon Sonuçları

$\bar{\omega}$	k	n	Noise	Outlier	F			M			VI		
					GA	kmeans	ward	GA	kmeans	ward	GA	kmeans	ward
0,05	2	50	0	0	**0.894**	0,754	0,766	**261**	536	517	**0.353**	0,709	0,666
0,05	2	50	0	1	**0.874**	0,727	0,748	**323**	597	559	**0.435**	0,802	0,735
0,05	2	50	0	2	**0,858**	0,730	0,745	**373**	599	577	**0,497**	0,816	0,761
0,05	2	50	1	0	**0,890**	0,746	0,753	**270**	551	539	**0,364**	0,735	0,713
0,05	2	50	1	1	**0.874**	0,719	0,733	**317**	613	590	**0,432**	0,827	0,779
0,05	2	50	1	2	**0,851**	0,719	0,726	**385**	620	614	**0,520**	0,839	0,822
0,05	2	50	2	0	**0,882**	0,731	0,742	**290**	584	562	**0,387**	0,773	0,749
0,05	2	50	2	1	**0,870**	0,725	0,743	**327**	601	574	**0,446**	0,817	0,770
0,05	2	50	2	2	**0,849**	0,700	0,704	**389**	661	660	**0,526**	0,890	0,876
0,05	2	100	0	0	**0,898**	0,752	0,767	**1015**	2187	2082	**0,362**	0,743	0,711
0,05	2	100	0	1	**0.886**	0,743	0,760	**1171**	2276	2151	**0.414**	0,783	0,746
0,05	2	100	0	2	**0,876**	0,745	0,750	**1279**	2262	2244	**0,454**	0,796	0,790
0,05	2	100	1	0	**0.896**	0,736	0,752	**1035**	2315	2202	**0.369**	0,786	0,756
0,05	2	100	1	1	**0.884**	0,738	0,752	**1164**	2321	2216	**0.422**	0,805	0,767
0,05	2	100	1	2	**0,876**	0,745	0,750	**1279**	2262	2244	**0,454**	0,796	0,790
0,05	2	100	2	0	**0,894**	0,743	0,749	**1060**	2262	2232	**0,377**	0,777	0,765
0,05	2	100	2	1	**0,883**	0,739	0,740	**1174**	2309	2315	**0,423**	0,811	0,807
0,05	2	100	2	2	**0,870**	0,727	0,735	**1327**	2416	2375	**0,477**	0,858	0,836
0,05	3	50	0	0	**0,789**	0,762	0,753	**345**	402	416	**0,644**	0,697	0,713
0,05	3	50	0	1	**0,776**	0,763	0,728	**378**	415	472	**0,696**	0,718	0,801
0,05	3	50	0	2	**0,756**	0,743	0,713	**419**	457	512	**0,766**	0,791	0,854
0,05	3	50	1	0	**0,766**	0,760	0,747	**381**	400	425	**0,708**	0,716	0,738
0,05	3	50	1	1	**0,753**	0,721	0,714	**413**	478	496	**0,755**	0,830	0,839
0,05	3	50	1	2	**0,739**	0,714	0,701	**446**	502	525	**0,819**	0,859	0,902
0,05	3	50	2	0	**0,763**	0,742	0,729	**387**	432	459	**0,719**	0,763	0,784
0,05	3	50	2	1	**0,749**	0,734	0,707	**417**	453	505	**0,775**	0,797	0,866
0,05	3	50	2	2	**0,736**	0,719	0,693	**450**	495	545	**0,828**	0,854	0,912
0,05	3	100	0	0	**0,822**	0,704	0,781	**1178**	1438	1499	**0,611**	0,666	0,677
0,05	3	100	0	1	**0,814**	0,784	0,768	**1241**	1505	1606	**0,647**	0,709	0,721
0,05	3	100	0	2	**0,803**	0,783	0,759	**1332**	1510	1699	**0,693**	0,727	0,763
0,05	3	100	1	0	**0,802**	0,792	0,777	**1305**	1430	1513	**0,663**	0,672	0,712
0,05	3	100	1	1	**0,788**	0,773	0,760	**1414**	1553	1649	**0,719**	0,741	0,769
0,05	3	100	1	2	**0,784**	0,768	0,757	**1461**	1603	1682	**0,738**	0,769	0,797
0,05	3	100	2	0	**0,805**	0,790	0,677	**1370**	1422	1601	0,697	**0,677**	0,748
0,05	3	100	2	1	**0,785**	0,771	0,752	**1434**	1580	1701	**0,729**	0,745	0,795
0,05	3	100	2	2	**0,784**	0,766	0,736	**1460**	1646	1830	**0,750**	0,769	0,838
0,05	4	50	0	0	**0.701**	0,676	0,694	**368**	496	480	0,895	0,891	**0.817**
0,05	4	50	0	1	**0,694**	0,669	0,675	**383**	527	521	0,927	0,927	**0,895**
0,05	4	50	0	2	**0,675**	0,663	0,656	**421**	551	576	0,988	0,966	**0,958**
0,05	4	50	1	0	**0,698**	0,669	0,675	**370**	501	506	0,903	0,917	**0,880**
0,05	4	50	1	1	**0,662**	0,656	0,660	**424**	535	545	1,018	0,967	**0,943**
0,05	4	50	1	2	**0,652**	0,632	0,648	**444**	591	584	1,068	1,056	**0,992**
0,05	4	50	2	0	0,657	0,662	**0,664**	**417**	515	531	1,022	0,936	**0,916**
0,05	4	50	2	1	0,641	**0,647**	0,636	**447**	559	577	1,081	**0,992**	1,021
0,05	4	50	2	2	**0,639**	0,629	0,630	**460**	602	608	1,107	1,069	**1,045**
0,05	4	100	0	0	**0.727**	0,682	0,692	**1364**	1975	1988	0,887	0,920	**0.849**
0,05	4	100	0	1	**0.712**	0,678	0,686	**1456**	2034	2056	0,937	0,952	**0.888**
0,05	4	100	0	2	**0,713**	0,666	0,677	**1470**	2131	2139	0,955	0,999	**0,929**

Ek 1. Simülasyon Sonuçları (devam)

$\bar{\omega}$	k	n	Noise	Outlier	F			M			VI		
					GA	kmeans	ward	GA	kmeans	ward	GA	kmeans	ward
0,05	4	100	1	0	**0.705**	0.674	0.684	**1473**	1996	2048	0.962	0.948	**0.888**
0,05	4	100	1	1	**0.689**	0,667	0,669	**1568**	2077	2137	1,007	0,985	**0.957**
0,05	4	100	1	2	**0,680**	0,662	0,659	**1627**	2141	2238	1,050	1,021	**1,003**
0,05	4	100	2	0	**0,689**	0,669	0,661	**1547**	2062	2187	1,003	0,967	**0,966**
0,05	4	100	2	1	**0,677**	0,652	0,662	**1622**	2191	2190	1,057	1,034	**0,976**
0,05	4	100	2	2	**0,663**	0,650	0,646	**1710**	2230	2349	1,114	1,056	**1,045**
0,05	5	50	0	0	**0,637**	0,600	0,613	**353**	587	582	1,057	1,102	**1,021**
0,05	5	50	0	1	**0,624**	0,586	0,604	**373**	619	617	1,104	1,169	**1,068**
0,05	5	50	0	2	**0,617**	0,573	0,588	**388**	653	663	1,137	1,230	**1,130**
0,05	5	50	1	0	**0,602**	0,592	0,600	**385**	588	601	1,167	1,135	**1,073**
0,05	5	50	1	1	**0.593**	0,577	0,590	**402**	622	627	1,205	1,205	**1,131**
0,05	5	50	1	2	**0,592**	0,564	0,578	**410**	665	672	1,208	1,260	**1,192**
0,05	5	50	2	0	0,574	0,585	**0.590**	**411**	600	611	1,256	1,163	**1.119**
0,05	5	50	2	1	0,572	0,573	**0.584**	**421**	622	630	1,279	1,225	**1.174**
0,05	5	50	2	2	0,562	**0,568**	0,563	**440**	652	687	1,320	1,255	**1,240**
0,05	5	100	0	0	**0,649**	0,618	0,623	**1405**	2273	2338	1,124	1,093	**1,038**
0,05	5	100	0	1	**0,646**	0,612	0,615	**1427**	2331	2434	1,142	1,129	**1,080**
0,05	5	100	0	2	**0,640**	0,605	0,607	**1471**	2410	2526	1,179	1,175	**1,123**
0,05	5	100	1	0	**0,628**	0,608	0,608	**1477**	2323	2414	1,211	1,130	**1,106**
0,05	5	100	1	1	**0,616**	0,601	0,607	**1548**	2377	2461	1,240	1,176	**1,128**
0,05	5	100	1	2	**0,608**	0,596	0,599	**1598**	2430	2552	1,294	1,214	**1,165**
0,05	5	100	2	0	0,602	**0,609**	0,602	**1575**	2313	2461	1,293	1,135	**1,129**
0,05	5	100	2	1	0,588	**0,606**	0,593	**1651**	2365	2571	1,340	**1,159**	1,175
0,05	5	100	2	2	0,585	**0,588**	**0,588**	**1683**	2498	2599	1,367	1,235	**1,218**
0,001	2	50	0	0	**0,997**	0,856	0,858	**8**	326	322	**0,013**	0,361	0,348
0,001	2	50	0	1	**0,974**	0,853	0,849	**71**	331	340	**0,104**	0,398	0,398
0,001	2	50	0	2	**0,954**	0,850	0,862	**123**	340	313	**0,180**	0,427	0,381
0,001	2	50	1	0	**0,992**	0,843	0,840	**20**	353	360	**0,027**	0,400	0,397
0,001	2	50	1	1	**0,970**	0,829	0,818	**75**	383	406	**0,118**	0,466	0,471
0,001	2	50	1	2	**0,952**	0,814	0,812	**124**	417	424	**0,188**	0,526	0,502
0,001	2	50	2	0	**0,989**	0,837	0,837	**27**	366	367	**0,037**	0,416	0,409
0,001	2	50	2	1	**0,972**	0,824	0,822	**70**	393	398	**0,109**	0,482	0,469
0,001	2	50	2	2	**0,951**	0,814	0,807	**126**	418	435	**0,192**	0,524	0,531
0,001	2	100	0	0	**0,995**	0,859	0,866	**49**	1292	1236	**0,023**	0,368	0,347
0,001	2	100	0	1	**0,982**	0,853	0,856	**196**	1350	1322	**0,078**	0,403	0,392
0,001	2	100	0	2	**0,973**	0,852	0,857	**293**	1354	1314	**0,120**	0,424	0,406
0,001	2	100	1	0	**0,990**	0,840	0,850	**97**	1454	1369	**0.036**	0,421	0,382
0,001	2	100	1	1	**0,980**	0,833	0,839	**203**	1511	1469	**0,087**	0,456	0,436
0,001	2	100	1	2	**0,970**	0,826	0,834	**308**	1576	1518	**0,135**	0,496	0,467
0,001	2	100	2	0	**0,989**	0,839	0,845	**108**	1462	1414	**0,041**	0,429	0,403
0,001	2	100	2	1	**0,972**	0,839	0,843	**278**	1465	1436	**0,111**	0,448	0,431
0,001	2	100	2	2	**0,968**	0,827	0,835	**325**	1572	1508	**0,142**	0,510	0,474
0,001	3	50	0	0	**0,989**	0,951	0,987	**19**	100	21	**0,037**	0,119	0,044
0,001	3	50	0	1	0,964	0,928	**0,965**	64	145	**62**	**0,123**	0,198	0,124
0,001	3	50	0	2	**0,949**	0,910	0,942	**91**	184	102	**0,186**	0,266	0,194
0,001	3	50	1	0	0,946	0,900	**0,950**	89	179	**83**	0,151	0,252	**0,136**
0,001	3	50	1	1	**0,929**	0,885	0,920	**119**	205	136	**0,220**	0,316	0,243
0,001	3	50	1	2	**0,907**	0,854	0,902	**160**	270	170	**0,302**	0,415	0,307

Ek 1. Simülasyon Sonuçları (devam)

$\bar{\omega}$	k	n	Noise	Outlier	F			M			VI		
					GA	kmeans	ward	GA	kmeans	ward	GA	kmeans	ward
0,001	3	50	2	0	**0.933**	0.865	0.924	**111**	243	127	**0.193**	0.338	0.207
0,001	3	50	2	1	**0.914**	0,856	0,907	**144**	266	159	**0.260**	0,380	0,273
0,001	3	50	2	2	**0,898**	0,838	0,896	**176**	299	180	**0,328**	0,462	0,329
0,001	3	100	0	0	**0,993**	0,946	0,992	**43**	437	49	**0,028**	0,130	0,032
0,001	3	100	0	1	**0,983**	0,934	0,982	**111**	535	122	**0,073**	0,180	0,078
0,001	3	100	0	2	**0,973**	0,921	0,971	**191**	645	197	**0,116**	0,234	0,125
0,001	3	100	1	0	**0,960**	0,917	0,958	**265**	646	282	**0,120**	0,209	0,129
0,001	3	100	1	1	**0,953**	0,897	0,952	**311**	768	324	**0,157**	0,287	0,158
0,001	3	100	1	2	0,945	0,899	**0,949**	374	763	**349**	0,198	0,298	**0,185**
0,001	3	100	2	0	**0,957**	0,890	0,951	**282**	836	331	**0,130**	0,285	0,158
0,001	3	100	2	1	**0,940**	0,878	0,935	**400**	918	438	**0,199**	0,337	0,214
0,001	3	100	2	2	**0,932**	0,862	0,925	**458**	1093	511	**0,240**	0,388	0,255
0,001	4	50	0	0	**0,981**	0,813	0,827	**23**	307	280	**0,063**	0,386	0,340
0,001	4	50	0	1	**0,959**	0,790	0,812	**54**	368	314	**0,132**	0,473	0,409
0,001	4	50	0	2	**0,939**	0,782	0,798	**82**	383	348	**0,201**	0,517	0,469
0,001	4	50	1	0	**0,892**	0,780	0,813	**133**	359	303	**0,298**	0,485	0,384
0,001	4	50	1	1	**0,878**	0,771	0,796	**153**	394	341	**0,353**	0,531	0,454
0,001	4	50	1	2	**0,860**	0,738	0,775	**181**	447	386	**0,419**	0,654	0,532
0,001	4	50	2	0	**0,860**	0,771	0,797	**171**	369	324	**0,402**	0,512	0,446
0,001	4	50	2	1	**0,838**	0,744	0,778	**202**	424	370	**0,467**	0,616	0,521
0,001	4	50	2	2	**0,835**	0,734	0,757	**211**	445	415	**0,504**	0,673	0,596
0,001	4	100	0	0	**0,987**	0,805	0,821	**64**	1320	1196	**0,050**	0,418	0,349
0,001	4	100	0	1	**0,974**	0,798	0,814	**134**	1377	1256	**0,098**	0,454	0,387
0,001	4	100	0	2	**0,959**	0,786	0,807	**220**	1504	1328	**0,150**	0,500	0,423
0,001	4	100	1	0	**0,913**	0,780	0,810	**432**	1482	1263	**0,254**	0,493	0,386
0,001	4	100	1	1	**0,907**	0,776	0,808	**464**	1482	1295	**0,278**	0,525	0,410
0,001	4	100	1	2	**0,897**	0,768	0,798	**519**	1591	1380	**0,333**	0,567	0,457
0,001	4	100	2	0	**0,886**	0,761	0,804	**565**	1561	1302	**0,343**	0,560	0,410
0,001	4	100	2	1	**0,879**	0,764	0,798	**607**	1566	1355	**0,376**	0,565	0,448
0,001	4	100	2	2	**0,856**	0,762	0,787	**728**	1623	1448	**0,462**	0,587	0,489
0,001	5	50	0	0	**0,977**	0,701	0,729	**23**	486	432	**0,068**	0,661	0,561
0,001	5	50	0	1	**0,948**	0,693	0,712	**52**	506	474	**0,157**	0,716	0,630
0,001	5	50	0	2	**0,930**	0,683	0,699	**73**	531	512	**0,224**	0,774	0,693
0,001	5	50	1	0	**0,849**	0,694	0,717	**148**	466	448	**0,408**	0,710	0,608
0,001	5	50	1	1	**0,825**	0,686	0,703	**174**	496	490	**0,486**	0,751	0,664
0,001	5	50	1	2	**0,824**	0,674	0,691	**180**	538	524	**0,507**	0,816	0,723
0,001	5	50	2	0	**0.790**	0,691	0,708	**204**	472	461	**0.569**	0,727	0,639
0,001	5	50	2	1	**0,782**	0,673	0,697	**216**	512	488	**0,615**	0,803	0,692
0,001	5	50	2	2	**0,763**	0,670	0,681	**240**	537	537	**0,680**	0,833	0,763
0,001	5	100	0	0	**0,976**	0,706	0,722	**97**	1952	1819	**0,088**	0,656	0,589
0,001	5	100	0	1	**0,954**	0,699	0,715	**188**	2021	1894	**0,158**	0,696	0,628
0,001	5	100	0	2	**0,950**	0,698	0,710	**206**	2045	1956	**0,185**	0,724	0,665
0,001	5	100	1	0	**0,847**	0,685	0,715	**614**	2049	1876	**0,430**	0,753	0,611
0,001	5	100	1	1	**0,833**	0,679	0,713	**674**	2111	1914	**0,482**	0,785	0,638
0,001	5	100	1	2	**0,840**	0,671	0,703	**659**	2179	2009	**0,483**	0,830	0,684
0,001	5	100	2	0	**0.812**	0,686	0,713	**743**	1965	1899	**0.572**	0,752	0,615
0,001	5	100	2	1	**0.806**	0,679	0,707	**774**	2067	1927	**0.601**	0,790	0,666
0,001	5	100	2	2	**0,800**	0,675	0,700	**814**	2122	2022	**0,623**	0,816	0,702