
GAZİ ÜNİVERSİTESİ
Gazi
1926

İKİLİ KÜMELEME ALGORİTMALARININ GÖRSEL VE SAYISAL AÇIDAN KARŞILAŞTIRILMASI

Ahmet KOCATÜRK

YÜKSEK LİSANS TEZİ
İSTATİSTİK ANABİLİM DALI

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

OCAK 2018

Ahmet KOCATÜRK tarafından hazırlanan "İKİLİ KÜMELEME ALGORİTMALARININ GÖRSEL VE SAYISAL AÇIDAN KARŞILAŞTIRILMASI" adlı tez çalışması aşağıdaki jüri tarafından OY BİRLİĞİ ile Gazi Üniversitesi İstatistik Anabilim Dalında YÜKSEK LİSANS TEZİ olarak kabul edilmiştir.

Danışman: Doç. Dr. Bülent ALTUNKAYNAK

İstatistik Anabilim Dalı, Gazi Üniversitesi

Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum. ...……

Başkan: Prof. Dr. Özgür YENİAY

İstatistik Anabilim Dalı, Hacettepe Üniversitesi

Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum. ……...

Üye: Doç. Dr. Hacı Hasan ÖRKCÜ

İstatistik Anabilim Dalı, Gazi Üniversitesi

Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum. ……...

Tez Savunma Tarihi: 25 / 01 / 2018

Jüri tarafından kabul edilen bu tezin Yüksek Lisans Tezi olması için gerekli şartları yerine getirdiğini onaylıyorum.

……..…….

Prof. Dr. Sena YAŞYERLİ

Fen Bilimleri Enstitüsü Müdürü

ETİK BEYAN

Gazi Üniversitesi Fen Bilimleri Enstitüsü Tez Yazım Kurallarına uygun olarak hazırladığım bu tez çalışmasında;

- Tez içinde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi,
- Tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu,
- Tez çalışmasında yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi,
- Kullanılan verilerde herhangi bir değişiklik yapmadığımı,
- Bu tezde sunduğum çalışmanın özgün olduğunu,

bildirir, aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim.

Ahmet KOCATÜRK

25 / 01 / 2018

İKİLİ KÜMELEME ALGORİTMALARININ GÖRSEL VE SAYISAL AÇIDAN KARŞILAŞTIRILMASI

(Yüksek Lisans Tezi)

Ahmet KOCATÜRK

GAZİ ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

Ocak 2018

ÖZET

Gen açıklama verilerinde benzer ifade yapılarına göre gen gruplarını belirlemek oldukça önemlidir. Bu veriler için yapılacak kümeleme analizlerinde son zamanlarda popüler olan ikili kümeleme yöntemleri kullanılmaktadır. İkili kümeleme yöntemlerinde farklı gen açıklama veri yapıları için çok sayıda algoritma önerilmiştir. Araştırmanın amacına göre elde edilecek ikili kümelerin etkinliğini ölçmek için bu algoritmaların performanslarına bakılması gerekir. Bu çalışmada en yaygın olarak kullanılan CC, Bimax, Plaid, Spectral, Quest ve Xmotif algoritmalarının performansları görsel ve sayısal olarak karşılaştırılmıştır. Bu algoritmaların görsel karşılaştırmasında ikili kümelerin ısı grafiklerine bakılmıştır. Sayısal karşılaştırılmasında ise varyans ölçüsü (VAR), ortalama karesel artık skoru (MSR), uygunluk indeksi (UI), ölçeklenen ortalama karesel artık skoru (SMSR), Chia ve Karuturi ikili küme skoru (CKSB), ortalama korelasyon ölçüsü (ACV), alt matris korelasyon ölçüsü (SCS), ortalama Spearman korelasyon değeri (ASR), Spearman ikili küme ölçüsü (SBM) ve sanal hata (VE) ikili küme değerlendirme ölçüleri hesaplanmıştır. Değerlendirme ölçüleri hesaplaması R fonksiyonları ile oluşturulmuş ve analizler bu kodlara uygulanarak gerçekleşmiştir. Farklı veri yapılarında karşılaştırma yapmak için yapay ve gerçek veriler kullanılmıştır. Yapay veri seti uygulamasında dört farklı senaryo ile ikili kümeler oluşturulmuştur. Bunlar ikili kümeler arasında örtüşme ve aykırı değerlerin olup olmadığı durumlardır. Gerçek veri seti uygulamasında ise maya verisi, lenf hücrelerinin gen ifadesini içeren insan verisi ve protein-protein etkileşim skorlarını içeren fare verisi kullanılmıştır. Yapılan analizler sonucunda hangi algoritmanın hangi veri setinde daha anlamlı ikili kümeler elde ettiği belirlenmiştir.

Bilim Kodu : 20514
Anahtar Kelimeler : İkili kümeleme, Gen açıklama verisi, Değerlendirme ölçüleri
Sayfa Adedi : 82
Danışman : Doç. Dr. Bülent ALTUNKAYNAK

VISUAL AND NUMERICAL COMPARISON OF BICLUSTERING ALGORITHMS

(M. Sc. Thesis)

Ahmet KOCATÜRK

GAZİ UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

January 2018

ABSTRACT

It is very important to identify gene groups according to similar expressions in gene expression data. Biclustering methods, which are popular recently, are used in the clustering analysis for this data. Numerous algorithms have been proposed for different gene expression data structures in biclustering methods. The performance of these algorithms needs to be examined in order to measure the effectiveness of the biclusters obtained for the purpose of the study. In this study, the performances of the most commonly used CC, Bimax, Plaid, Spectral, Quest and Xmotif algorithms are visually and numerically compared. In the visual comparison of these algorithms, the heatmaps of biclusters are looked at. In numerical comparison, variance measure (VAR), mean squared residual score (MSR), fitness index (UI), scaled mean squared residual score (SMSR), Chia and Karuturi bicluster score (CKSB), average correlation measure (SPS), Spearman correlation coefficient (ASR), Spearman bicluster measure (SBM), and virtual error (VE) bicluster evaluation measures were calculated. Calculation of evaluation measures was made with R functions and analyzes were applied to these codes. Artificial and real data are used to compare different data structures. In the application of artificial data set, biclusters were formed with four different scenarios. These are cases where there is overlap between the biclusters and whether there are outliers. In the real data set, yeast, human data containing the gene expression of lymphoma cells and mouse data containing protein-protein interaction scores were used. As a result of the analysis, it was determined which algorithm obtained more meaningful biclusters in which data set.

Science Code : 20514
Key Words : Biclustering, Gene expression data, Evaluation measures
Page Number : 82
Supervisor : Assoc. Prof. Dr. Bülent ALTUNKAYNAK

TEŞEKKÜR

Çalışmam boyunca katkılarından, desteklerinden ve her türlü yardımından dolayı saygıdeğer tez danışmanım Doç. Dr. Bülent ALTUNKAYNAK'a verdiği emeklerinden dolayı teşekkür ederim.

Sadece tez çalışmam boyunca değil, hayatımın her anında bana yoldaş, sırdaş, arkadaş ve tam anlamıyla eş olan çok sevdiğim eşim Sevim KOCATÜRK'e çalışmam boyunca özverili bir şekilde destek olduğu için teşekkür ederim. İyi ki varsın.

Hayatım boyunca tüm davranışlarıyla örnek aldığım, akademisyen olmamdaki en büyük pay sahibi ve hiçbir zaman maddi ve manevi yardımlarını esirgemeyen ağabeyim Arş. Gör. Dr. Metin KOCATÜRK'e teşekkür ederim.

Araştırma sürecinde beni yalnız bırakmadan bana her türlü desteği sabırla veren değerli meslektaşlarım Arş. Gör. Mihraç KÜPELİ'ye, Arş. Gör. Esra ÖZKAN AKSU ve eşi Sinan Davut AKSU'ya teşekkür ederim.

Sevgili babam Mehmet Suat KOCATÜRK'e ve biricik annem Nuran KOCATÜRK'e, ablalarım Seyhan ÖRNEK ve Selin BÜLBÜL'e teşekkür ederim. Her ne kadar benimle gurur duyduğunuzu dile getirseniz de asıl gurur duyması gereken kişi benim, size sahip olduğum için.

İÇİNDEKİLER

Sayfa

Sayfa

ÇİZELGELERİN LİSTESİ

Çizelge **Sayfa**

ŞEKİLLERİN LİSTESİ

Şekil **Sayfa**

Şekil **Sayfa**

Şekil **Sayfa**

SİMGELER VE KISALTMALAR

Bu çalışmada kullanılmış simgeler ve kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

Simgeler	**Açıklamalar**
α	Alfa
β	Beta
δ	Delta
μ	Mü
π	Pi

Kısaltmalar	**Açıklamalar**
ACV	Ortalama korelasyon ölçüsü
ANOVA	Varyans analizi
ASR	Ortalama Spearman korelasyon değeri
BBC	Bayes ikili küme algoritması
CC	Cheng ve Church algoritması
CE	Kümeleme hatası
CKSB	Chia ve Kauturi ikili küme skoru
CRAN	Bilgisayar araştırma savunma ağı
CSI	Campello soft indeksi
DBF	Dijital Beamforming algoritması
FABIA	İkili küme edinmek için faktör analizi
FLOC	Esnek örtüşen ikili kümeleme algoritması
GO	Gen ontolojisi
GOWE	Gen ontoloji ağırlıklı zenginleştirme skoru
ISA	Yinelemeli işaret algoritması
LSS	Katman kareler toplamı
MPM	Metabolik yol haritaları
MSR	Ortalama karesel artık değeri

Kısaltmalar	Açıklamalar
OPSM	Sıra korumalı alt matrisler
QUBIC	Nitel ikili küme algoritması
PCC	Pearson korelasyon katsayısı
PPI	Protein-protein etkileşim
SBM	Spearman ikili küme ölçüsü
SCS	Spearman alt matris korelasyon ölçüsü
SMSR	Ölçeklenen ortalama karesel artık skoru
UI	Uygunluk indeksi
VAR	Varyans ölçüsü
VE	Sanal hata

1. GİRİŞ

Son yıllarda gen açıklama (gene expression) verilerinin ikili kümeleme (biclustering) yaklaşımıyla analiz edilmesi yaygınlaşmıştır. Bunun nedeni farklı koşulların (conditions) alt kümeleriyle ilişkili gen kümelerini tespit etme çabasıdır. Bu sayede ortak-açıklama genleri (co-expressed or co-regulated) tanımlanabilmektedir. Ancak farklı gen açıklama verilerinin varlığı ve bu verilerin büyük hacimli veriler olması analiz işlemini oldukça zorlaştırmaktadır. Bu nedenle literatürde ikili kümeleme için birçok farklı sezgisel algoritma önerilmiştir. Bunlardan bazıları Cheng ve Church (2000) tarafından önerilen CC algoritması, Prelic, Bleuler, Zimmermann, Wille, Bühlmann, Gruissem, Hennig, Thiele ve Zitzler (2006) tarafından önerilen Bimax algoritması, Lazzeroni ve Owen (2000) tarafından önerilen Plaid algoritması, Murali ve Kasif (2003) tarafından önerilen Quest ve Xmotif algoritmaları ve Kluger, Basri ve Gerstein (2003) tarafından önerilen Spectral algoritması örnek olarak verilebilir. Bununla birlikte, bu algoritmaların etkinliğine ilişkin çalışmalar sınırlı sayıdadır.

Prelic ve diğerleri (2006) ikili kümeler içeren yapay gen açıklama verileri üretmişler ve CC, Bimax, OPSM, SAMBA, Xmotif ve ISA ikili kümeleme algoritmalarını karşılaştırmışlardır. Karşılaştırma ölçüsü olarak algoritmalardan elde edilen ikili kümeler ile yapay verideki ikili kümelerin benzerlikleri kullanılmıştır. Farklı veri yapılarında algoritmaların farklı sonuçlar verdiği görülmüştür. Gerçek veri üzerinden yaptıkları karşılaştırmada ise ISA, SAMBA ve OPSM algoritmalarının daha iyi sonuçlar verdiğini ifade etmişlerdir.

Das ve Idicula (2011) ortalama kare artığı (Mean Squared Residue, MSR) eşiğini ve MSR fark eşiğini kullanan MSRT, MSRDT ve ISIMSRDT ikili kümeleme algoritmalarını CC, SEBI, FLOC ve DBF algoritmalarıyla karşılaştırmıştır. Karşılaştırma için iki gerçek veri ve ortalama MSR, ortalama hacim ve gen sayılarını kullanmışlardır. Bu üç algoritmanın diğerlerine göre daha iyi sonuçlar verdiğini göstermişlerdir.

Bhattacharya, Krenitsky, Huyck, Solleti, Lunger ve Pryhuber (2012) farklı gen açıklama verileri için kümeleme ve ikili kümeleme algoritmalarını farklı küme doğrulama indeksleri açısından karşılaştırmışlardır. Çalışmada delta, OPSM, SAMBA, ISA, Bimax ve bioNMF

ikili kümeleme algoritmaları dikkate alınmıştır. Elde edilen sonuçlar en iyi performansı SAMBA, en kötü performansı ise delta algoritmasının verdiğini göstermiştir.

Li, Guo, Wu, Shi, Cheng, ve Tao (2012), Bimax, FABIA, ISA, QUBIC ve SAMBA ikili kümeleme algoritmalarını iki farklı veri setini (GDS1620 ve Pathway) kullanarak karşılaştırmışlardır. Karşılaştırmalarda ikili kümelerin değerlendirilmesi için Gen ontoloji ağırlıklı kuvvetlendirme skoru (Gene ontology weighted enrichment score, GOWE) ve protein-protein etkileşim skoru (protein-protein interaction score, PPI) kullanılmıştır. GDS1620 verisi için ISA algoritması en iyi sonucu verirken Pathway verisi için en iyi skorlar Bimax algoritmasından elde edilmiştir.

Zhao, Wee-Chung Liew, Wang ve Yan (2012), GO, metabolik yol haritaları (metabolic pathway maps, MPM) ve PPI skorlarını kullanarak ikili kümeleme algoritmalarını karşılaştırmış ve farklı senaryolar için BBC (Bayesian Biclustering Algorithm) nin iyi performans gösterdiği sonucuna ulaşmışlardır.

Eren, Deveci, Küçüktunç ve Çatalyürek (2012), ikili kümeleme algoritmalarını yapay veriler üzerinden küme skorunu (set score) kullanarak karşılaştırmışlardır. Küme skorları başlangıç kümesiyle algoritma sonunda elde edilen kümenin uyumuna dayalıdır. Benzer şekilde Padilha ve Campello (2017) kümeleme hatası (clustering error, CE) ve Campello soft indeksi (CSI) kullanarak ikili kümeleme algoritmalarını yapay ve gerçek veriler üzerinden karşılaştırmışlardır. Sonuç olarak her iki çalışmada da farklı senaryolar için farklı sonuçlar elde edilmiştir (Tüm senaryolar için geçerli etkin bir algoritma elde edilememiştir).

Bu tez çalışmasında amaç ikili kümeleme algoritmalarının daha geniş bir şekilde karşılaştırılmasını yapmaktır. Çalışmada CC, Bimax, Plaid, Quest, Spectral ve Xmotif ikili küme algoritmaları varyans ölçüsü (VAR), ortalama karesel artık skoru (MSR), uygunluk indeksi (UI), ölçeklenen ortalama karesel artık skoru (SMSR), Chia ve Karuturi ikili küme skoru (CKSB), ortalama korelasyon ölçüsü (ACV), alt matris korelasyon ölçüsü (SCS), ortalama Spearman korelasyon değeri (ASR), Spearman ikili küme ölçüsü (SBM) ve sanal hata (VE) ikili küme değerlendirme skorları bakımından yapay ve gerçek veriler kullanılarak farklı veri yapıları açısından karşılaştırılmıştır. Elde edilen sonuçlar ısı

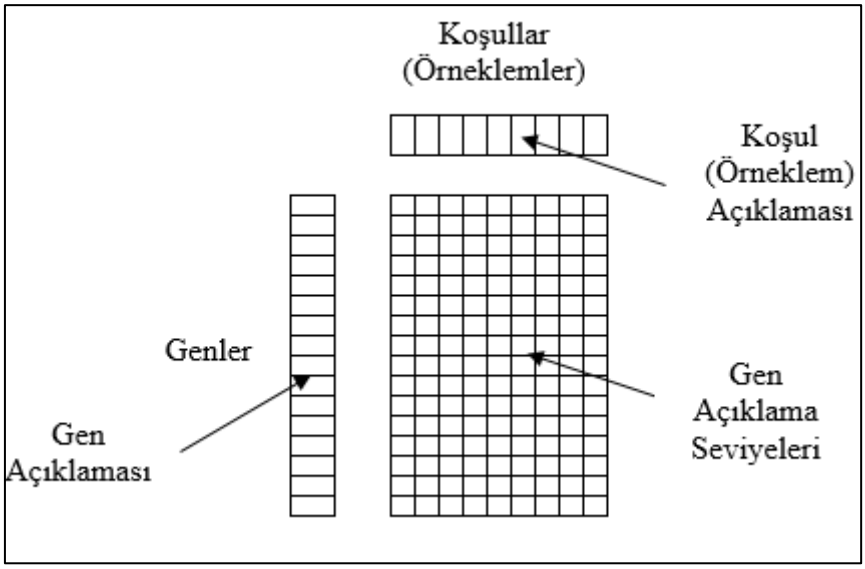
grafikleri ve tablolar aracılığıyla sunulmuştur. Çalışmada değerlendirme ölçüleri için R fonksiyonları oluşturulmuş ve tüm analizler R program kodlarıyla uygulanmıştır.

Tez çalışmasının ikinci bölümünde çalışmaya dâhil edilen ikili küme algoritmaları tanıtılmaktadır. Üçüncü bölümde ise ikili kümelerin değerlendirilmesinde kullanılabilecek değerlendirme ölçüleri verilmekte ve bu ölçülerin nasıl hesaplandığı küçük veri matrisi üzerinden gösterilmektedir. Dördüncü bölümde yapay ve gerçek veriler üzerinden yapılan uygulamalar ve sonuçları yer almaktadır. Bu bölümde yapay verilerin üretilmesiyle ilgili aşamalar ve sonuçların karşılaştırılmasına ilişkin çıktılar verilmiştir. Son bölümde çalışmadan elde edilen sonuçlar verilmiş ve önerilerde bulunulmuştur.

2. İKİLİ KÜMELEME YÖNTEMLERİ

2.1. Tanım

İkili kümeleme terimi ilk olarak Mirkin (1996) tarafından kullanılarak bir gen açıklama veri matrisinde hem satır hem de sütun kümelerinin eş zamanlı kümelenmesi olarak belirtilmiştir. Gen açıklama veri matrisleri birçok koşullardan (örneklemden) elde edilen binlerce genin açıklamasını içeren çok boyutlu verilerdir. Bir gen açıklama verisi Şekil 2.1'de gösterilmiştir.



Şekil 2.1. Gen açıklama verisi

Gen açıklama verilerinin analizi çalışmanın amacına göre farklılık gösterebilmektedir. Örneğin, belirli koşullara (örneklere) göre gen açıklama değerlerini kullanarak genleri gruplamak veya belirli genlere göre koşulları (örnekleri) gruplamak istenebilir. Bu tür amaçlar için klasik kümeleme algoritmaları kullanılabilmektedir. Ancak son yıllarda hem genleri hem de koşulları aynı anda kümeleyerek gen örüntüleri ortaya çıkartmak daha önemli hale gelmiştir. Bu amaç için ikili kümeleme algoritmaları geliştirilmiştir. Bir ikili küme yapısı aşağıdaki gibi ifade edilebilir.

$$B = \begin{pmatrix} b_{11} & b_{12} & . & . & . & b_{1|J|} \\ b_{21} & b_{22} & . & . & . & b_{2|J|} \\ . & . & . & & & . \\ . & . & & . & & . \\ . & . & & & . & . \\ b_{|I|1} & b_{|I|2} & . & . & . & b_{|I||J|} \end{pmatrix}$$

Burada b_{ij} i inci satır (gen) j inci sütun (koşul veya örneklem) elemanını göstermektedir. Bu veri matrisinden yola çıkarak ikili kümeleme algoritmalarında yaygın olarak kullanılan bazı formüller aşağıdaki gibi verilebilir.

$$b_{Ij} = \frac{1}{|I|}\sum_{i=1}^{|I|} b_{ij} \tag{2.1}$$

$$b_{iJ} = \frac{1}{|J|}\sum_{j=1}^{|J|} b_{ij} \tag{2.2}$$

$$b_{IJ} = \frac{1}{|I||J|}\sum_{i=1}^{|I|}\sum_{j=1}^{|J|} b_{ij} \tag{2.3}$$

Burada, b_{Ij} j inci sütunun ortalaması, b_{iJ} i inci satırın ortalaması ve b_{IJ} genel ortalamayı göstermektedir.

<u>İkili küme türleri</u>

Gen verilerinin içerisinde aldığı değerler bakımından farklı değerler alabilen ikili küme türleri bulunabilir. Bir ikili kümenin değerleri sabitlerden oluşabileceği gibi satırlara veya sütunlara göre artan yapıda olabilir. Sadece satırlar bazında toplamsal veya çarpımsal artışlar gösterebileceği gibi aynı durumu sütunlar bazında da gösterebilir. Dahası toplamsal ve çarpımsal değerlerin karmasından (cohorent) oluşan bir ikili küme yapısı da mümkündür. Bu ikili küme türleri aşağıda açıklanmıştır.

Sabit ikili küme türleri

Sabit değerli ikili kümeler kendi içerisinde üçe ayrılır. Bunlar; sabit, satır sabit ve sütun sabit ikili kümelerdir. Sabit ikili küme elemanları $b_{ij} = \pi$ şeklinde değerler alırken, satır sabit ikili kümeler $b_{ij} = \pi + \beta_j$ (toplamsal) veya $b_{ij} = \pi \times \beta_j$ (çarpımsal) şeklinde değerler alır. Benzer şekilde sütun sabit ikili kümelerin değerleri ise $b_{ij} = \pi + \beta_i$ (toplamsal) veya $b_{ij} = \pi \times \beta_i$ (çarpımsal) şeklinde ifade edilebilir.

Sabit ikili küme türlerinin örnek gösterimi Şekil 2.2'de verilmiştir.

1	1	1	1	1
1	1	1	1	1
1	1	1	1	1
1	1	1	1	1
1	1	1	1	1

(a)

1	2	3	4	5
1	2	3	4	5
1	2	3	4	5
1	2	3	4	5
1	2	3	4	5

(b)

1	1	1	1	1
2	2	2	2	2
3	3	3	3	3
4	4	4	4	4
5	5	5	5	5

(c)

Şekil 2.2. Sabit ikili küme gösterimi (a) sabit değerler (b) sabit satırlar (c) sabit sütunlar

Toplamsal ve çarpımsal ikili küme türleri

Aguilar-Ruiz (2005), gen verileri içerisindeki ikili kümeler için sabit ikili kümelerden farklı bir yapı önermiştir. Bu yapıları ikili küme elemanları arasındaki sayısal ilişkiyi inceleyerek tanımlamıştır. Sabit olmayan her bir ikili küme (sütun elemanları aynı kalmak koşuluyla) satır elemanları için parametre değeri π_i olsun. Toplamsal model için ikili kümelerin satır elemanlarına, her bir sütuna β_j katsayısı eklenerek matematiksel model ($b_{ij} = \pi_i + \beta_j$) oluşturulur. İyi bir toplamsal ikili küme yapısı için satırlar arasındaki sayısal ilişki benzer olmalıdır. Çarpımsal model için de benzer durum geçerlidir. Çarpımsal modelin parametresi α_j katsayısı her bir sütun ile çarpılarak model ($b_{ij} = \pi_i \times \alpha_j$) oluşturulur. Her iki modelin ikili küme içerisinde aynı anda bulunması durumunda ise sütun elemanlarına hem toplamsal modeldeki β_j katsayısı eklenir hem de çarpımsal modeldeki α_j katsayısı çarpılır. Bu durumda toplamsal ve çarpımsal ikili küme modeli (

$b_{ij} = \pi_i \times \alpha_j + \beta_j$) oluşturulur. Toplamsal ve çarpımsal ikili kümelerin örnek gösterimi Şekil 2.3'de gösterilmiştir.

1	2	0	3	2
2	3	1	4	3
3	4	2	5	4
4	5	3	6	5
5	6	4	7	6

(a)

1	2	6	3	9
2	4	12	6	18
3	6	18	9	27
4	8	24	12	36
5	10	30	15	45

(b)

1	3	7	5.3	15
2	5	13	7.3	21
3	7	19	9.3	27
4	9	25	11	32
5	11	31	13	38

(c)

Şekil 2.3. Toplamsal ve çarpımsal ikili küme gösterimi (a) toplamsal (b) çarpımsal (c) toplamsal ve çarpımsal

Görüldüğü gibi gen açıklama verileri farklı yapılarda ikili küme türleri barındırabilmektedir. Bu nedenle farklı ikili küme algoritmaları geliştirilmiştir. En yaygın kullanılan ikili kümeleme algoritmaları; Cheng ve Church (CC), Bimax, Plaid, Quest, Spectral ve Xmotif olarak verilebilir. Bu yöntemler aşağıda anlatılmıştır.

2.1. CC Algoritması

Cheng ve Church (2000)'a göre; bir ikili kümenin anlamlı olması, ikili kümenin tüm elemanların varyansını temsil eden Ortalama Karesel Artık Değeri (Mean Squared Residue MSR, H)'nin minimum değer almasına bağlıdır:

$$H(I,J) = \frac{1}{|I||J|} \sum_{i \in I, j \in J} (b_{ij} - b_{iJ} - b_{Ij} + b_{IJ})^2 \tag{2.4}$$

İkili kümedeki b_{ij} elemanı için artık değeri $b_{ij} - b_{iJ} - b_{Ij} + b_{IJ}$ formülü ile hesaplanmaktadır. Bir ikili kümenin anlamlılığı, küme içerisindeki elemanların birbirine olan benzerliğini hesaplayan *H* değeri ile belirlenir. *H* değerinin küçük olması küme elemanlarının benzerliğinin yüksek, *H* değerinin büyük olması ise küme elemanların benzerliğinin düşük olduğunu ifade eder. CC algoritmasında bir δ sabiti belirlenerek *H* değeri bu sabitle karşılaştırılır. *H* değeri δ sabitinden küçük olduğunda elde edilen ikili küme δ-ikili küme (δ-bicluster) olarak ifade edilir. CC algoritmasının temel amacı, H skoru minimum olan maksimum büyüklükte ikili kümeler bulmaktır.

CC algoritmasında veri matrisi tek bir ikili küme olarak işleme başlanır ve satır sütun silme işlemleriyle devam edilir. Satır ve sütunların silinme işlemlerinden sonra, rastgele bir satırın veya sütunun eklenmesi ile H değeri, δ değerini geçmeyecek şekilde artırılır. H değeri δ değerine yaklaşıncaya kadar satır veya sütunlar eklenir. Algoritmada ilk ikili küme belirlenir. Daha sonra veri kümesinden oluşturulan ilk ikili küme çıkarılır. Veri matrisi güncellenerek ikinci iterasyona geçilir. Algoritmanın adımları Şekil 2.4'de verilmiştir.

Girdiler:
A : Veri matrisi
δ : En büyük kabul edilebilir ortalama kare artığı skoru (maximum acceptable mean squared residue score), $\delta \geq 0$
α : Çoklu satır silme için eşik değer, $\alpha > 0$
Adımlar:

1. $A_{IJ} = A$
2. a_{iJ} , a_{Ij} ve a_{IJ} değerlerini hesapla
3. $H(I,J)$ değerini hesapla.
4. Eğer $H(I,J) \leq \delta$ ise A_{IJ} matrisini ikili küme olarak al aksi halde sonraki adıma git
5. Aşağıdaki koşulu sağlayan $i \in I$ satırlarını sil
   $$\frac{1}{|J|}\sum_{j\in J}(a_{ij} - a_{iJ} - a_{Ij} + a_{IJ})^2 > \alpha H(I,J)$$
6. a_{iJ} , a_{Ij} ve a_{IJ} ve $H(I,J)$ değerlerini hesapla
7. Aşağıdaki koşulu sağlayan $j \in J$ sütunlarını sil
   $$\frac{1}{|I|}\sum_{i\in I}(a_{ij} - a_{iJ} - a_{Ij} + a_{IJ})^2 > \alpha H(I,J)$$
8. Eğer tüm satır ve sütun silme işlemi bittiyse sonlandır aksi halde A_{IJ} matrisini A'dan çıkarıp geriye kalan matris için adım 1'e dön.

Şekil 2.4. CC algoritmasının adımları

2.2. Bimax Algoritması

Bimax algoritması, iki değer alan veri matrislerinde ikili kümelerin bulunması için etkili bir algoritmadır. 0 ve 1 değerlerinden oluşan veri matrislerinde 1'lerden oluşan alt matrislerin aranmasına dayalı olan Bimax algoritması hızlı ve basit bir yaklaşımdır (Rodriguez-Baena ve diğerleri, 2011). Prelic ve diğerleri (2006) tarafından temeli oluşturulan Bimax ikili kümeleme algoritması iki değer alan verilerde iyi performans göstermektedir. Gen açıklama verilerinde belli bir eşik değerine göre iki değer alan veri matrislerinin oluşturulmasıyla Bimax yöntemi kullanılabilmektedir. Bu yöntemin diğer algoritmalara göre önemli bir avantajı, aykırı değerlerden etkilenmemesidir. Algoritmanın işleyişi Şekil 2.5'de verilmiştir.

Girdiler:
E_{MxN}: MxN boyutlu veri matrisi
Adımlar:
1. E matrisinden örtüşen U ve V alt matrisleri elde etmek için rastgele bir m satır elemanı seç.
2. Sütunlar kümesinde ilk N_U ve N_V iki alt kümesi oluştur.
3. Sadece N_U alt kümesinde bulunan sütunlarla uyumlu satır elemanlarını M_U alt kümesine, N_U ve N_V alt kümelerinin örtüşen kısımlarında bulunan satır elemanlarını M_W alt kümesine ve sadece N_V alt kümesinde bulunan sütunlarla uyumlu satır elemanlarını M_V alt kümesine böl.
4. Satırlardan elde edilen M_U, M_W ve M_V alt kümelere göre uyumlu sütun elemanlarında bulunan N_U ve N_V alt kümelerini güncelle.
5. U ve V alt matrisleri birbirinden tamamen ayrışana kadar bölme işlemini tekrarla.

Şekil 2.5. Bimax algoritmasının adımları

Veri matrisi önce sütun kümesi içerisinde 1 ve 0 değerlerini ayıracak şekilde iki alt kümeye bölünür. Referans olarak herhangi bir satır elemanı ele alınarak algoritmanın ilk adımı gerçekleşir. Algoritmanın ikinci adımında aynı yöntem satır kümesi için uygulanır. Satır kümesinde 1 değerine sahip elemanlar yer değiştirir. Algoritmanın her aşamasında satır ve sütun elemanlarının sıraları yeniden düzenlenerek alt matris oluşturulmaya başlanır. Algoritma en büyük alt matrisi oluşturana kadar bu işlemleri tekrarlar. Veri matrisinde bölünme işlemi gerçekleştikten sonra 0 değerine sahip satır ve sütun elemanları silinerek sadece 1 değerine sahip ikili küme oluşturulur.

2.3. Plaid Algoritması

Bu model, ikili kümeleme için ideal olarak veri matrisinin temel köşegenleri boyunca alt grupları oluşturur. Bunlar, kümelerin üst üste gelmesine olanak tanıyan iki taraflı küme analizi biçimidir. Örtüşen modeller, iki taraflı kümeler içinde iki yönlü ANOVA modelleri de içerir (Lazzeroni ve Owen, 2002). Bu algoritma veri matrislerindeki katmanların toplamını modellemektedir ve hataların minimum yapılmasına dayalıdır (Turner ve diğerleri, 2005). Plaid modelin varsayımları şu şekilde verilebilir.

- Bir gen birden fazla kümeye ait olabilir.
- Her bir gen kümesi, bazı örneklere (ancak hepsi değil) göre tanımlanmıştır.
- Genler küme örnekleri ve küme genlerini aynı modelde örnekler.

Plaid model algoritmasında satır ve sütunları bir araya getirdikten sonra büyük veri matrisi içerisinde dikdörtgensel bölge oluşturularak bir alt matris şeklinde kümeleme (blok) yapılır. Her blok kendi içerisinde birbirine yakın değerlere (grafiklerde ise renklere) sahip

olur. Cebirsel olarak bloklar içerisindeki genler şu şekilde ifade edilebilir:

$$Y_{ij} = \mu_0 + \sum_{k=1}^{K} \mu_k r_{ik} c_{jk} \tag{2.5}$$

Eşitlik 2.5'te μ_0 arkaplan rengi, μ_k ise k bloğunda bulunan rengi temsil eder. Eğer k bloğunda i. gen mevcutsa r_{ik} değeri 1, aksi takdirde 0 olur. Benzer şekilde j. örnek k bloğu içerisinde mevcutsa c_{jk} değeri 1, aksi takdirde 0 olur (Lazzeroni ve Owen, 2002). Eğer veride bulunan her gen bir küme içerisinde bulunuyorsa ya da her örnek bir küme içerisinde bulunuyorsa Eşitlik 2.6'daki kısıtlar sağlanır:

$$\sum_{k=1}^{K} r_{ik} = 1 \text{ ve } \sum_{k=1}^{K} c_{ik} = 1 \tag{2.6}$$

Bazı gerçek veri setlerinde ise bir gen birden fazla küme içerisinde bulunabilir. Böyle bir durumda kümeler arasında örtüşme olur. Örtüşme olan kümeler için Eşitlik 2.6'daki kısıtlar yerine Eşitlik 2.7'daki kısıtlar dahil edilir:

$$\sum_{k=1}^{K} r_{ik} \geq 2 \text{ ve } \sum_{k=1}^{K} c_{jk} \geq 2 \tag{2.7}$$

Eşitlik 2.5'deki μ_k değeri ikili küme için veri seti içerisindeki tüm satır elemanlarının ve tüm sütun elemanlarının olduğu durumlarda kullanılan doğrusal model denklemidir. Gen ifade veri matrislerinde ise genler kümesi içerisinde bazı örnekler kümesi etkisini gösterir. Bunun tersi olarak örnekler kümesi içerisinde de genlerin bir kısmı etkili olmaktadır. Bu etkileşimi göstermek için doğrusal denklemde bazı parametrelerin eklenmesi gerekir.

$$Y_{ij} = \mu_0 + \sum_{k=1}^{K} (\mu_k + a_{ik}) r_{ik} c_{jk} \tag{2.8}$$

$$Y_{ij} = \mu_0 + \sum_{k=1}^{K} (\mu_k + b_{jk}) r_{ik} c_{jk} \tag{2.9}$$

$$Y_{ij} = \mu_0 + \sum_{k=1}^{K} (\mu_k + a_{ik} + b_{jk}) r_{ik} c_{jk} \tag{2.10}$$

Eşitlik 2.8'de a_{ik} parametresi ikili küme için oluşturulan katmanda satırların etkisini, Eşitlik 2.9'da b_{jk} parametresi ise sütunların etkisini göstermektedir. Plaid model olarak tanımlanın doğrusal denklemde ise Eşitlik 2.10'de gösterildiği gibi hem satırların hem de sütunların etkisini belirten a_{ik} ve b_{jk} parametreleri kullanılmıştır. Bu parametreler k katmanının satır ya da sütun değerleri için etkisini hesaplar. Eğer $|\mu_k+a_{ik}|$ değeri büyükse satır elemanlarının etkisi, $|\mu_k+b_{jk}|$ değeri büyükse sütun elemanlarının etkisi daha fazla olur. Bu modelin çözümü için sezgisel iteratif tabanlı algoritmalar geliştirmiştir (Busygin ve diğerleri, 2008).

Girdiler:
A : Veri matrisi
μ : Arkaplan ve blok etkisi
a_{ik} : Satır etkisi
b_{ik} : Sütun etkisi
r_{ik} : Satırlar için gösterge değişkeni, r_{ik} = 0,1
c_{jk} : Sütunlar için gösterge değişkeni, c_{ik} = 0,1
Adımlar:

1. A matrisinden elde edilen kalıntı matrisini (Z) hesapla.
2. Başlangıç değeri olarak r_{ik}^{0} ve c_{jk}^{0} değerlerini 0 olarak ayarla. $r_{ik}^{(n-1)}$ ve $c_{jk}^{(n-1)}$ değerlerine göre sıradan en küçük kareler yöntemini kullanarak $\mu^{(n)}$, $a_{ik}^{(n)}$ve $b_{ik}^{(n)}$ etkilerinin tahmin edicilerini hesapla.
3. $\mu^{(n)}$, $a_{ik}^{(n)}$, $b_{ik}^{(n)}$ ve $c_{jk}^{(n-1)}$ etkilerini tahmin edicileri ile sıradan en küçük kareler yöntemini kullanarak $r_{ik}^{(n)}$ etkisini hesapla.
4. $\mu^{(n)}$, $a_{ik}^{(n)}$, $b_{ik}^{(n)}$ ve $r_{jk}^{(n-1)}$ etkilerini tahmin edicileri ile sıradan en küçük kareler yöntemini kullanarak $c_{ik}^{(n)}$ etkisini hesapla.
5. Eğer 0.5+n/2(N-T) değeri 0.5'ten büyükse $r_{ik}^{(n)}$ ve $c_{ik}^{(n)}$ etkilerinden herhangi birini ekle.
6. Eğer 0.5-n/2(N-T) değeri 0.5'ten küçükse $r_{ik}^{(n)}$ ve $c_{ik}^{(n)}$ etkilerinden herhangi birini ekle aksi halde üçüncü adıma dön.
7. $\mu^{(N+1)}$, $a_{ik}^{(N+1)}$ ve $b_{ik}^{(N+1)}$ etkilerini hesapla.
8. Aday katman kareler toplamını her bir satır için hesapla.

$$LSS_{aday} = \sum_{i=1}^{I}\sum_{j=1}^{J}(\mu + a_i + \hat{b}_j)\hat{r}_i\hat{c}_j$$

9. Satır elemanlarındaki maksimum LSS_{aday} değeri için ikili kümeyi kabul et ve adım 2'ye dön.

Şekil 2.6. Plaid algoritmasının adımları

2.4. Xmotif Algoritması

Gen açıklama verilerinde benzer satırları ve sütunları bir arada kümelemek için bu algoritma Xmotif adı verilen alt matrisleri bulmaya çalışmaktadır. Veri matrisinde birçok Xmotif ikili kümeler bulunabilir. İkili kümeleme işleminde Xmotif'leri maksimum büyüklükte olması beklenir. Ancak her gen veya her koşul Xmotif'ler içerisinde olmayabilir. En geniş Xmotif'leri bulmak için tekrarlayıcı bir algoritma yapısı vardır. Algoritma, en geniş Xmotif kümesini bulduktan sonra tüm veri setinde bulunan bu kümeyi çıkararak geriye kalan veri matrisinden tekrar bir Xmotif kümesi bularak işleme devam

eder. Tüm alt matrisler bulununcaya kadar algoritmadaki iterasyon devam eder (Murali ve Kasif, 2003). Algoritmanın adımları Şekil 2.7'de verilmiştir.

Girdi:
A_{SxD} : SxD boyutlu sıralı veri matrisi
Adımlar:
1. A matrisinin satır elemanlarından rasgele bir örnek (*c*) seç.
2. A matrisinin sütun elemanlarından rasgele *D* alt kümesi seç.
3. Her bir *g* geni için, eğer *g*, *D* alt kümesi içindeki tüm örnekler ve *c*'de *s* durumunda ise (*g*,*s*) çiftini G_{ij} kümesine ekle.
4. *Cij'*yi *Gij* kümesindeki tüm gen ifadelerinde *c* ile uyumlu örnek kümesi olarak al.
5. Eğer C_{ij} *axn* boyutlu örnekten daha küçük ise (C_{ij} , G_{ij}) çiftini çıkar.
6. $|G_{ij}|$ maksimum olana kadar (*C* , *G*) motifini çalıştır.

Şekil 2.7. Xmotif algoritmasının adımları

Xmotif ikili küme yönteminde birçok faydalı özellik bulunmaktadır. Bu yaklaşımda bir gen birden fazla Xmotif içinde bulunabilir. Model içerisinde gen açıklama seviyesi çoklu durumlar için düzenlenebilir. Örneğin bulunan bir Xmotif ikili kümesinin elemanına ait olan bir durumun başka bir ikili küme içerisine dâhil edilmesine izin verilebilir.

2.5. Quest Algoritması

Murali ve Kasif (2003) tarafından temeli oluşturulan bu algoritma ilk olarak aynı sorulara benzer veya aynı cevapları veren bireylerin soru gruplarına göre kümelenmesi için geliştirilmiştir. Bu algoritma, değerler sınıflama (nominal) ölçeğinde verilmişse Xmotif algoritmasına benzer şekilde çalışır. Algoritma, sıralı (ordinal) ölçekler için önceden belirlenmiş bir parametreyle (*d*) ayarlanmış bir boyut aralığında benzer cevaplar arar. Bunun için belirlenen aralık, seçilen yanıtlayıcı değişken için başlangıç değerinden parametre değeri kadar daha düşük ve daha yüksek olacak şekilde belirlenir. Sürekli değişkenler için ise seçilen dağılımın çeyreklik değerlerine göre belirlenir. Yaygın olarak kullanılan normal skorlar olduğu için dağılım da normal dağılım olarak kullanılır (Kaiser,2011).

Girdi:
A_{SxD} : SxD boyutlu veri matrisi
Adımlar:
1. A matrisinin satır elemanlarından rasgele bir örnek (*c*) seç.
2. A matrisinin sütun elemanlarından rasgele *D* alt kümesi seç.
3. Her bir *g* geni için, Eğer *g, D* alt kümesi içindeki tüm örnekler ve *c*'de *s* durumunda ise (*g*,*s*) çiftini G_{ij} kümesine ekle.
4. C_{ij}'yi G_{ij} kümesindeki tüm gen ifadelerinde *c* ile uyumlu örnek kümesi olarak al.
5. Eğer C_{ij} *axn* boyutlu örnekten daha küçük ise (C_{ij} , G_{ij}) çiftini çıkar.
6. $|G_{ij}|$ maksimum olana kadar (*C* , *G*) motifini çalıştır.

Şekil 2.8. Quest algoritmasının adımları

2.6. Spectral Algoritması

Kluger ve diğerleri (2003) tarafından geliştirilen bu algoritma gen ve koşulların bir alt kümesini dama tahtası yapısına benzer olarak tanımlamaya çalışır. Dama tahtası yapısı veri matrisindeki sabit değerli ikili kümelerin bir kombinasyonu olarak tanımlanabilir.

Spectral ikili kümeleme algoritması, gen açıklama verilerindeki yapının tek değer ayrıştırmasına göre ele alınmasını varsayar. Örneğin dört bloktan oluşan 4×6 boyutlu bir A matrisi olsun. Her blok ca, cb, da, db gibi sabit değerlere sahip olsun. $u = (c, c, d, d)$ ve $v = (a, a, a, b, b, b)$ vektörleri sırasıyla sınıflandırma öz genini ve öz diziyi göstersin. A matrisinin öz gen ve öz dizilerin çarpımı ($A = u^t v$) sonucunda matematiksel model Eşitlik 2.11'de gösterilmiştir.

$$A = \begin{pmatrix} ca & ca & ca & cb & cb & cb \\ ca & ca & ca & cb & cb & cb \\ da & da & da & db & db & db \\ da & da & da & db & db & db \end{pmatrix} = \begin{pmatrix} c \\ c \\ d \\ d \end{pmatrix} \begin{pmatrix} a & a & a & b & b & b \end{pmatrix} \tag{2.11}$$

Spectral ikili kümelemede, normalleştirilmiş gen açıklama verisinin tekil değer ayrıştırması Eşitlik 2.12'de gösterilen iki farklı özvektör probleminin çözümü ile sağlanabilir.

$$A^T A\omega = \lambda\omega \text{ ve } AA^T z = \lambda z \tag{2.12}$$

Burada ω ve z sırasıyla öz gen ve öz dizi değerleridir. Eğer normalleştirilmiş veri dama

tahtası yapısında ise aynı öz değere (λ) sahip en az bir çift sabit öz gen ve öz dizi vardır. Algoritmanın adımları Şekil 2.9'da verilmiştir.

Girdi:
A_{NxM} : NxM boyutlu veri matrisi
Adımlar:
1. Bağımsız satır ve sütunların ölçeklenmesi (Independent Rescaling of Rows and Columns, IRRC), ikili rasgelelik ya da logaritmik etkileşim yöntemleri kullanarak normalleştirilmiş gen açıklama matrisi (A) üret.
2. Normalleştirilmiş matrise tek değer ayrıştırması uygula. Eğer normalizasyon yöntemi olarak IRRC ya da ikili rasgelelik kullanıldıysa ortak en geniş öz değere sahip satır ve sütun çiftini çıkar.
3. Seçilen her satır ve sütun öz vektör çifti için tek boyutlu k-means kümeleme yöntemi ya da seçilen öz vektörlere normalleştirilmiş verinin yansımasını uygula.

Çıktı:
En iyi uygun adımdaki satır ve sütun öz vektörler çiftinin blok matris sonucu olarak rapor et.

Şekil 2.9. Spectral algoritmasının adımları

3. İKİLİ KÜME DEĞERLENDİRME ÖLÇÜLERİ

Önceki bölümde, ikili kümeleme algoritmalarından bazıları tanıtılmıştı. Bu algoritmaların hangilerinin hangi durumda daha iyi olduğunu tespit edebilmek için elde edilen ikili kümelerin karşılaştırılması gerekir. Bu amaçla geliştirilen ölçülere ikili küme değerlendirme ölçüleri adı verilir. Bu ölçülerden bazıları burada verilmiş ve sayısal bir örnek üzerinden elde hesaplanışları gösterilmeye çalışılmıştır.

3.1. Varyans Ölçüsü

Hartigan (1972) tarafından ikili kümeleri değerlendirme ölçüsü olarak küme içi varyans kullanılmıştır. İkili kümelerin varyanslarının toplamına dayalı olan bu ölçü Eşitlik 3.1'de verilmiştir.

$$VAR(B) = \sum_{i=1}^{|I|} \sum_{j=1}^{|J|} (b_{ij} - b_{IJ})^2 \tag{3.1}$$

Bu ölçü, sabit elemanlı ikili kümelerin tespit edilmesinde etkili olmaktadır (Pontes ve diğerleri, 2015).

3.2. Ortalama Karesel Artık Skoru

İkili kümelerin kalitesini ölçmek için kullanılan ortalama karesel artık skoru (MSR) yaygın kullanılan bir yaklaşımdır (Cheng ve Church, 2000). Bu ölçü Eşitlik 3.2'de verilmiştir.

$$MSR(B) = \frac{1}{|I||J|} \sum_{i=1}^{|I|} \sum_{j=1}^{|J|} \left(b_{ij} - b_{iJ} - b_{Ij} + b_{IJ} \right)^2 \tag{3.2}$$

Bu ölçü ikili kümede yer alan satır ve sütun değerlerinin tutarlılığını incelemektedir. Bu değer küçüldükçe ikili kümelemenin başarısı artar. Eğer bu değer δ (delta) gibi bir değerden küçükse bu ikili kümeye δ-ikili küme adı verilmektedir. Bu ölçü veri matrisindeki değerlerin ölçeklendirilmesine aşırı duyarlıdır ve bu durumda ikili kümenin değerlendirmesinde yetersiz olduğu görülmüştür (Aguilar-Ruiz, 2005). Önceki bölümde

anlatılan çarpımsal ikili küme durumunda, bu ölçünün parametre değerine bağlı olduğu ve varyansın yüksek çıkmasından dolayı ikili kümeleri değerlendirmede yetersiz olduğu ve toplamsal ikili küme durumunda kullanılması gerektiği sonucuna ulaşılmıştır.

3.3. Uygunluk İndeksi

Yip ve diğerleri (2004) tarafından önerilen uygunluk indeksi (UI) ölçüsü ikili kümenin sütunlarının uygunluğunu inceler. Sütun elemanlarının uygunluklarının toplamına dayalı olan bu ölçü özellikle sabit değerli ikili kümelerde iyi sonuçlar vermektedir. Her bir sütun değeri için uygunluk indeksi Eşitlik 3.3'de verilmiştir.

$$R_{Ij} = 1 - \frac{\sigma_{Ij}^2}{\sigma_j^2} \tag{3.3}$$

Burada σ_{Ij}^2 (yerel varyans) ikili kümedeki sütun değerlerinin varyansı ve σ_j^2 (genel varyans) tüm veri setindeki sütun değerlerinin varyansıdır. Yerel varyans küçüldükçe indeks değeri büyür. Olabilecek en yüksek uygunluk indeksi yerel varyansın sıfır olması durumunda gerçekleşir.

3.4. Ölçeklenen Ortalama Karesel Artık Skoru

Mukhopadhyay ve diğerleri (2009) tarafından geliştirilen ölçeklenen ortalama karesel artık skoru (SMSR) çarpımsal ikili kümeler için uygun bir değerlendirme ölçüsüdür. MSR'nin çarpımsal ikili kümeler için ortaya çıkan dezavantajını gidermek için geliştirilmiştir. Bu ölçü Eşitlik 3.4'de verilmiştir.

$$SMSR(B) = \frac{1}{|I||J|} \sum_{i \in I, j \in J} \frac{(b_{iJ} \times b_{Ij} - b_{ij} \times b_{IJ})^2}{b_{iJ}^2 \times b_{Ij}^2} \tag{3.4}$$

Bu ölçü toplamsal ikili küme türlerinde etkin değildir.

3.5. Chia ve Karuturi Ölçüsü

Chia ve Karuturi (2010) tarafından önerilen bu ölçü ikili kümelerin doğruluğunu, bir veri matrisi içerisinde sütun elemanlarından ikili küme içerisinde olanlar ile olmayanlar arasındaki benzerliğe bakarak değerlendirir. İkili kümedeki satır elemanlarının sütun elemanları ile etkileşimi incelenirken MSR skoru kullanılır. Bu skor kullanılarak ikili kümelerin satır etkisi $T_k(b)$ ve sütun etkisi $B_k(b)$ belirlenir. Satır ve sütun etkisi sırasıyla Eşitlik 3.5 ve Eşitlik 3.6 ile hesaplanır.

$$T_k(b) = \frac{1}{|I(b)|} \sum_{i=1, i \in b}^{I(b)} {b_{iJk}}^2 - \frac{MSR_k(b)}{J_k(b)} \tag{3.5}$$

$$B_k(b) = \frac{1}{|J_k(b)|} \sum_{j=1, j \in b}^{J_k(b)} {b_{Ijk}}^2 - \frac{MSR_k(b)}{I(b)} \tag{3.6}$$

Burada *k* alt indisi, ikili küme içerisinde olan elemanlar için 1, olmayanlar için 2 değerini alır. $I(b)$ değeri, ikili küme içerisindeki satır sayısını, $J_k(b)$ değeri ise ikili küme içindeki ve dışındaki sütun sayısını gösterir. Satır ve sütun etkileri için ikili kümedeki elemanlar ile diğerleri için ayrı ayrı hesaplanması gerekir.

Hesaplanan satır ve sütun etkileri kullanılarak Chia ve Karuturi uygunluk skoru (CKSB) aşağıdaki gibi hesaplanır.

$$CKSB(b) = LOG\left(\frac{\max(T_1(b)+\alpha, B_1(b)+\alpha)}{\max(T_2(b)+\alpha, B_2(b)+\alpha)}\right) \tag{3.7}$$

Burada α değeri 0 ile 1 arasında belirlenen bir ölçek faktörüdür. Çok küçük ikili kümelerin satır ve sütun etkisini dengelemek için kullanılır. CKSB skoru ne kadar yüksekse ikili kümeler o kadar anlamlıdır.

3.6. Korelasyon Tabanlı Ölçüler

Korelasyon katsayısı geleneksel kümeleme yöntemlerinde olduğu gibi farklı mikrodizi verilerinin bulunduğu analiz türlerinde de yoğun olarak kullanılmaktadır. Gen ifade

verilerinin bulunduğu mikrodizi veri setinin büyüklüğüne bakılmaksızın genler arasındaki benzerlik ölçüsünü korelasyon katsayısı ile belirlemek mümkündür. Ayrıca genler arasında pozitif yönlü korelasyon kadar negatif yönlü korelasyon da bulunabilir. İkili kümeler için değerlendirme ölçülerinden korelasyon tabanlı olanlar; Pearson korelasyon katsayısı (PCC), Alt matris korelasyon skoru (SCS), ortalama korelasyon değeri (ACV), Spearman ortalama Rho değeri (ASR) ve Spearman ikili küme ölçüsü (SBM) olarak sıralanabilir (Pontes ve diğerleri, 2015).

3.6.1. Ortalama korelasyon ölçüsü

İki değişken arasındaki Pearson korelasyon katsayısı, iki değişkenin kovaryansının standart sapmalarının çarpımlarına bölünmesi ile elde edilir. Bu değer -1 ile +1 arasında değer alır ve mutlak değerce 1'e yaklaştığında ilişkinin güçlendiğini 0'a yaklaştığında ise ilişkinin zayıfladığını ifade eder. Bu değerin negatif çıkması ters yönlü, pozitif çıkması ise aynı yönlü ilişki olduğunu göstermektedir. Burada dikkat edilmesi gereken nokta bu ölçünün doğrusal ilişkiyi belirlemede kullanılmasıdır. i_1 ve i_2 ikili küme içerisindeki herhangi iki satırı göstermek üzere bu katsayı Eşitlik 3.8' ile hesaplanır.

$$r(i_1, i_2) = \frac{\sum_{j=1}^{|J|} (b_{i_1 j} - b_{i_1 J})(b_{i_2 j} - b_{i_2 J})}{\sqrt{\sum_{j=1}^{|J|} (b_{i_1 j} - b_{i_1 J})^2 \sum_{j=1}^{|J|} (b_{i_2 j} - b_{i_2 J})^2}} \tag{3.8}$$

Burada $b_{i_1 j}$ ve $b_{i_2 j}$ değerleri, j sütununda bulunan i_1 ve i_2 satır eleman değerlerini, $b_{i_1 J}$ ve $b_{i_2 J}$ değerleri ise i_1 *ve* i_2 satır elemanlarının ortalamasını göstermektedir. Bu katsayı her bir satır (gen) çifti için hesaplanır ve Eşitlik 3.9'da verilen ortalama korelasyon ölçüsü elde edilir.

$$AC(B) = \frac{\sum_{i_1=1}^{|I|-1} \sum_{i_2=i_1+1}^{|I|} r(i_1, i_2)}{\binom{|I|}{2}} \tag{3.9}$$

Burada korelasyon değeri sadece satırlar arasındaki ilişkiyi ölçtüğü için sütunlar arasındaki ilişkiyi dikkate almamaktadır. Bu problemi ortadan kaldırmak için Yang ve diğerleri (2011) ve Teng ve Chan (2008), çalışmalarında sütunlar için de korelasyon değerini kullanmıştır.

Bir başka ortalama korelasyon ölçüsü Teng ve Chan (2008) tarafından Eşitlik 3.10'daki şekilde önerilmiştir.

$$ACV(B) = \max\left\{ \frac{\sum_{i_1=1}^{|I|}\sum_{i_2=1}^{|I|}|r(i_1,i_2)|-|I|}{|I|^2-|I|}, \frac{\sum_{j_1=1}^{|J|}\sum_{j_2=1}^{|J|}|r(j_1,j_2)|-|J|}{|J|^2-|J|} \right\} \quad (3.10)$$

Burada $r(i_1,i_2)$ değeri, herhangi i_1 ve i_2 satır çiftleri arasındaki korelasyon değerini, $r(j_1,j_2)$ değeri herhangi j_1 ve j_2 sütun çiftleri arasındaki korelasyon değerini göstermektedir.

Bu ölçü ise 0 ile 1 arasında değer alır. 1'e yaklaştıkça ikili kümenin kalitesi artar.

3.6.2. Alt matris korelasyon skoru

Yang ve diğerleri (2011), hem satırlar hem de sütunlar üzerinden bir değerlendirme ölçüsü geliştirmek amacıyla korelasyon katsayısını kullanmıştır. Satırlar ve sütunlar için hesaplanan korelasyon katsayıları sırasıyla Eşitlik 3.11 ve Eşitlik 3.12'deki gibi tanımlanır.

$$S_{satır} = \min_{i_1 \in I}(S_{i_1 J}), \qquad S_{i_1 J} = 1 - \frac{\sum_{i_2 \neq i_1, i_2 \in I} \left| r(x_{i_1 J}, x_{i_2 J}) \right|}{|I|-1} \quad (3.11)$$

$$S_{sütun} = \min_{j_1 \in J}(S_{I j_1}), \qquad S_{I j_1} = 1 - \frac{\sum_{j_2 \neq j_1, j_2 \in J} \left| r(x_{I j_1}, x_{I j_2}) \right|}{|J|-1} \quad (3.12)$$

Burada $r(x_{i_1 J}, x_{i_2 J})$ satırlar arasındaki, $r(x_{I j_1}, x_{I j_2})$ de sütunlar arasındaki Pearson korelasyon değerini göstermektedir. $S_{satır}$ skoru, ikili küme içerisindeki satırlar üzerindeki, $S_{sütun}$ skoru ikili küme içerisindeki sütunlar üzerindeki korelasyon derecesini gösterir. Bu değerlerden minimum olanı alınarak alt matris korelasyon ölçüsü elde edilir. Yang ve diğerleri (2011), belirli bir eşik değeri parametresi δ (delta) belirleyerek $S(I,J) < \delta$ koşulunu sağlayan ikili kümelere δ-corbicluster ismini vermiştir.

3.6.3. Ortalama Spearman korelasyon değeri

Ayadi ve diğerleri (2009) tarafından önerilen bu ölçü iki değişkene atanan sıra sayıları arasındaki korelasyonu kullanır. Bu değer Pearson korelasyon katsayısının aksine bir dağılım varsayımı gerektirmemektedir. Bununla beraber Pearson korelasyon katsayısına göre daha az hassas olduğu açıktır. Ortalama Spearman korelasyon değeri (ASR) Eşitlik 3.13'de verilmiştir.

$$ASR(B) = 2 \times \max\left\{ \frac{\sum_{i_1 \in I} \sum_{i_2 \geq i_1+1, i_2 \in I} r_s(i_1, i_2)}{|I|(|I|-1)}, \frac{\sum_{j_1 \in J} \sum_{j_2 \geq j_1+1, j_2 \in J} r_s(j_1, j_2)}{|J|(|J|-1)} \right\} \tag{3.13}$$

Burada $r_s(i_1, i_2)$ değeri iki satır arasındaki, $r_s(j_1, j_2)$ değeri iki sütun arasındaki Spearman korelasyonunu göstermektedir. Eğer ikili küme içerisinde bulunan satır veya sütun elemanları pozitif yönde güçlü bir ilişkide ise ASR +1 değerine yakındır. Bunun tam tersi olarak ASR'nin -1'e yakın olması da negatif yönlü güçlü bir ilişki olduğunu gösterir.

3.6.4. Spearman ikili küme ölçüsü

Flores ve diğerleri (2013) tarafından önerilen bu ölçü, toplamsal ve çarpımsal ikili kümelerin değerlendirilmesinde kullanılır. Bu ölçümün hesaplanması iki aşamadan oluşur. Birinci aşama, Spearman korelasyon katsayılarının hesaplanması, ikinci aşama ise bu katsayıların kullanılmasıyla Eşitlik 3.14'deki değerin elde edilmesidir.

$$SBM(B_{ij}) = \alpha\left(B_{ij}\right) \times r^G_{B_{IJ}} \times \beta\left(B_{ij}\right) \times r^C_{B_{IJ}} \tag{3.14}$$

Burada $r^G_{B_{IJ}}$ ve $r^C_{B_{IJ}}$ değerleri ikili küme içerisindeki satırlarda ve sütunlarda gözlenen eğilimlerin ifadesini özetler. Bu değerler Eşitlik 3.15 ve 3.16'daki gibi hesaplanır.

$$r^G_{B_{IJ}} = \frac{2}{|I| \times (|I| - 1)} \sum_{i_1=1}^{|I|} \sum_{i_2=i_1+1}^{|I|} |r_s(i_1, i_2)| \tag{3.15}$$

$$r^C_{B_{IJ}} = \frac{2}{|J| \times (|J| - 1)} \sum_{j_1=1}^{|J|} \sum_{j_2=j_1+1}^{|J|} |r_s(j_1, j_2)| \tag{3.16}$$

Burada $r(i_1, i_2)$ ve $r(j_1, j_2)$ değerleri i_1 ve i_2 satırları ve j_1 ve j_2 sütunları arasındaki Spearman korelasyon katsayısının mutlak değerlerini göstermektedir.

Eşitlik 3.14'de gösterilen $\alpha(B_{ij})$ ve $\beta(B_{ij})$ değerleri güvenilirlik katsayılarını gösterir. Bu değerler hem satırlardaki hem de sütunlardaki gözlenen eğilimlerin etkisini ağırlıklandırır. Korelasyon katsayılarının hesaplanmasında ağırlıklandırma, farklı boyutlarda örnek büyüklüğü kullanıldığı için gereklidir. Güvenilirlik katsayısını belirleyen α ve β parametreleri 0 ile 1 arasında bir değer alır. β katsayısı sütun etkisini belirleyeceği için yüksek güvenilirlikte olması beklendiğinden genellikle varsayılan değeri 1 olur. Benzer şekilde α katsayısı, sütun sayısının belirlenecek bir eşik değeri üzerinde olması durumunda 1 olur. Aksi durumda, ikili kümenin sütun sayısının tüm verideki sütun sayısına oranı olarak belirlenir (Flores ve diğerleri, 2013).

SBM için herhangi bir değer aralığı belirtilmemiştir. İkili kümelerin kalitesini artırmak için SBM'nin maksimize edilmesi gerekir, ancak herhangi bir değer aralığı belirtilmemiştir. Dolayısıyla bu değere bakarak bir ikili kümenin ne kadar iyi olduğu söylenemez. Ancak birden fazla ikili kümeden hangisinin diğerine göre daha iyi olduğunu belirlemede kullanılır.

3.7. Standartlaştırılmış Ölçüler

İkili kümelemede etkili olan önemli bir nokta veri matrisi içerisindeki değerlerin aralığıdır. Mikrodizi verilerde genlere ait değer aralıkları birbirinden farklı olabilir. İkili kümeleme, genlerin sayısal değerlerinden çok eğilimlerini dikkate alır. Bu nedenle genlerle ilgili

hesaplama yapmadan önce verilerin aynı aralıkta olmasını sağlamak için standartlaştırma işlemleri uygulanabilir. Standartlaştırma işlemi Eşitlik 3.17'de verildiği gibi yapılır.

$$\hat{b}_{ij} = \frac{b_{ij} - \mu_{g_i}}{\sigma_{g_i}} \quad , \quad 1 \le i \le |I| \quad , \quad 1 \le j \le |J| \tag{3.17}$$

Burada μ_{g_i} ve σ_{g_i} sırasıyla i inci genin (satır) ortalamasını ve standart sapmasını göstermektedir.

Standartlaştırma iki avantaj sağlar. Bunlardan birincisi tüm satırların benzer aralığa sahip olması ve ikincisi ise tüm koşullarda (sütunlarda) homojenliğin sağlanması ve bu sayede grafiksel gösterimlerin daha anlaşılabilir olmasıdır.

Standartlaştırılmış değerlere dayalı olarak önerilen sanal hata değerlendirme ölçüsü aşağıda verilmiştir.

Sanal hata

Sanal Hata (VE) ölçüsünün temel amacı ikili küme içerisindeki satır elemanlarının genel eğilimlerinin nasıl olduğunu belirlemektir. Bu yöntemde ikili küme içerisine yapay bir satır eklenir. Eklenen bu satırdaki her eleman j inci sütun elemanlarının ortalamasıdır. Her bir sütun elemanı ile eklenen bu satır elemanları arasındaki farklar hesaplanarak VE skoru Eşitlik 3.18'deki gibi bulunur.

$$VE(B) = \frac{1}{|I||J|} \sum_{i=1}^{|I|} \sum_{j=1}^{|J|} \left| \hat{b}_{ij} - b_{Ij} \right| \tag{3.18}$$

İkili kümede satır elemanları ne kadar benzer olursa VE değeri de o kadar küçük olur. Toplamsal ve çarpımsal yapıyı aynı anda barındıran ikili küme yapılarında bu ölçü iyi sonuçlar vermemektedir.

Çizelge 3.1'de değerlendirme ölçülerinin hangi ikili küme yapılarında kullanılması gerektiğini göstermektedir.

Çizelge 3.1. Değerlendirme ölçülerinin kullanıldığı ikili küme yapıları

Ölçü	İkili Küme Yapısı
VAR	Sabit
MSR	Toplamsal
UI	Sabit
SMSR	Çarpımsal
CKSB	Sabit, Toplamsal ve Çarpımsal
ACV	Toplamsal ve Çarpımsal
SCS	Toplamsal ve Çarpımsal
ASR	Toplamsal ve Çarpımsal
SBM	Toplamsal ve Çarpımsal
VE	Toplamsal veya Çarpımsal

3.8. Sayısal Örnekler

3.8.1. İkili küme yapılarının oluşturulması

Değerlendirme ölçülerinin farklı ikili küme yapılarında uygulanmasını göstermek için küçük bir yapay veri seti oluşturulmuştur. Veri seti 10×10'luk matris şeklinde, ortalaması 0 ve varyansı 1 olan normal dağılımdan rasgele üretilmiştir. Veri matrisi Şekil 3.1'de gösterilmiştir.

	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}
x_1	0.49	0.57	0.20	0.20	0.78	-0.55	-2.51	-0.52	-1.04	-0.90
x_2	0.06	0.58	0.67	0.68	0.68	-0.25	2.12	-1.08	-0.31	-0.21
x_3	0.61	0.60	0.77	0.13	0.13	-0.19	-1.50	-0.39	-0.61	-0.98
x_4	0.58	0.05	0.99	0.36	0.36	-1.07	1.24	-0.53	-0.89	0.56
x_5	0.69	0.34	0.19	0.37	0.37	0.15	-0.03	1.07	-0.87	0.12
x_6	-1.04	-0.05	-0.97	0.08	0.08	1.95	0.49	0.67	-0.06	0.14
x_7	-1.88	-3.10	0.09	-0.10	-0.10	-1.30	1.57	0.70	-1.58	2.05
x_8	0.29	-1.15	0.43	-0.58	-0.58	-0.62	-0.98	-1.05	-0.15	-1.24
x_9	-0.75	-0.55	-0.70	-2.36	-2.36	0.05	0.83	0.74	-1.53	0.08
x_{10}	-1.88	-0.10	1.11	-1.07	-1.07	0.62	-0.83	-0.12	-0.88	-1.31

Şekil 3.1. 10×10 boyutlu yapay veri seti

Veri seti içerisine 5×5 boyutlu altı farklı yapıda ikili küme eklenerek yapay veri matrisleri elde edilmiştir. Veri setinde sadece ilk beş satır ve beş sütun elemanları ikili küme yapılarına göre değiştirilmiş, ikili küme elemanı olmayan diğer elemanlar aynı kalmıştır. Altı farklı ikili küme yapıları; sabit, yukarı düzenlenmiş sabit, toplamsal, çarpımsal,

içerisinde hem toplamsal hem de çarpımsal modeli bulunduran ve plaid model ikili kümedir.

Veri seti içerisinde oluşturulan ikili küme yapıları Şekil 3.2'de gösterilmiştir.

	y_1	y_2	y_3	y_4	y_5
x_1	0.00	0.00	0.00	0.00	0.00
x_2	0.00	0.00	0.00	0.00	0.00
x_3	0.00	0.00	0.00	0.00	0.00
x_4	0.00	0.00	0.00	0.00	0.00
x_5	0.00	0.00	0.00	0.00	0.00

(a)

	y_1	y_2	y_3	y_4	y_5
x_1	5.00	5.00	5.00	5.00	5.00
x_2	5.00	5.00	5.00	5.00	5.00
x_3	5.00	5.00	5.00	5.00	5.00
x_4	5.00	5.00	5.00	5.00	5.00
x_5	5.00	5.00	5.00	5.00	5.00

(b)

Şekil 3.2. Sabit ikili küme yapıları (a) ortalama değerli ikili küme (b) yukarı düzenlenmiş ikili küme

Sabit ikili küme yapılarından veri setinin ortalama değerini ve maksimum değerinden daha büyük sabit bir değer alan ikili kümeler Şekil 3.3'de gösterilmiştir. Veri matrisi ortalaması 0 ve varyansı 1 olan normal dağılım olduğu için sabit ikili kümelerin elemanlarından birincisi için $\pi = 0$, ikincisi için $\pi = 5$ alınmıştır.

	y_1	y_2	y_3	y_4	y_5
x_1	-0.17	-0.26	-0.95	-1.53	0.26
x_2	-0.04	-0.69	-0.71	-1.72	0.12
x_3	-0.12	-0.41	-0.86	-1.60	0.21
x_4	-0.31	0.23	-1.22	-1.32	0.42
x_5	-0.17	-0.23	-0.96	-1.52	0.27

(a)

	y_1	y_2	y_3	y_4	y_5
x_1	-0.37	-0.34	-1.18	-1.67	0.11
x_2	-1.19	-1.17	-2.00	-2.49	-0.71
x_3	-0.66	-0.63	-1.47	-1.96	-0.17
x_4	0.56	0.59	-0.25	-0.74	1.04
x_5	-0.32	-0.29	-1.12	-1.62	0.17

(b)

	y_1	y_2	y_3	y_4	y_5
x_1	0.03	-0.09	0.05	-0.04	-0.03
x_2	0.15	-0.53	0.29	-0.23	-0.17
x_3	0.07	-0.25	0.13	-0.11	-0.08
x_4	-0.11	0.40	-0.22	0.17	0.13
x_5	0.02	-0.07	0.04	-0.03	-0.02

(c)

Şekil 3.3. Toplamsal ve çarpımsal ikili küme yapıları (a) toplamsal ve çarpımsal modelin aynı anda bulunduğu ikili küme (b) toplamsal ikili küme (c) çarpımsal ikili küme

Toplamsal ve çarpımsal ikili küme oluşturulurken $\alpha_{5\times1}$ ve $\beta_{5\times1}$ katsayı matrisleri ortalaması 0 ve varyansı 1 olan standart normal dağılımdan rasgele olarak üretilmiştir. Sabit olmayan ikili küme elemanlarına oluşturulan bu katsayı matrisleri eklenip çarpılmıştır. Böylece hem toplamsal hem de çarpımsal modeli içinde bulunduran ikili küme yapısı oluşturulmuştur. Toplamsal ve çarpımsal modelin ayrı ayrı bulunduğu ikili küme yapılarını oluşturmak için matematiksel modeldeki ($b_{ij} = \pi_i \times \alpha_j + \beta_j$) parametrelerde

bazı değişiklikler yapılmıştır. Toplamsal model oluşturmak için modeldeki α_j katsayı matrisi elemanları 1 olarak alınmıştır. Böylece toplamsal model ($b_{ij} = \pi_i + \beta_j$) elde edilmiştir. Çarpımsal modeli elde etmek için β_j katsayı matrisi, elemanları 0 olarak alınmış ve çarpımsal model ($b_{ij} = \pi_i \times \alpha_j$) elde edilmiştir.

	y_1	y_2	y_3	y_4	y_5
x_1	-1.53	-3.18	-2.11	0.33	-1.42
x_2	-3.18	-4.82	-3.75	-1.31	-3.07
x_3	-2.11	-3.75	-2.69	-0.25	-2.00
x_4	0.33	-1.31	-0.25	2.19	0.44
x_5	-1.42	-3.07	-2.00	0.44	-1.32

Şekil 3.4. Plaid model ikili küme yapısı

İkili küme yapılarından biri olan Plaid model ikili kümenin elemanları ise yine benzer şekilde ortalaması 0 varyansı 1 olan standart normal dağılımdan üretilen arkaplan etkisi matrisi, küme etkisi matrisi, satır ve sütun etkisi matrisi eklenerek oluşturulmuştur.

3.8.2. İkili küme yapıları için hesaplanan değerlendirme ölçüleri

6 farklı ikili küme yapısı için 11 farklı değerlendirme ölçüsü kullanılmıştır. Her bir ikili küme yapısı için değerlendirme ölçüleri hesaplanmış ve örnek olarak toplamsal ve çarpımsal yapıdaki ikili kümenin değerlendirme ölçüleri hesaplamalarının uygulamalı gösterimi aşağıda verilmiştir. Daha sonra diğer ikili küme yapılarının tamamı R Project dilinde R fonksiyonları oluşturularak hesaplanmıştır.

Toplamsal ve çarpımsal ikili kümenin değerlendirme ölçülerinin hesaplanması

VAR Skoru

Toplamsal ve çarpımsal ikili küme yapısının VAR skoru hesaplanması için öncelikle ikili küme elemanlarının ortalaması bulunur. Daha sonra her bir ikili küme elemanlarının ortalama değerden farkı alınarak kareleri toplanır. Böylece ikili kümenin VAR skoru hesaplanmış olur.

$$\overline{b_{ij}} = -0.5316$$

$$VAR(B) = \sum_{i=1}^{|5|}\sum_{j=1}^{|5|}(b_{ij} - (-0.5316))^2 = 10.7893$$

MSR Skoru

İkili kümenin satır ve sütun ortalamaları bulunur. İlk sütunda bulunan satırların ortalaması örnek olarak aşağıda gösterilmiştir.

$$b_{I1} = \frac{1}{|I|}\sum_{i=1}^{|5|} b_{i1} = \frac{1}{|5|} \times \left[(-0.1666) + (-0.0409) + (-0.1224) + (-0.3089) + (-0.1748)\right] = -0.1627$$

Hesaplanan ilk sütun ortalamasına benzer olarak diğer sütunlardaki ortalama değerler de bulunur.

$$b_{Ij} = \frac{1}{|I|}\sum_{i=1}^{|I|} b_{ij} = \begin{bmatrix} -0.1627 & -0.2730 & -0.9405 & -1.5389 & 0.2571 \end{bmatrix}$$

İlk satırda bulunan sütunların ortalaması örnek olarak aşağıda gösterilmiştir.

$$b_{1J} = \frac{1}{|J|}\sum_{j=1}^{|5|} b_{1j} = \frac{1}{|5|} \times \left[(-0.1666) + (-0.2597) + (-0.9478) + (-1.5331) + (0.2615)\right] = -0.5291$$

Hesaplanan ilk satır ortalamasına benzer olarak diğer satırlardaki ortalama değerler de bulunur.

$$b_{iJ} = \frac{1}{|J|}\sum_{j=1}^{|J|} b_{ij} = \begin{bmatrix} -0.5291 \\ -0.6087 \\ -0.5571 \\ -0.4391 \\ -0.5240 \end{bmatrix}$$

Her bir satır ve sütun için hesaplanan ortalama değerler MSR skoru formülünde yerine konulur. İkili kümenin her bir elemanından satır ve sütun ortalamaları çıkarılır, genel ortalama eklenerek kareleri toplanır ve MSR skoru elde edilmiş olur.

$$MSR(B) = \frac{1}{|I||J|}\sum_{i=1}^{|I|}\sum_{j=1}^{|J|}\left(b_{ij} - b_{iJ} - b_{Ij} + b_{IJ}\right)^2 = 0.0272$$

UI (Uygunluk İndeksi)

Uygunluk indeksi hesaplamasında MSR'de olduğu gibi satır elemanlarının ortalaması bulunur ve her bir ikili küme elemanı bulunduğu satır ortalamasından farkları alınıp, kareleri toplanarak uygunluk indeksi formülünde yerine konularak hesaplanır. Son olarak ikili kümenin uygunluk indeksi bulunması için her satır için bulunan uygunluk indekslerinin ortalaması alınır.

$$\sigma_{I1}^2 = \sum_{i=1}^{5}(b_{i1} - b_{I1})^2 = 0.0000$$

İkili kümenin ilk satırlarından elde edilen yerel varyanslar tüm satırlar için bulunur. İlk sütun için bulunan tüm satırların yerel varyansı aşağıda gösterilmiştir.

$$\sigma_{I1}^2 = \begin{bmatrix} 0.0000 \\ 0.0148 \\ 0.0016 \\ 0.0214 \\ 0.0001 \end{bmatrix} = 0.0380$$

Bulunan yerel varyans değerleri genel varyans değerine oranlanarak 1'den çıkarılır ve her sütun değeri için uygunluk indeksi hesaplanır.

$$R_{I1} = 1 - \frac{\sigma_{I1}^2}{\sigma_1^2} = 1 - \frac{0.0380}{10.7893} = 0.9965$$

Her sütun için bulunan uygunluk indekslerinin ortalaması alınarak ikili kümenin genel

uygunluk indeksi hesaplanmış olur.

$$R_{IJ} = 1 - \frac{\sigma_{IJ}^2}{\sigma_J^2} = 0.9860$$

SMSR Skoru

MSR skoru için bulunan ikili kümenin satır ve sütun ortalama değerleri bu skorun hesaplanmasında da gereklidir. İkili kümenin birinci satır ve birinci sütununda bulunan eleman için SMSR skoru hesaplaması aşağıdaki gibidir.

$$SMSR(B_{11}) = \frac{(b_{1J} \times b_{I1} - b_{11} \times b_{IJ})^2}{b_{1J}^2 \times b_{I1}^2}$$

$$SMSR(B_{11}) = \frac{(-0.5291) \times (-0.1627) - (-0.1666) \times (-0.5316))^2}{(-0.5291)^2 \times (-0.1627)^2} = 0.0008$$

Her bir ikili küme elemanı için bu skor hesaplandıktan sonra ikili küme için genel SMSR skoru bulunur.

$$SMSR(B) = \frac{1}{|I||J|} \sum_{i \in I, j \in J} \frac{(b_{iJ} \times b_{Ij} - b_{ij} \times b_{IJ})^2}{b_{iJ}^2 \times b_{Ij}^2} = 0.4001$$

CKSB Skoru

CKSB skorunun hesaplanmasında ilk olarak ikili kümelerin satır etkisi $T_k(b)$ ve sütun etkisi $B_k(b)$ belirlenir. Satır ve sütun etkisi aşağıdaki gibi hesaplanır:

$$T_1(b) = \frac{1}{|I(b)|} \sum_{i=1, i \in b}^{I(b)} {b_{iJ1}}^2 - \frac{MSR_1(b)}{J_1(b)} = 0.2846$$

$$B_1(b) = \frac{1}{|J_1(b)|} \sum_{j=1, j \in b}^{J_1(b)} {b_{Ij1}}^2 - \frac{MSR_1(b)}{I(b)} = 0.6785$$

$$T_2(b) = \frac{1}{|I(b)|}\sum_{i=1, i\in b}^{I(b)} {b_{iJ2}}^2 - \frac{MSR_2(b)}{J_2(b)} = 0.2238$$

$$B_2(b) = \frac{1}{|J_2(b)|}\sum_{j=1, j\in b}^{J_2(b)} {b_{Ij2}}^2 - \frac{MSR_2(b)}{I(b)} = -0.1411$$

Satır ve sütun etkileri belirlendikten sonra ölçek faktörü $\alpha = 0.001$ alınarak CKSB aşağıdaki gibi hesaplanır.

$$CKSB(b) = LOG\left(\frac{\max(T_1(b)+\alpha, B_1(b)+\alpha)}{\max(T_2(b)+\alpha, B_2(b)+\alpha)}\right) = 0.4691$$

ACV Skoru

ACV skoru hesaplamasında öncelikle ikili küme elemanlarının Pearson korelasyon katsayısı matrisi bulunur. Birinci satır elemanları ile ikinci satır elemanları arasındaki korelasyon katsayısı hesaplaması aşağıdaki gibidir.

$$r(i_1, i_2) = \frac{\sum_{j=1}^{|J|}(b_{i_1 j} - b_{i_1 J})(b_{i_2 j} - b_{i_2 J})}{\sqrt{\sum_{j=1}^{|J|}(b_{i_1 j} - b_{i_1 J})^2 \sum_{j=1}^{|J|}(b_{i_2 j} - b_{i_2 J})^2}} = 0.9315$$

Bütün satırlar arasındaki Pearson korelasyon katsayısı hesaplandıktan sonra ikili kümenin korelasyon matrisi Şekil 3.5'deki gibi olur.

	x_1	x_2	x_3	x_4	x_5
x_1	1.00	0.93	0.93	0.93	1.00
x_2	0.93	1.00	0.97	0.73	0.92
x_3	0.93	0.97	1.00	0.87	0.99
x_4	0.93	0.73	0.87	1.00	0.94
x_5	1.00	0.92	0.99	0.94	1.00

Şekil 3.5. İkili küme için oluşturulan korelasyon matrisi

Korelasyon matrisi kullanılarak ortalama korelasyon skoru hesaplanır.

$$AC(B) = \frac{\sum_{i_1=1}^{|I|-1} \sum_{i_2=i_1+1}^{|I|} r(i_1, i_2)}{\binom{|I|}{2}} = 9.2106$$

Burada korelasyon değeri sadece satırlar arasındaki ilişkiyi ölçmektedir. Benzer işlemler uygulanarak sütunlar için de korelasyon matrisi hesaplanır ve sütunların da ortalama korelasyon skoru hesaplanarak genel ortalama korelasyon ölçüsü bulunur.

$$ACV(B) = \max\left\{ \frac{\sum_{i_1=1}^{|I|} \sum_{i_2=1}^{|I|} |r(i_1, i_2)| - |I|}{|I|^2 - |I|}, \frac{\sum_{j_1=1}^{|J|} \sum_{j_2=1}^{|J|} |r(j_1, j_2)| - |J|}{|J|^2 - |J|} \right\} = 0.2105$$

SCS Skoru

Satır ve sütunlar için oluşturulan Pearson korelasyon katsayısı matrisinin elemanlarının mutlak değeri alınarak aşağıdaki formülde yerine yazılır:

$$S_{i_1 J} = 1 - \frac{\sum_{i_2 \neq i_1, i_2 \in I} |r(x_{i_1 J}, x_{i_2 J})|}{|I| - 1} = 1 - \frac{|0.2105|}{|5| - 1} = 0.9474$$

$$S_{I j_1} = 1 - \frac{\sum_{j_2 \neq j_1, j_2 \in J} |r(x_{I j_1}, x_{I j_2})|}{|J| - 1} = 1 - \frac{|-0.15|}{|5| - 1} = 0.9625$$

$$SCS = \min(S_{i_1 J}, S_{I j_1}) = 0.9474$$

ASR Skoru

Bu ölçü iki değişkene atanan sıra sayıları arasındaki korelasyonu kullandığı için ikili küme elemanlarının sıra sayıları matrisi belirlenir. Sıra sayıları matrisi aşağıda gösterilmiştir.

	y_1	y_2	y_3	y_4	y_5
x_1	-0.17	-0.26	-0.95	-1.53	0.26
x_2	-0.04	-0.69	-0.71	-1.72	0.12
x_3	-0.12	-0.41	-0.86	-1.60	0.21
x_4	-0.31	0.23	-1.22	-1.32	0.42
x_5	-0.17	-0.23	-0.96	-1.52	0.27

(a)

	y_1	y_2	y_3	y_4	y_5
x_1	17	14	8	3	23
x_2	19	11	10	1	20
x_3	18	12	9	2	21
x_4	13	22	6	5	25
x_5	16	15	7	4	24

(b)

Şekil 3.6. İkili küme ve sıra sayı matrisinin gösterimi (a) ikili küme matrisi (b) ikili kümenin sıra sayıları matrisi

Pearson korelasyon katsayısı kullanılarak bulunan ACV skorundaki benzer işlemler Ortalama Spearman korelasyon değeri (ASR) için de yapılır ve aşağıdaki gibi bulunur.

$$ASR(B) = 2 \times \max\left\{ \frac{\sum_{i_1 \in I} \sum_{i_2 \geq i_1+1, i_2 \in I} r_s(i_1, i_2)}{|I|(|I|-1)}, \frac{\sum_{j_1 \in J} \sum_{j_2 \geq j_1+1, j_2 \in J} r_s(j_1, j_2)}{|J|(|J|-1)} \right\}$$

$$ASR(B) = 2 \times \max\{(0.4470), (-0.1020)\} = 0.8941$$

SBM Skoru

Bu ölçümün hesaplanmasında ilk olarak önceden bulunan Spearman korelasyon katsayıları kullanılır. Daha sonra $\alpha(B_{ij})$ ve $\beta(B_{ij})$ katsayıları varsayılan parametre değerleri 1 alınarak aşağıdaki değer elde edilir.

Öncelikle ikili küme içerisindeki satırlarda ve sütunlarda gözlenen eğilimlerin ifadesini özetleyen $r^G_{B_{IJ}}$ ve $r^C_{B_{IJ}}$ değerleri bulunur.

$$r^G_{B_{IJ}} = \frac{2}{|I| \times (|I|-1)} \sum_{i_1=1}^{|I|} \sum_{i_2=i_1+1}^{|I|} |r(i_1, i_2)| = \frac{2}{|5| \times (|5|-1)} 9.2106 = 0.9210$$

$$r^C_{B_{IJ}} = \frac{2}{|J| \times (|J|-1)} \sum_{j_1=1}^{|J|} \sum_{j_2=j_1+1}^{|J|} |r(j_1, j_2)| = \frac{2}{|5| \times (|5|-1)} 10 = 1.0000$$

$$SBM(B_{ij}) = \alpha(B_{ij}) \times r^G_{B_{IJ}} \times \beta(B_{ij}) \times r^C_{B_{IJ}} = (1) \times (0.9210) \times (1) \times (1) = 0.9210$$

VE Skoru

Sanal Hata (VE) ölçüsünde, ikili küme içerisindeki satır elemanlarının genel eğilimlerinin nasıl olduğunu belirlemek için, satırlardaki her eleman *j* inci sütun elemanlarının ortalaması olacak şekilde yapay satır eklenir. Her bir sütun elemanı ile eklenen bu satır elemanları arasındaki farklar hesaplanarak VE skoru aşağıdaki gibi bulunur.

$$VE(B) = \frac{1}{|I||J|}\sum_{i=1}^{|I|}\sum_{j=1}^{|J|}\left|\hat{b}_{ij} - b_{Ij}\right|$$

$$= \left|(-0.1666) - (-0.1627)\right| + \ldots + \left|(-0.2708) - (-0.2571)\right| = 0.1160$$

3.8.3. İkili küme yapıları için değerlendirme ölçülerinin R programı ile uygulaması

İkili küme yapılarından toplamsal ve çarpımsal model için hesaplanan 10 farklı değerlendirme ölçüsünün nasıl bulunduğu bir önceki bölümde gösterilmiştir. Bu bölümde R programı kullanılarak ikili küme yapılarının değerlendirme ölçüsü hesaplanmıştır. Hesaplanan değerlendirme ölçüleri Çizelge 3.1'de verilmiştir.

Çizelge 3.2. İkili küme yapıları için R programında bulunan değerlendirme ölçüleri

	VAR	MSR	U I	SMSR	CKSB	ACV	SCS	ASR	SBM	VE
Sabit İkili Küme	0.0000	0.0000	-	-	-1.3687	-	-	-	-	0.0000
Yukarı Düzenlenmiş Sabit İkili Küme	0.0000	0.0000	-	0.0000	2.0293	-	-	-	-	0.0000
Toplamsal ve Çarpımsal İkili Küme	10.7893	0.0272	0.9859	0.4001	0.4691	0.2136	0.9474	0.9600	0.9210	0.1159
Toplamsal İkili Küme	18.4668	0.0000	0.9119	54.9437	0.6025	0.2500	0.0000	1.0000	1.0000	0.4241
Çarpımsal İkili Küme	0.8425	0.0272	0.8206	2.1377	-1.6931	-0.1500	0.0000	0.2000	0.9273	0.1159

Sabit ve yukarı düzenlenmiş sabit ikili kümeler için değerlendirme ölçülerinden uygunluk indeksi (UI), ortalama korelasyon ölçüsü (ACV), alt matris korelasyon skoru (SCS), ortalama Spearman korelasyon değeri (ASR) ve Spearman ikili küme ölçüsü (SBM) bulunamamıştır. Bunun nedeni sabit değerli ikili kümelerde varyansın sıfır olmasıdır.

İkili küme yapılarından toplamsal ve çarpımsal (koherent) ikili küme, toplamsal ikili küme yapısındaki varyans değerleri yüksek çıkmıştır. Bunun sebebi ise ikili küme elemanlarına

eklenmiş olan parametre değerlerinden kaynaklanmaktadır. Çarpımsal ikili küme elemanlarına eklenen parametre değeri olmadığı için varyansı küçüktür.

Ortalama karesel artık skoru (MSR) çarpımsal ikili kümeler haricinde tüm ikili küme yapılarında 0 çıkmıştır. Bunun nedeni eklenen parametre değerlerinin satırlar veya sütunlar arasındaki sayısal ilişkiyi değiştirmemesidir.

Değerlendirme ölçülerinden Chia ve Karuturi ikili küme skoru (CKSB) ve uygunluk indeksi (UI) dışındaki tüm ölçüler varyans ölçüsü ile doğru orantılıdır. Varyansı yüksek olan ikili kümenin diğer tüm değerlendirme ölçüsü (CKSB ve UI hariç) diğer ikili küme yapılarına göre yüksek çıkmıştır. En uygun ikili kümenin sanal hata (VE) ölçüsünün sıfır olması gerekir.

Chia ve Karuturi ikili küme skorlarında pozitif değerli olanlar anlamlı ikili kümeler olarak değerlendirilir. Buna göre, ortalamaya (0) göre düzenlenmiş sabit ikili kümeler için bu ölçü anlamlı değilken, yukarı (5) düzenlenmiş ikili küme yapısı için anlamlıdır.

4. UYGULAMA

Bu bölümde, Bölüm 2'de tanıtılan ikili kümeleme algoritmaları Bölüm 3'de belirtilen ikili küme değerlendirme ölçüleri bakımından karşılaştırılmıştır. Bu amaç için ilk olarak farklı senaryolara uygun şekilde yapay veri matrisleri üretilmiştir. Daha sonra ikili kümeleme çalışmalarında iyi bilinen gerçek veri setleri kullanılarak karşılaştırmalar yapılmıştır. Tüm kodlamalarda R Project v3.4.4 ücretsiz yazılım programı kullanılmıştır. İkili kümeleme algoritmalarının veri matrislerine uygulanarak ikili kümelerin elde edilmesinde "CRAN: http://cran.r-project.org/" üzerinden erişimi sağlanan "biclust" paketinden yararlanılmıştır. Veri matrislerinin oluşturulması ve ikili küme değerlendirme ölçülerinin hesaplanması için gerekli olan R kodları ise bu çalışmada oluşturulmuştur. Bu çalışmada kullanılan tüm kodlar eklerde verilmiştir.

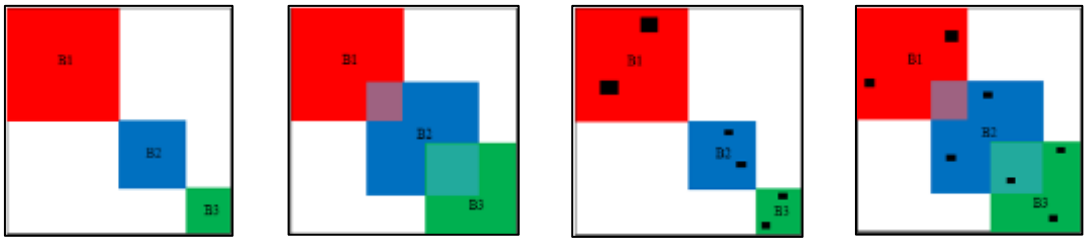
Bu bölümde dikkate alınan tasarım (algoritmalar, ölçüler ve veri yapıları) aşağıdaki tabloda özetlenmiştir.

Çizelge 4.1. Uygulama Tasarımı

Algoritmalar	CC, Bimax, Plaid, Quest, Spectral, Xmotif
Değerlendirme ölçüleri	VAR, MSR, UI, SMRS, CKSB, ACV, SCS, ASR, SBM, VE
Yapay veri matrisleri (1000×1000)	Örtüşme ve aykırı değer yok Örtüşme var ve aykırı değer yok Örtüşme yok ve aykırı değer var Örtüşme ve aykırı değer var
Gerçek veri	Maya verisi (2884×17) İnsan verisi (4096×96) Fare verisi (1080×77)

4.1. Yapay Veri ile Uygulama

Kullanılacak algoritmaların performansı karşılaştırılırken dört farklı senaryo üzerinden simülasyon çalışması yapılmıştır. İkili kümelerde örtüşmenin olduğu ve olmadığı durumlar ve aykırı değerlerin bulunup bulunmadığı durumlar dikkate alınmıştır. Bu veri matrislerine ait yapılar Şekil 4.1'de gösterilmiştir.



Şekil 4.1. Farklı durumlar için oluşturulan ikili kümelerin gösterimi. (a) örtüşmenin ve aykırı değerlerin bulunmadığı durum (b) örtüşmenin olduğu ve aykırı değerlerin olmadığı durum (c) örtüşmenin olmadığı ve aykırı değerlerin olduğu durum (d) örtüşmenin ve aykırı değerlerin olduğu durum

Yapay veri matrislerinin oluşturulmasında ilk adım farklı parametrelere sahip normal dağılımlar kullanılarak 1000×1000 boyutlu bir başlangıç matrisini oluşturmaktır. Bu başlangıç matrisi mümkün olduğu kadar heterojen bir yapıda türetilmektedir. Bunun nedeni daha sonra matris içerisine yerleştirilecek ikili kümelerden daha baskın bir küme yapısının ortaya çıkmasını engellemektir. Sonraki adım önceki bölümlerde anlatılan ikili küme yapılarına uygun olarak farklı boyutlarda 3 ikili küme yapısının oluşturulması ve veri matrisinin içerisine Şekil 4.1'de verilen senaryolara uygun olarak yerleştirilmesidir. Yapay veri matrisi oluşturulduktan sonra bir program parçasıyla satırlar kendi içlerinde ve sütunlar da kendi içlerinde rastgele yer değiştirmektedir. Bu sayede veri matrisi ikili kümelerin kolayca belirlenemeyeceği karmaşık bir hale gelmektedir. Sonraki aşama ikili kümeleme algoritmalarının uygulanması ve elde edilen ikili kümelerle başlangıç veri matrisindeki ikili kümelerin karşılaştırılmasıdır. Aykırı değerler için ikili kümelerde bulunan bazı değerler rastgele olarak uç değerlerle değiştirilmiştir.

Bazı algoritmaların kullanılabilmesi için veri için ön işleme adımları gereklidir. Örneğin Bimax algoritmasının uygulanabilmesi için veri matrisinin ikili değer (binary) alması gerekir (Verma, 2010). Bu nedenle Bimax uygulanmadan önce belli bir eşik değerine göre matris iki değer alan bir yapıya dönüştürülmüştür. Benzer şekilde, Xmotif algoritmasının kullanılması için veri setindeki değişkenlerin kesikli olması gerekmektedir. Bu yüzden sürekli değişkenlere sahip veri seti, kesikli değişkenler olacak şekilde dönüşüm işlemi uygulanmıştır.

İkili kümeleme algoritmaları sezgisel yöntemler oldukları için belirli girdi parametrelerine bağlı olarak çalışırlar. Her ne kadar yazılımlarda her bir algoritma için varsayılan

parametre değerleri bulunsa da bu değerlerin belirlenmesinde literatürde önerilen bazı yaklaşımların dikkate alınması gerekir.

CC algoritması için parametre değerleri

Bu algoritma için delta (δ) parametresi (kabul edilen maksimum H skoru) veri matrisindeki elemanların birinci çeyreklik değerine yakın bir değer olarak alınmıştır (Cheng ve Church, 2000). Bir diğer parametre olan alfa (α) parametresi (ölçek faktörü) varsayılan değeri 1,5 olmasına rağmen zaman performansından daha az etkilenmesi için 1 olarak ayarlanmıştır. Alfa parametresinde her iki değer alınarak uygulama yapıldığında aynı sonuçlar çıkmıştır. Parametrenin minimum aldığı değer seçilerek en kısa zamanda çözüme ulaşılması sağlanmıştır. Varsayılan bir diğer parametre ise maksimum ikili küme sayısıdır. Beklenen ikili küme sayısı 3 olduğu için bu parametrenin varsayılan değeri 3 olarak ayarlanmıştır. Dört farklı durum için de aynı parametre değerleri kullanılmıştır.

Bimax algoritması için parametre değerleri

Bimax algoritması için dönüşüm yapılan veri matrisi ile birlikte beklenen ikili kümelerin satır ve sütun elemanlarının minimum değerleri 200 olarak belirlenmiştir. Bunun nedeni yapay veride üretilen ikili kümelerden en küçüğünün boyutunun 200×200 olmasıdır. Bu algoritmadaki maksimum ikili küme sayısı 3 olarak belirlenmiştir. Bimax algoritması için de dört farklı durumdaki parametre değerleri aynı alınmıştır.

Plaid algoritması için parametre değerleri

Plaid algoritmasındaki varsayılan parametre değerlerinde değişiklik olarak sadece arka plan katmanı olarak düşünülecek bir değer girilmiştir. Beklenen ikili kümeleri bulmak için belirlenen bu değer, matris içerisindeki elemanların ikili kümeye dâhil olması istenilen değerlerin bir arada olmasını sağlamaktadır. Plaid algoritmasının diğer parametre değerleri varsayılan değerlerinde olup değiştirilmemiştir.

Xmotif algoritması için parametre değerleri

Xmotif algoritmasında bulunan parametreler Quest algoritmasındaki parametreler ile aynıdır. Tek fark olarak belirli bir aralık ölçüsü bulunmamaktadır. Belirli bir aralık ölçüsünün olmamasının sebebi ise Xmotif algoritması kesikli değerlere sahip veri matrisinde çalışmasıdır. Dört farklı durum için veri matrisinde dönüşüm benzer şekilde yapılmış ve parametre değerleri de aynı alınmıştır.

Quest algoritması için parametre değerleri

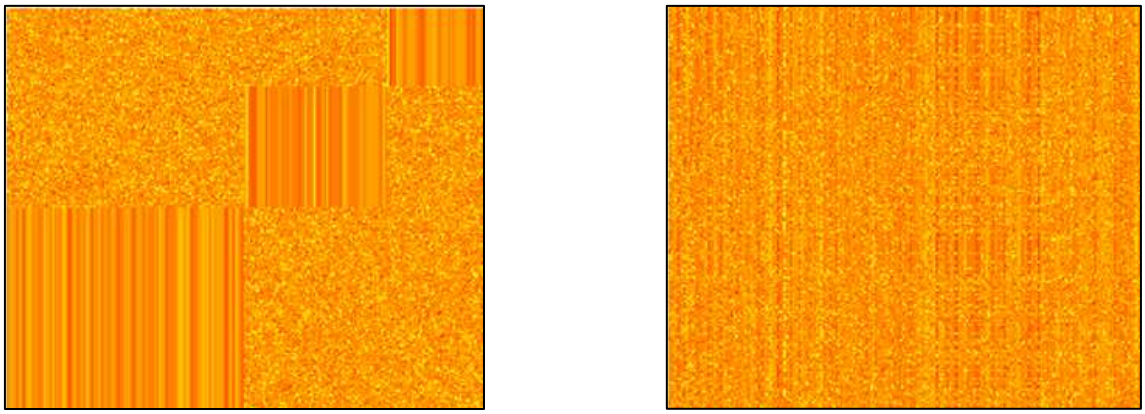
Quest algoritmasında varsayılan parametre değerlerine ek olarak belirli bir aralık ölçüsü ayarlanmıştır. İkili kümelerin oluşturacağı elemanlar için belirlenecek bu aralık değeri 1 olarak alınmıştır. Benzerlik ölçüsü olarak da düşünülen bu parametre değeri veri seti içerisindeki varyans değerine göre belirlenmiştir. Böylece veri matrisi içerisinde oluşturulan ikili kümelerin bulunacağı elemanların tamamı benzerlik ölçüsü içerisinde olacaktır.

Spectral algoritması için parametre değerleri

Spectral algoritması ikili kümeleri bulmak için parametre olarak özdeğer sayısı ve kümeler arası varyans değerinin ayarlanması gerekir. Algoritmanın ilk adımında kullandığı ve daha sonraki adımlarında eleman değerlerine göre değişim gösteren özdeğer sayısı varsayılan değeri 1 olarak alınmıştır. Kümeler arası varyans değeri ikili kümelerin dışında kalacak elemanların belirlendiği aralık içerisinde olmasını sağlamak amacıyla ikili kümelerin olduğu durumlara göre farklı seçilmiştir. Minimum satır ve minimum sütun sayısını belirten parametreleri ise en küçük ikili küme matrisinin boyutları dikkate alınarak 200 olarak ayarlanmıştır. Böylece beklenen üç ikili küme içerisinde daha küçük boyutta ikili kümeler üretilmeyecek ve istenilen ikili kümelerin bulunması sağlanacaktır.

4.1.1. Senaryo 1: Örtüşme ve aykırı değerlerin olmadığı durum

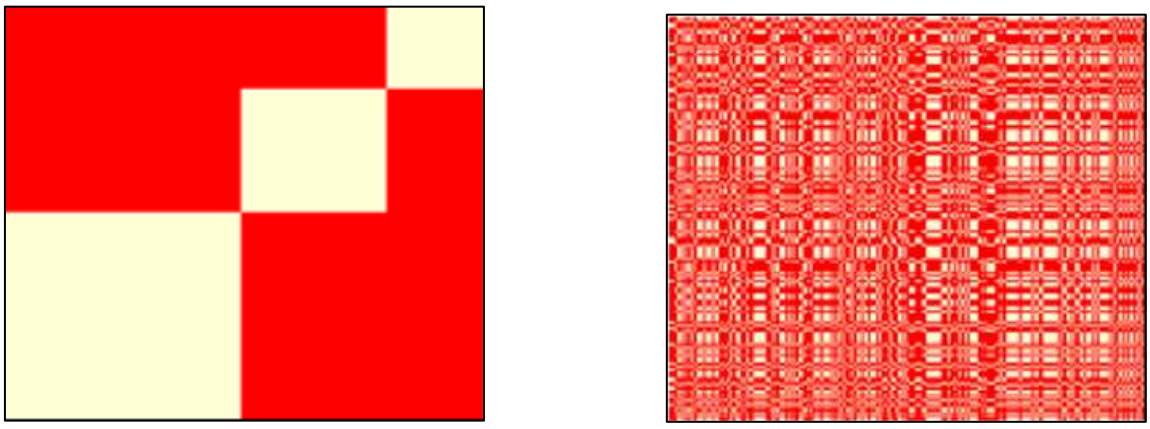
Senaryo 1 için üretilen veri matrisinin başlangıç durumu ve karıştırıldıktan sonraki durumu Şekil 4.2'de gösterilmiştir.



(a) (b)

Şekil 4.2. Senaryo 1 için oluşturulan veri matrisinin ısı grafiği. (a) Üretilen veri matrisinin ısı grafiği (b) Karıştırılmış veri matrisinin ısı grafiği

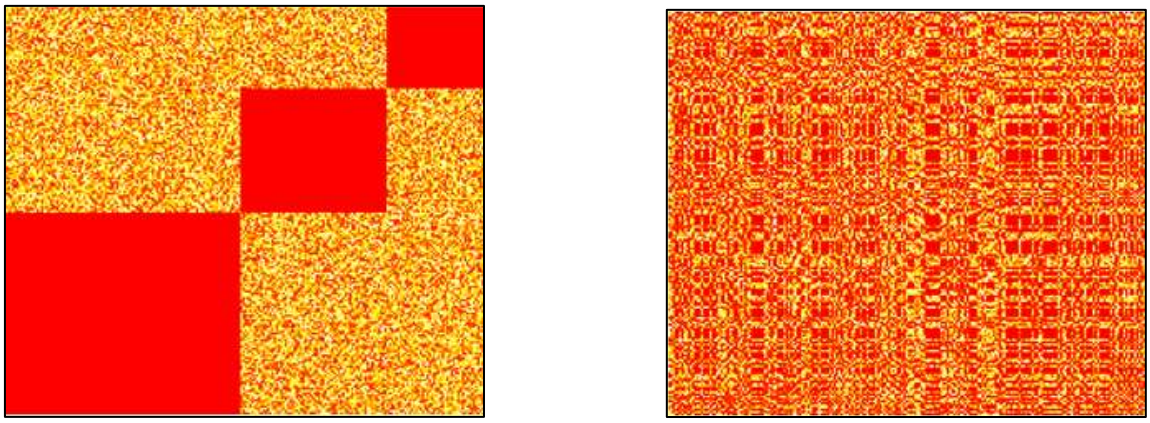
Isı grafiğindeki renklendirme matris içerisindeki sayıların değerine göre en küçük değerdeki sayılar koyu renkten (kırmızıdan) en büyük değerdeki sayılar açık renge (beyaza) göre renk geçişini göstermektedir. İkili kümelerde bulunan eleman değerleri veri setindeki diğer elemanların değerlerine çok yakın olduğu için ısı grafiğinde açık bir şekilde görülememektedir.



(a) (b)

Şekil 4.3. Senaryo 1 için oluşturulan iki değerli veri matrisinin ısı grafiği. (a) Üretilen veri matrisinin ısı grafiği (b) Karıştırılmış veri matrisinin ısı grafiği

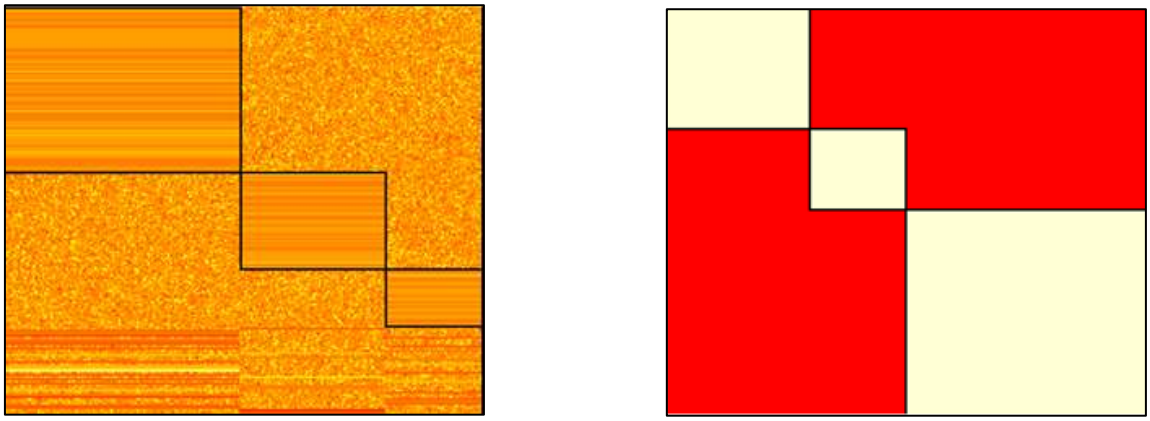
İki değerli veri matrisinde ikili kümelerin aldığı değer 1, diğerlerinin aldığı değer 0 olduğu için ikili küme elemanları açık renkte, diğerleri koyu renkte gösterilmiştir.



(a) (b)

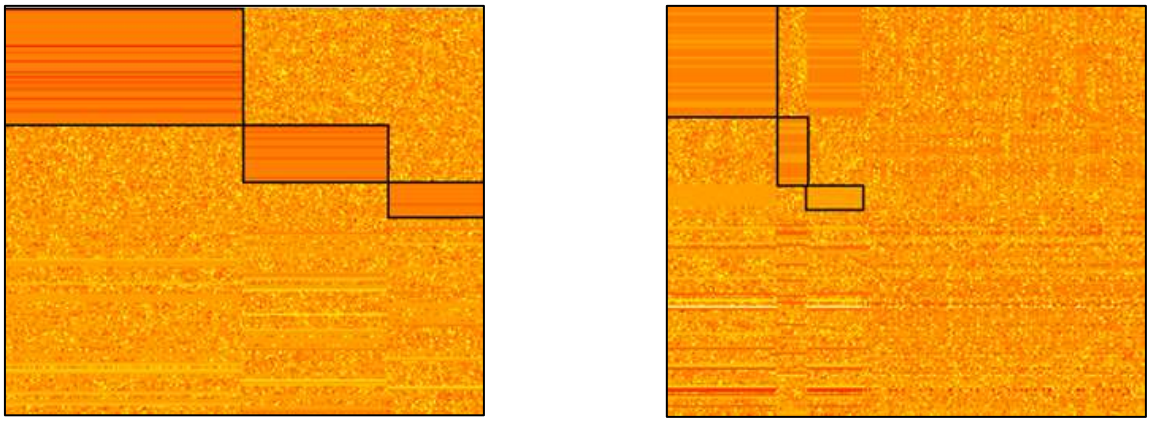
Şekil 4.4. Senaryo 1 için oluşturulan kesikli değerli veri matrisinin ısı grafiği. (a) Üretilen veri matrisinin ısı grafiği (b) Karıştırılmış veri matrisinin ısı grafiği

Veri matrisi oluşturulduktan sonra 6 farklı algoritma ile ikili kümeler elde edilmiştir. Beklenen ikili kümelere en yakın sonuca ulaşan algoritmaların sonuçları karşılaştırılmıştır. Algoritmaların elde ettikleri ikili kümelerin ısı grafikleri aşağıda gösterilmiştir.



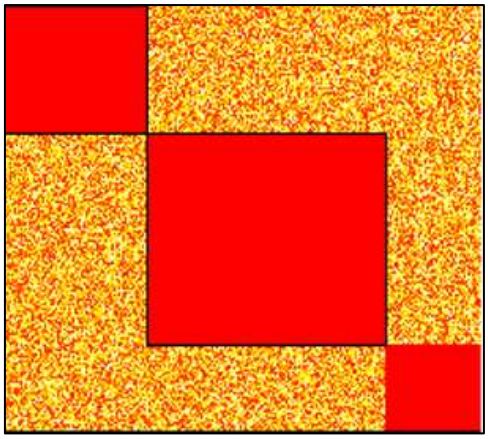
Şekil 4.5. Senaryo 1 için CC ve Bimax algoritmalarından elde edilen ısı grafikleri

CC algoritması ile 400×500, 236×300 ve 141×200 boyutlu ikili kümeler elde edilmiştir. Bimax algoritmasında ise birinci ikili küme 300×300, ikinci ikili küme 200×200 ve üçüncü ikili küme 500×500 boyutlu matris olarak bulunmuştur. Bimax algoritmasındaki veri matrisi içerisinde ikili kümelerin değeri 1, diğerlerinin 0 olduğu için ikili kümeler açık renkte gösterilmiştir.



Şekil 4.6. Senaryo 1 için Plaid ve Quest algoritmalarından elde edilen ısı grafikleri

Plaid algoritması 285×500 boyutlu,137×300 boyutlu ve 88×200 boyutlu olan ikili kümeleri, Quest algoritması ise 272×300 boyutlu,168×63 boyutlu ve 56×196 boyutlu olan ikili kümeleri elde etmiştir. Her iki algoritma da üç ikili küme bulunmasına rağmen beklenen boyutlarda ikili kümeler bulunamamıştır.



Şekil 4.7. Senaryo 1 için Xmotif algoritmasından elde edilen ısı grafiği

Spectral algoritması ikili küme bulamamıştır. Çünkü ikili kümelerin dama tahtası yapısında olduğu durumlarda çalışan Spectral algoritması normal dağılımdan gelen veri seti için uygun değildir. Xmotif algoritması ise 500×500 boyutlu ve 300×300 boyutlu ikili kümeleri elde etmiştir. Xmotif algoritması için kesikli değişkene dönüştürülen veri matrislerinde ikili kümelerin elemanları 1 değeri almış, geriye kalan elemanlar ise 2 ve 10 arasındaki kategorilere ayrılmıştır. Böylece beklenen ikili kümeler bulunmuştur. Isı grafiğinde Xmotif algoritmasında bulunan ikili kümelerin diğer algoritmalardaki ikili

kümelere göre farklı renkte olmasının sebebi ikili küme elemanlarının veri matrisi içerisinde en düşük kategorik değere sahip olmasıdır.

Bu grafiklere bakarak bir yorum yapmak gerektiğinde gerçek duruma en yakın sonucu Bimax algoritmasının en uzak sonucu ise Quest algoritmasının verdiği söylenebilir.

Her bir algoritma ile bulunan ikili kümelerin değerlendirme ölçüleri hesaplanmıştır. Bu ölçülere bakılarak hangi algoritmanın hangi durumda daha iyi sonuçlar verdiği yorumlanabilir. Başlangıç veri matrisine görsel açıdan en yakın sonuçları veren ikili kümeleme algoritmalarının on farklı değerlendirme ölçüsüne göre karşılaştırılması aşağıdaki çizelgede verilmiştir.

Çizelge 4.2. Algoritmalardan elde edilen ikili küme değerlendirme ölçüleri

	VAR	MSR	U I	SMSR	CKSB	ACV	SCS	ASR	SBM	VE
CC	182 788	0.0000	0.9980	0.0000	0.6687	0.4979	0.9990	0.9959	0.9910	0.8019
Bimax	0.0000	0.0000	-	0.0000	4.6151	-	-	-	-	0.0000
Plaid	117 323	0.0000	0.9976	0.0000	1.1462	0.4979	0.9990	0.9959	0.9910	0.9456
Quest	3464	0.0041	0.9962	1.0503	0.0462	0.4699	0.9895	0.9401	0.9490	0.4534
Xmotif	0.0000	0.0000	-	0.0000	-3.5754	-	-	-	-	0.0000

İkili küme algoritmalarından Bimax ve Xmotif algoritmasıyla elde edilen ikili kümelerin eleman değerleri sabit olduğu için varyans ölçüsü 0 çıkmıştır. Buna bağlı olarak korelasyon ölçüsüne dayalı olan değerlendirme ölçüleri bu algoritmalar için bulunamamıştır. Elemanları sabit olan ikili kümeler için değerlendirme ölçüsü olarak CKSB ölçüsüne bakılabilir. Bimax algoritması ile elde edilen ikili kümenin CKSB ölçüsü 4.61 değerinde çıkmıştır. Xmotif algoritması ile elde edilen ikili kümenin CKSB ölçüsünün negatif çıkmasının sebebi ise ikili küme elemanlarının küme dışındaki diğer elemanlara göre daha düşük değerde olmasıdır. Kesikli değerli veri matrislerindeki ikili kümeleri değerlendirmek için CKSB skorunun mutlak değerce yorumlanması daha anlamlı olacaktır.

CC, Plaid ve Quest algoritmalarından elde edilen ikili kümeleri değerlendirme ölçülerine göre karşılaştırılırsa her ölçü için benzer sonuçlar çıktığı görülmektedir. Örneğin CC ve Plaid algoritmaları ile elde edilen ikili kümelerin varyansı çok yüksek, MSR skorları ise

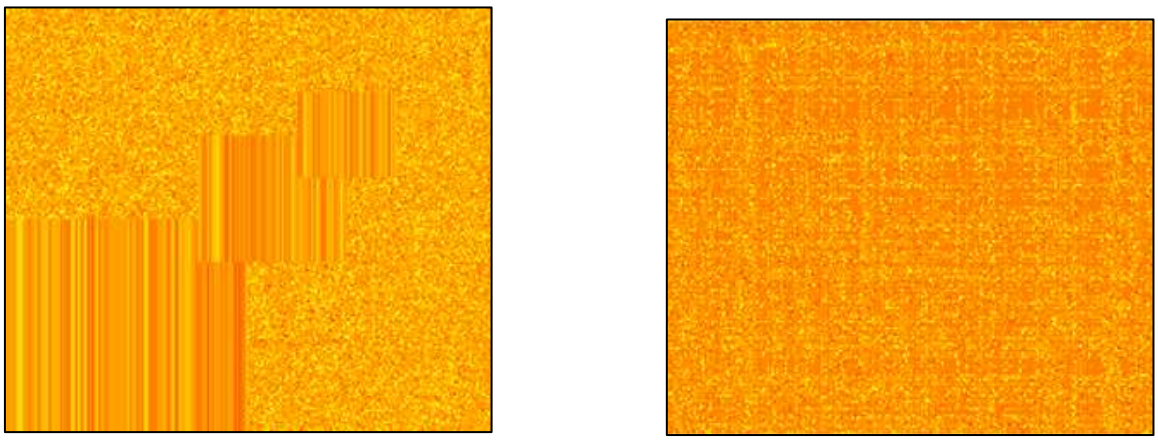
0'a çok yakın bir değerde çıkmıştır. Quest algoritması ile elde edilen ikili kümenin ise varyans değeri diğer algoritmalara göre daha düşük, MSR skoru daha yüksek çıkmıştır. Bu üç algoritma ile elde edilen ikili kümelerin uygunluk indeksi birbirine çok yakın değerler almakla birlikte maksimum uygunluk indeksi skoruna çok yakın çıkmıştır.

CKSB ve VE ölçüsünde Plaid algoritması ile elde edilen ikili kümenin skoru diğer algoritmalardan elde edilen ikili kümelere göre (Bimax ve Xmotif hariç) daha yüksek çıkmıştır. CKSB skorunun yüksek ve VE skorunun düşük olması istenilen en iyi sonucu vermektedir. Plaid algoritması ile elde edilen ikili küme diğer algoritmalarla elde edilen ikili kümelere göre CKSB skorunda daha iyi ve VE skorunda daha kötü sonuç vermiştir.

Genel olarak elde edilen değerlendirme ölçülerine bakılarak CC ve Plaid algoritması ile elde edilen ikili kümenin diğer algoritmalar ile elde edilen ikili kümelere göre daha iyi sonuçlar verdiği görülmektedir. Ayrıca CKSB ve VE skoruna göre CC algoritması ile elde edilen ikili kümenin Plaid algoritması ile elde edilen ikili kümeye göre daha iyi olduğu görülmektedir.

4.1.2. Senaryo 2: Örtüşmenin olduğu ve aykırı değerlerin olmadığı durum

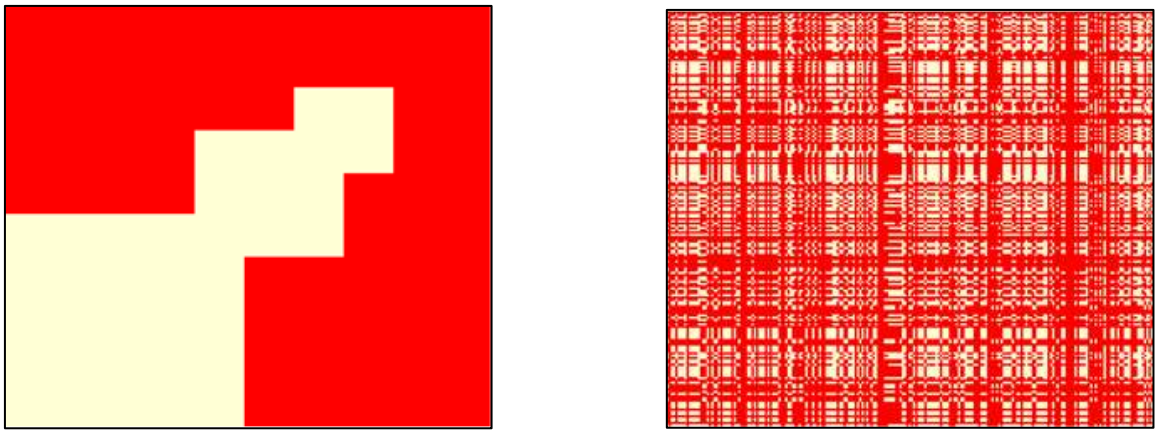
Senaryo 2 için üretilen veri matrisinin başlangıç durumu ve karıştırıldıktan sonraki durumu Şekil 4.8'de gösterilmiştir.



(a) (b)

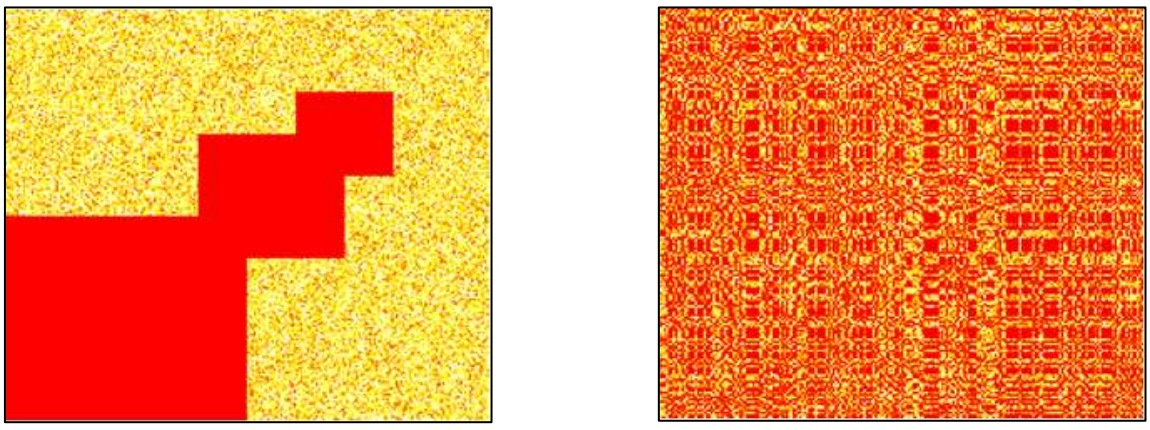
Şekil 4.8. Senaryo 2 için oluşturulan veri matrisinin ısı grafiği. (a) Üretilen veri matrisinin ısı grafiği (b) Karıştırılmış veri matrisinin ısı grafiği

Üretilen veri matrisi içerisindeki ikili kümelerin örtüşme olduğu durum veri matrisi karıştırılmadan önceki halinde görülmektedir. İki değerli ve kesikli değerli veri matrisinde örtüşmenin olduğu ikili kümeler ve üretilen veri matrisinin karıştırılmış hali Şekil 4.9 ve Şekil 4.10'da gösterilmiştir.



(a) (b)

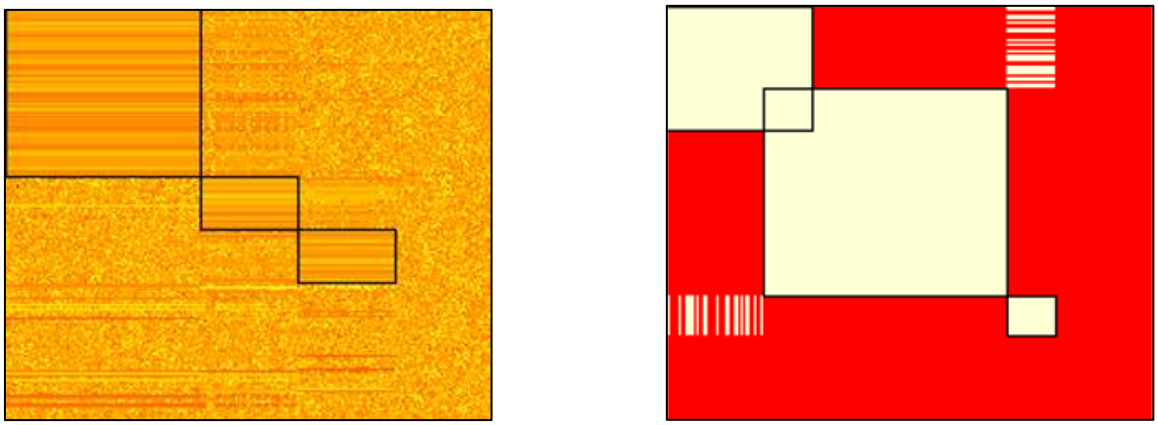
Şekil 4.9. Senaryo 2 için oluşturulan iki değerli veri matrisinin ısı grafiği. (a) Üretilen veri matrisinin ısı grafiği (b) Karıştırılmış veri matrisinin ısı grafiği



(a) (b)

Şekil 4.10. Senaryo 2 için oluşturulan kesikli değerli veri matrisinin ısı grafiği. (a) Üretilen veri matrisinin ısı grafiği (b) Karıştırılmış veri matrisinin ısı grafiği

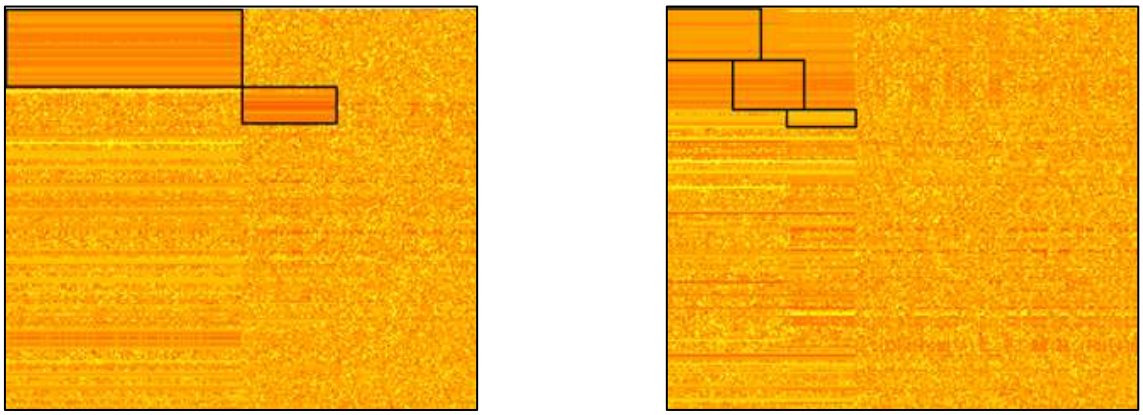
Örtüşen ikili kümelerin bulunduğu veri matrisi oluşturulduktan sonra algoritmalar ile ikili kümeler elde edilmiştir. Algoritmalardan elde edilen ikili kümelerin ısı grafikleri aşağıda gösterilmiştir.



Şekil 4.11. Senaryo 2 için CC ve Bimax algoritmalarından elde edilen ısı grafikleri

CC algoritması ile elde edilen ikili kümelerde örtüşme gerçekleşmemiştir. Algoritma örtüşmeleri dikkate almadan 405×400 boyutlu, 130×200 boyutlu ve 127×199 boyutlu ikili kümeler elde etmiştir.

Bimax algoritmasındaki bulunan ikili kümelerden 300×300 boyutlu ve 500×500 boyutlu ikili kümeler hem beklenen boyutta elde edilmiş hem de örtüşme beklendiği gibi gerçekleşmiştir. Ancak 300×300 boyutlu ikili küme ile 200×200 boyutlu ikili kümenin örtüşmesi sağlanamamış, örtüşmenin bulunduğu elemanlar üçüncü ikili kümenin dışında kalmıştır. Böylece elde edilen üçüncü ikili küme örtüşmenin olmadığı 100×100 boyutlu ikili küme olarak bulunmuştur.

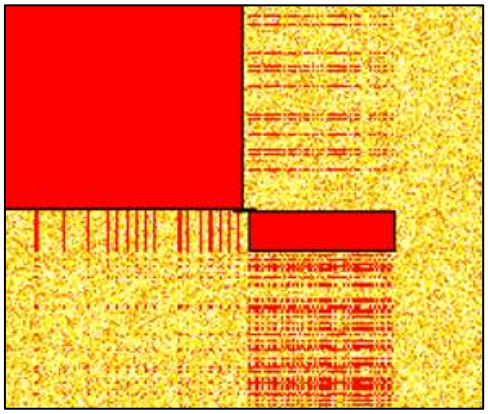


Şekil 4.12. Senaryo 2 için Plaid ve Quest algoritmalarından elde edilen ısı grafikleri

Plaid algoritması beklenen birinci ikili kümede örtüşen satır elemanları ikili kümeye dâhil edilmeyerek 192×500 boyutlu ikili küme olarak bulmuştur. Bulunan diğer ikili küme ise

92×200 boyutlu olup birinci ikili küme ile hem satır hem de sütun elemanları örtüşmeyecek şekilde elde edilmiştir.

Quest algoritması ile elde edilen sonuçlarda örtüşmeyen üç küçük boyutlu ikili kümeler bulunmuştur. Quest algoritmasından bulunan ikili kümelerin sütun elemanları beklendiği gibi örtüşme sağlarken, satır elemanları örtüşme olmadan elde edilmiştir. Elde edilen ikili kümelerde sütun elemanları örtüşme sağlasa bile birbirinden farklı ikili kümeler olmadığı için beklenen örtüşme durumu olmamıştır.



Şekil 4.13. Senaryo 2 için Xmotif algoritmasından elde edilen ısı grafiği

Xmotif algoritması örtüşmenin olmadığı durumda elde edilen ikili kümelere benzer sonuç vermiştir. 500×500 boyutlu ve 100×400 boyutlu ikili kümeler elde edilmiş ve bu kümelerin satır ve sütun elemanlarında örtüşme sağlanmamıştır.

Karşılaştırılan 6 farklı algoritmadan sadece Spectral algoritması ile ikili küme elde edilememiştir. Diğer algoritmalarda ise örtüşmenin olduğu durumda bulunan ikili kümelerden Bimax algoritması satır ve sütunlardaki örtüşmeyi ve Quest algoritması sadece sütunlardaki örtüşmeyi sağlamıştır. CC, Plaid ve Xmotif algoritlarından elde edilen ikili kümeler arasında örtüşme sağlanamamıştır. Beklenen ikili kümelere en yakın bulunan ikili kümeler Bimax algoritması ile elde edilmiştir.

Grafikler incelendiğinde gerçek duruma en yakın sonucu Bimax algoritmasının en uzak sonucu ise Quest algoritmasının verdiği söylenebilir.

Örtüşmenin olduğu durumda beş farklı algoritmadan elde edilen ikili kümeler beklenen durumu karşılamamasına rağmen bulunan ikili kümelerin 10 farklı değerlendirme ölçüsü hesaplanmıştır.

Çizelge 4.3. Örtüşmenin olduğu durumda algoritmalardan elde edilen ikili küme değerlendirme ölçüleri

	VAR	MSR	U I	SMSR	CKSB	ACV	SCS	ASR	SBM	VE
CC	153 333	0.0000	0.9852	0.0000	0.8205	0.4974	0.9987	0.9949	0.9980	0.6641
Bimax	0.0000	0.0000	-	0.0000	3.1800	-	-	-	-	0.0000
Plaid	99 978	0.0021	0.9980	0.0009	1.5694	0.4969	0.9990	0.9908	0.9990	0.6768
Quest	7766	0.0054	0.9950	3.7282	0.0577	0.4859	0.9975	0.9401	0.9890	0.4732
Xmotif	0.0000	0.0000	-	0.0000	-3.9636	-	-	-	-	0.0000

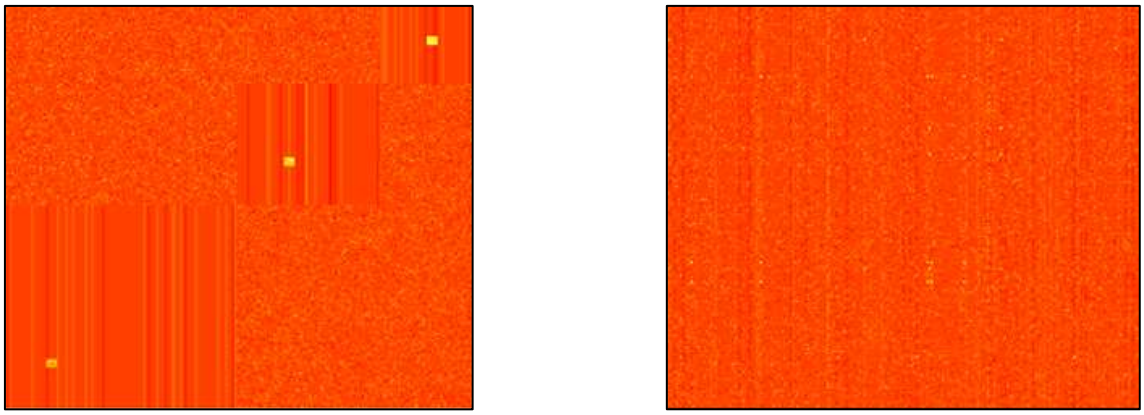
Bimax ve Xmotif algoritmasıyla elde edilen ikili kümelerde örtüşme durumu olsa bile eleman değerleri sabit olduğu için varyans ölçüsü 0 çıkmıştır. Buna bağlı olarak UI, ACV, SCS, ASR ve SBM skorları değerlendirme ölçüsü olarak kullanılamamıştır. Değerlendirme ölçüsü olarak sadece CKSB skoruna bakılabilir. Bimax algoritması ile elde edilen ikili kümenin CKSB ölçüsü 3.18 olarak bir önceki senaryoya göre düşük çıkmıştır. Ancak bulunan ikili küme bu ölçüye göre anlamlıdır. Bir önceki bölümde anlatıldığı gibi Xmotif algoritması ile elde edilen ikili kümenin CKSB skoru negatif olmasına rağmen anlamlı bir ikili küme olduğu söylenebilir.

CC, Plaid ve Quest algoritmalarından elde edilen ikili kümeleri değerlendirme ölçülerinde bir önceki senaryodaki değerlendirme ölçülerine göre düşüş olmuştur. Bunun sebebi ise bulunan ikili kümelerin örtüşmeden dolayı tam olarak elde edilememesidir. Bu üç algoritmadan UI, CKSB, SCS ve SBM ölçülerine göre en iyi sonuç Plaid algoritmasıyla elde edilen ikili kümede olmuştur.

Örtüşmenin olduğu ikili kümeler için genel olarak değerlendirme ölçülerine bakıldığında CC, Plaid ve Quest algoritması ile elde edilen ikili kümelerin birbirine çok yakın değerler aldığı ve diğer algoritmalar ile elde edilen ikili kümelere göre daha anlamlı sonuçlar verdiği görülmektedir.

4.1.3. Senaryo 3: Örtüşmenin olmadığı ve aykırı değerlerin olduğu durum

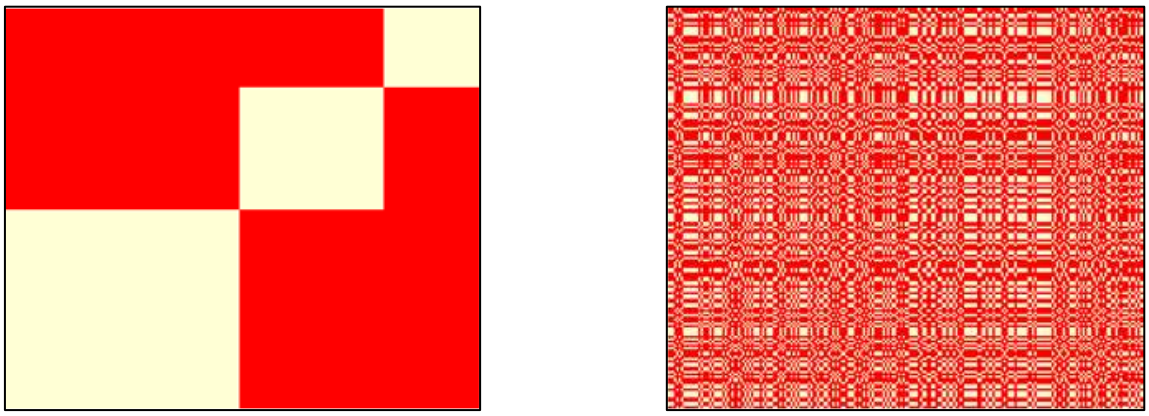
Senaryo 3 için üretilen veri matrisinin başlangıç durumu ve karıştırıldıktan sonraki durumu Şekil 4.14'de gösterilmiştir.



(a) (b)

Şekil 4.14. Senaryo 3 için oluşturulan veri matrisinin ısı grafiği. (a) Üretilen veri matrisinin ısı grafiği (b) Karıştırılmış veri matrisinin ısı grafiği

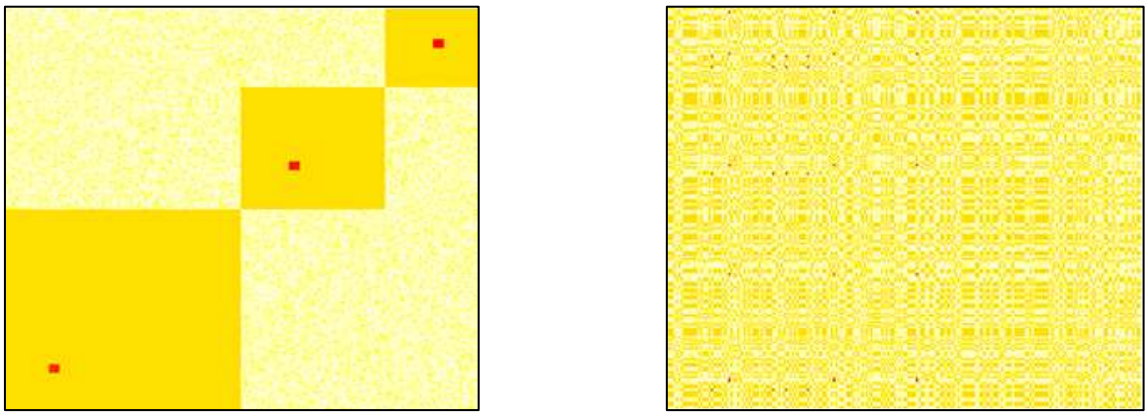
Isı grafiğindeki renklendirmenin bir önceki senaryolara göre koyu renkte çıkması aykırı değerlerin veri matrisine eklenmesinden kaynaklanmaktadır. Matris içerisindeki sayıların aykırı değerlere göre oldukça küçük çıkması ısı grafiğinin görüntüsünü değiştirmiştir. İkili küme elemanlarının sayı değerleri diğer elemanlara daha yakın olması ısı grafiğinde görünmesini zorlaştırmıştır.



(a) (b)

Şekil 4.15. Senaryo 3 için oluşturulan iki değerli veri matrisinin ısı grafiği. (a) Üretilen veri matrisinin ısı grafiği (b) Karıştırılmış veri matrisinin ısı grafiği

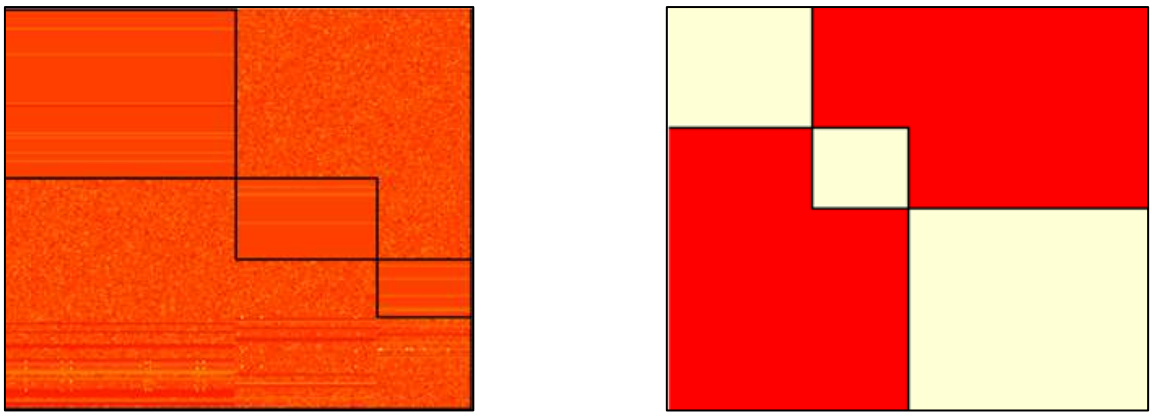
İki değerli veri matrisinde ikili kümelerin aldığı değer 1, diğerlerinin aldığı değer 0 olduğu için veri setine aykırı değerler eklense de dönüşüm işleminde aykırılığını yitirecek ve bu veri seti ile yapılan uygulama ilk senaryo ile aynı olacaktır.



(a) (b)

Şekil 4.16. Senaryo 3 için oluşturulan kesikli değerli veri matrisinin ısı grafiği. (a) Üretilen veri matrisinin ısı grafiği (b) Karıştırılmış veri matrisinin ısı grafiği

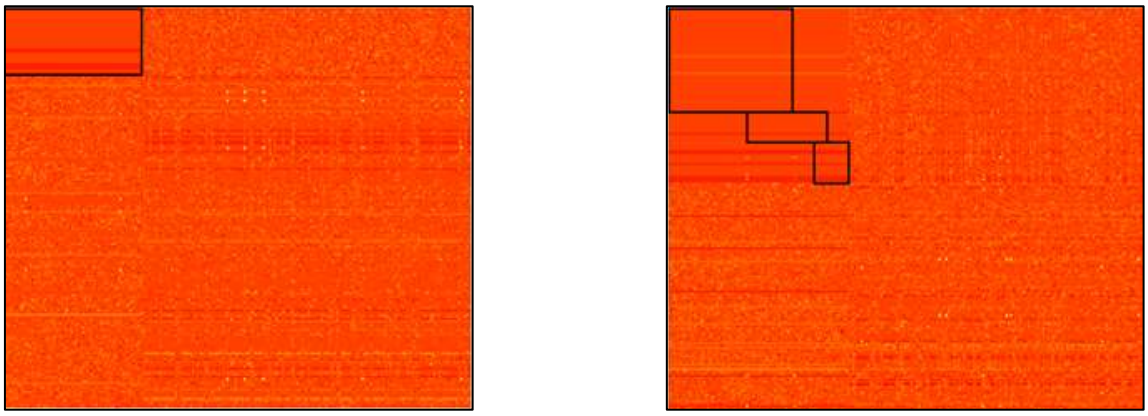
Bu senaryoda örtüşme olmadan sadece ikili kümeler içerisinde rastgele olarak aykırı değerlerin bulunduğu durum göz önüne alınmıştır. Her bir algoritmanın bulduğu ikili kümeler aşağıda gösterilmiştir.



Şekil 4.17. Senaryo 3 için CC ve Bimax algoritmalarından elde edilen ısı grafikleri

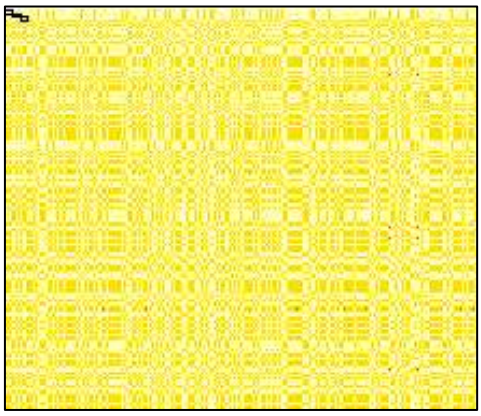
CC algoritması ile 424×500, 205×300 ve 143×200 boyutlu ikili kümeler elde edilmiştir. İlk senaryoya göre elde edilen ikili kümlerin satır elemanlarında azalma meydana gelmiştir. Çünkü aykırı değerlerin bulunduğu satır elemanları ikili kümeler içerisinde bulunmamaktadır. Bimax algoritması ise aykırı değerlerden etkilenmediği için ilk

senaryodaki sonuçlar ile aynı sonuçlar elde edilmiştir.



Şekil 4.18. Senaryo 3 için Plaid ve Quest algoritmalarından elde edilen ısı grafikleri

Plaid algoritması ile 164×300 boyutlu ikili küme elde edilmiştir. Bu ikili küme içerisinde aykırı değerler bulunmamaktadır. Quest algoritması ise 255×256 boyutlu,77×168 boyutlu ve 99×158 boyutlu olan ikili kümeleri elde etmiştir. Örtüşmenin olduğu ve aykırı değerlerin olmadığı durumda bulunan ikili kümeler ile örtüşmenin olmadığı ve aykırı değerlerin bulunduğu ikili kümeler birbirine benzer şekilde elde edilmiştir. Sadece kümelerin boyutlarında değişiklikler söz konusudur.



Şekil 4.19. Senaryo 3 için Xmotif algoritmasından elde edilen ısı grafiği

Xmotif algoritması aykırı değerleri ayrı bir küme olarak belirlemiştir. Spectral algoritması ise bu senaryoda ikili küme bulamamıştır. Bu grafiklere bakarak bir yorum yapmak gerektiğinde gerçek duruma en yakın sonucu Bimax algoritmasının en uzak sonucu ise Xmotif algoritmasının verdiği söylenebilir.

İkili kümeler elde edildikten sonra her bir algoritmanın bulduğu ikili kümelerin değerlendirme ölçüsüne göre karşılaştırılması yapılmıştır. Bu ölçüler Çizelge 4.4'de gösterilmiştir.

Çizelge 4.4. Aykırı değerlerin olduğu durumda algoritmalardan elde edilen ikili küme değerlendirme ölçüleri

	VAR	MSR	U I	SMSR	CKSB	ACV	SCS	ASR	SBM	VE
CC	192 077	0.0000	0.9980	0.0000	0.6010	0.4979	0.9990	0.9959	0.9950	0.7821
Bimax	0.0000	0.0000	-	0.0000	4.6151	-	-	-	-	0.0000
Plaid	57 750	0.0160	0.9966	0.0564	1.3138	0.4901	0.9983	0.9873	0.9840	0.8637
Quest	18 784	0.0011	0.9960	0.0064	0.0921	0.4941	0.9980	0.9961	0.9890	0.4617
Xmotif	0.0000	0.0000	-	0.0000	0.1980	-	-	-	-	0.0000

Aykırı değerler iki değerli veri matrisi içerisinde bulunmadığı için Bimax algoritmasıyla elde edilen ikili kümenin değerlendirme ölçüleri ilk senaryodaki ile aynıdır. Xmotif algoritması ile elde edilen ikili kümede değerlendirme ölçüsü olarak sadece CKSB skoru elde edilmiş, bu skor diğer senaryolara göre farklı çıkmıştır. Xmotif algoritmasından elde edilen ikili kümeler aslında beklenen ikili kümeler olmayıp aykırı değerlerin bulunduğu ikili kümelerdir. Veri matrisinde dönüşüm işlemi uygulandığında aykırı değerlerin bulunduğu kategorinin yüksek olması sonucunda CKSB skoru pozitif bir sayı çıkmıştır. Ancak ikili küme boyutları çok küçük olduğu için bu skora göre anlamlılığı düşük ikili kümeler olarak söylenebilir.

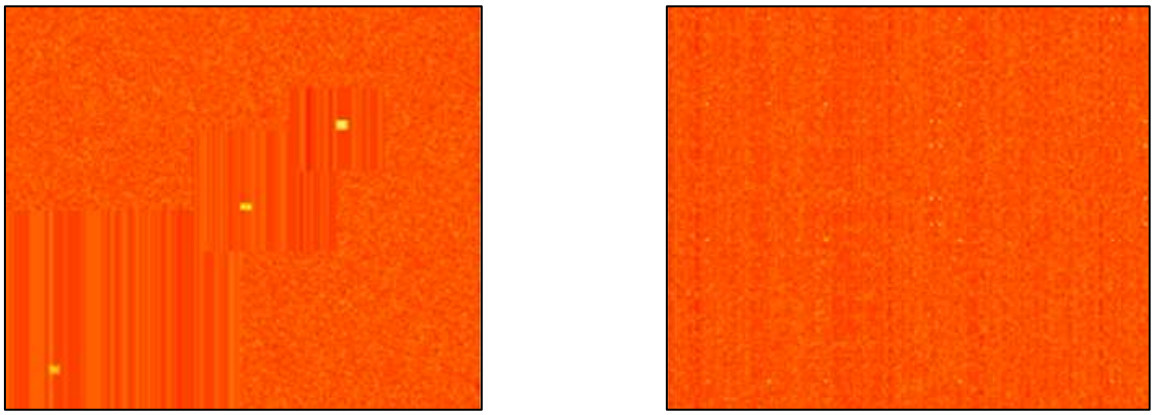
CC ve Quest algoritmaları için elde edilen varyans değeri ilk senaryodakine göre artış göstermiştir. Ancak Plaid algoritmasından elde edilen varyans değeri düşmüştür. Algoritmalardan elde edilen ikili kümeler içerisinde aykırı değerler olmadığından dolayı değerlendirme ölçülerinden çıkan sonuçlar önceki senaryolara göre benzerlik göstermektedir. VE skorunun CC ve Plaid algoritmalarından elde edilen ikili kümeler için yüksek değerde çıkması bu ikili kümelerin anlamlı olmadığını gösterir. Ancak sadece VE skoruna göre ikili kümeleri değerlendirmek doğru olmaz. Diğer değerlendirme ölçülerinin çoğunda ikili kümeler yüksek derecede anlamlı çıkmıştır. VE skorunun yüksek çıkmasının ve bu yüzden anlamsız çıkmasının sebebi ise ikili küme içerisindeki aykırı değerlerden kaynaklanmaktadır.

Değerlendirme ölçülerinden MSR, SMSR, UI, SCS ve SBM skorlarına göre CC

algoritmasından, Plaid ve Quest algoritmalarına göre daha anlamlı sonuçlar elde edilmiştir.

4.1.4. Senaryo 4: Örtüşmenin ve aykırı değerlerin olduğu durum

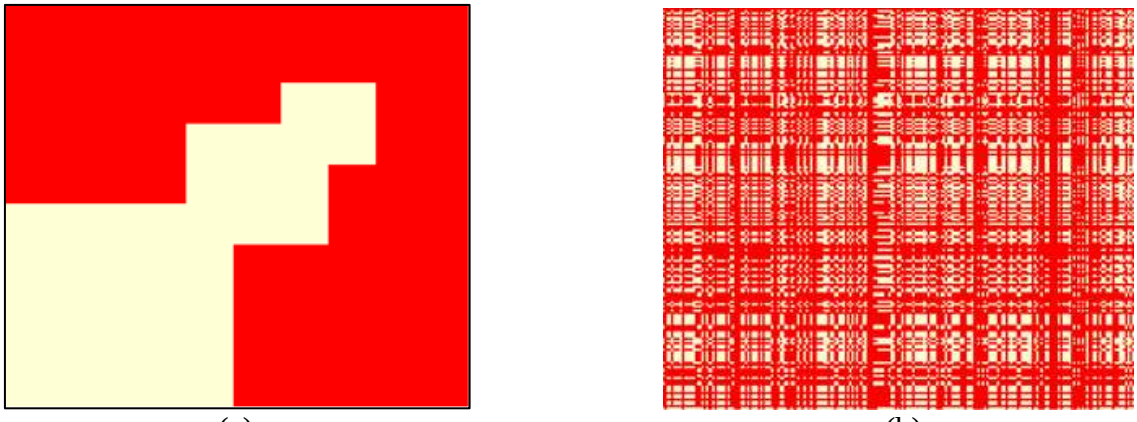
Senaryo 4 için üretilen veri matrisinin başlangıç durumu ve karıştırıldıktan sonraki durumu Şekil 4.20'de gösterilmiştir.



(a) (b)

Şekil 4.20. Senaryo 4 için oluşturulan veri matrisinin ısı grafiği. (a) Üretilen veri matrisinin ısı grafiği (b) Karıştırılmış veri matrisinin ısı grafiği

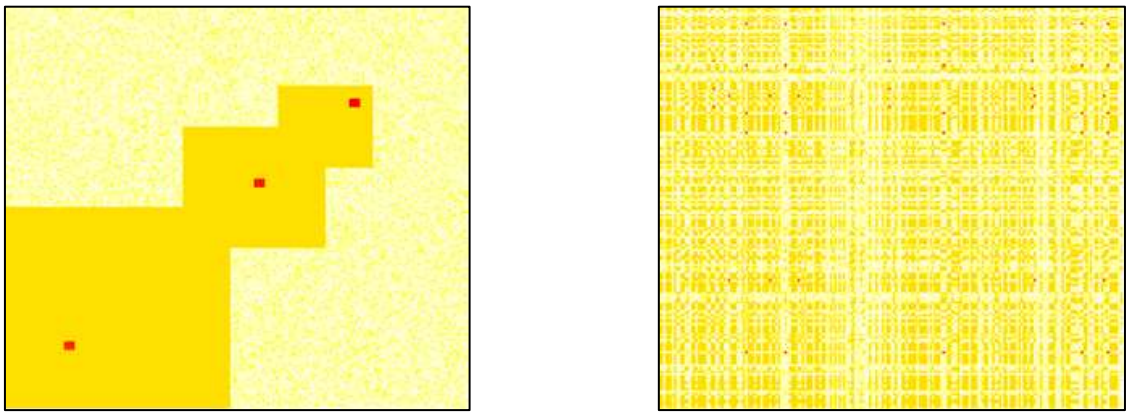
Üretilen veri matrisi içerisindeki örtüşme ve aykırı değerlerin bulunduğu veri matrisi ısı grafiğinde görülmektedir. İki değerli ve kesikli değerli veri matrisinde örtüşmenin ve aykırı değerlerin olduğu ikili kümeler ve üretilen veri matrisinin karıştırılmış hali aşağıda gösterilmiştir.



(a) (b)

Şekil 4.21. Senaryo 4 için oluşturulan iki değerli veri matrisinin ısı grafiği. (a) Üretilen veri matrisinin ısı grafiği (b) Karıştırılmış veri matrisinin ısı grafiği

Dördüncü senaryoda elde edilen iki değerli veri matrisi ikinci senaryodaki ile aynıdır. Çünkü aykırı değerler iki değerli matris içerisinde dönüşüme uğradığından anlamını yitirmiştir.

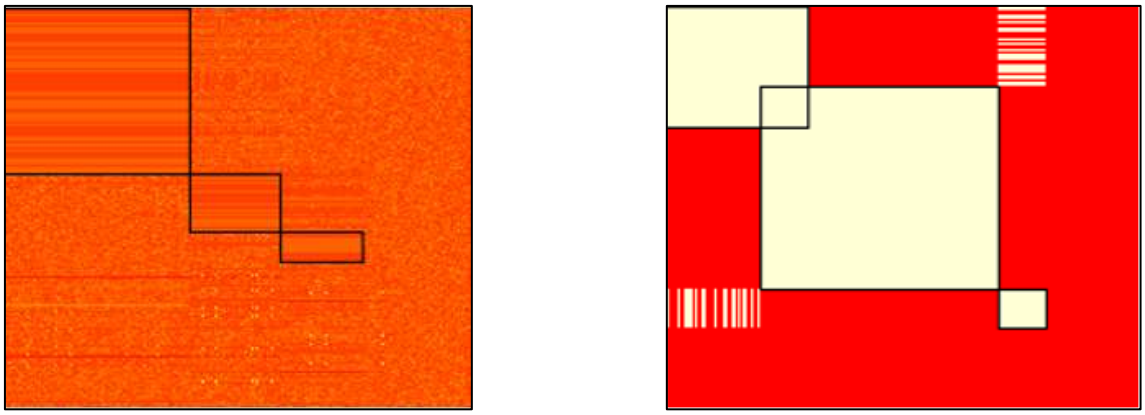


(a) (b)

Şekil 4.22. Senaryo 4 için oluşturulan kesikli değerli veri matrisinin ısı grafiği (a) Üretilen veri matrisinin ısı grafiği (b) Karıştırılmış veri matrisinin ısı grafiği

İkili kümelerin elde edilmesi

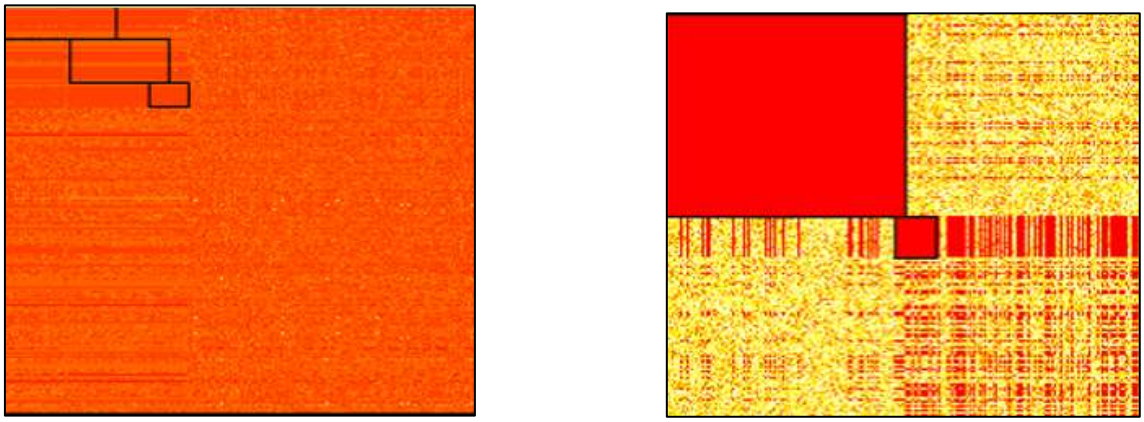
İkili kümeler arasında örtüşmenin olduğu ikinci senaryoda kullanılan veri matrisine aykırı değerler eklenerek dördüncü senaryo oluşturulmuştur. Hem örtüşmenin olduğu hem de aykırı değerlerin bulunduğu veri matrisinde algoritmaların ikili kümeleri bulmada nasıl bir performans göstereceği bu senaryoda incelenmiştir. Algoritmalardan elde edilen ikili kümeler aşağıda gösterilmiştir.



Şekil 4.23. Senaryo 4 için CC ve Bimax algoritmalarından elde edilen ısı grafikleri

CC algoritması ile elde edilen ikili kümelerde örtüşme gerçekleşmemiştir. Ayrıca aykırı değerler ikili kümeler içerisinde değildir. Birinci olarak 413×400 boyutlu ikili küme, ikinci olarak 142×188 boyutlu ikili küme ve üçüncü olarak 79×174 boyutlu ikili küme elde edilmiştir. Bulunan ikili kümelerin örtüştüğü kısımdaki tüm satır ve sütun elemanları ikili kümeler dışında kalmıştır.

Bimax algoritmasından elde edilen sonuç ikinci senaryo ile aynıdır. Çünkü bu senaryo için eklenen aykırı değerler Bimax algoritmasında kullanılan veri matrisinde dönüşümden dolayı etkisini kaybetmiş ve ikili kümelerdeki örtüşmenin olduğu durum olan ikinci senaryo sonuçları çıkmıştır.



Şekil 4.24. Senaryo 4 için Quest ve Xmotif algoritmalarından elde edilen ısı grafikleri

Quest algoritması ile elde edilen ikili kümeler ise ikinci ve üçüncü senaryoda olduğu gibi beklenen birinci ikili kümenin elemanlarını kapsayacak şekilde örtüşmeyen 78×242, 107×205 ve 60×179 boyutlu ikili kümeler bulunmuştur. Quest algoritmasından bulunan ikili kümelerin sadece sütun elemanları örtüşme sağlamıştır.

Xmotif algoritması ile 500×500 boyutlu ve 101×85 boyutlu ikili kümeler bulunmuştur. Bu ikili küme sadece sütun elemanları ile örtüşme sağlamıştır. Ayrıca aykırı değerler dönüşüm işleminde farklı kategoride oldukları için ikili küme elemanları arasında bulunmamaktadır.

Plaid algoritması ve Spectral algoritması ile dördüncü senaryoda ikili küme elde edilememiştir.

Bu grafiklere bakarak bir yorum yapmak gerektiğinde gerçek duruma en yakın sonucu Bimax algoritmasının en uzak sonucu ise Quest algoritmasının verdiği söylenebilir. Senaryo 4 için elde edilen değerlendirme ölçüleri aşağıda gösterilmiştir.

Çizelge 4.5. Örtüşme ve aykırı değerlerin olduğu durumda algoritmalardan elde edilen ikili küme değerlendirme ölçüleri

	VAR	MSR	U I	SMSR	CKSB	ACV	SCS	ASR	SBM	VE
CC	166 988	0.0000	0.9661	0.0000	0.8801	0.4974	0.9987	0.9949	0.9980	0.8309
Bimax	0.0000	0.0000	-	0.0000	3.1800	-	-	-	-	0.0000
Quest	7 556	0.0325	0.9958	0.1489	0.2776	0.4542	0.9981	0.9103	0.9980	0.5540
Xmotif	0.0000	0.0000	-	0.0000	-3.9636	-	-	-	-	0.0000

Dördüncü senaryoda Plaid ve Spectral algoritması ile ikili küme elde edilemediğinden Çizelge 4.5'de sadece dört farklı algoritma ile elde edilen ikili kümlerin değerlendirme ölçüleri bulunmaktadır. Ayrıca Bimax ve Xmotif algoritmaları ile elde edilen ikili kümelerin değerlendirme ölçüsü sadece CKSB skoru ile bulunmuştur.

CC ve Quest algoritmaları ile elde edilen ikili kümelerin değerlendirme ölçüleri karşılaştırıldığında CC algoritması ile elde edilen ikili kümenin varyansı Quest algoritması ile elde edilen ikili kümenin varyansına göre çok yüksek değerlidir. Buna bağlı olarak MSR skoru CC algoritması ile elde edilen ikili kümede Quest algoritması ile elde edilen ikili kümeye göre daha düşük çıkmıştır. Uygunluk indeksleri iki algoritma için de yeterli ölçüde yüksek çıkmış olsa da Quest algoritması ile elde edilen ikili kümede daha yüksektir. Ortalama korelasyon değeri (ACV) ikili kümelerin elemanları arasındaki ilişkinin yüksek olmasa da anlamlı ölçüde çıkmıştır. Ortalama Spearman değeri (ASR) ve Spearman ikili küme ölçüsü (SBM) 1'e çok yakın değerler çıkarak ikili kümelerin anlamlı olduğunu işaret eder. CC algoritması ile elde edilen ikili kümenin değerlendirme ölçüleri içerisinde anlamsız olarak düşünülecek tek ölçü VE skorudur. Bunun sebebi ise VE skorunun toplamsal ve çarpımsal ikili küme yapısında iyi sonuç vermemesinden kaynaklanmaktadır.

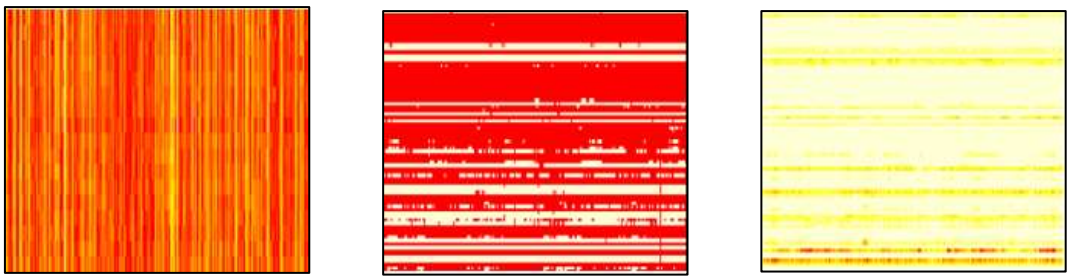
4.2. Gerçek Veri ile Uygulama

Bu bölümde üç farklı veri seti kullanılarak algoritmalardan elde edilen ikili kümelerin değerlendirme ölçülerine göre karşılaştırılması yapılmıştır.

İlk olarak 2884 gen ve 17 koşuldan oluşan “Saccharomyces cerevisiae” maya veri matrisi kullanılmıştır. İkinci olarak insandaki lenf hücrelerinin gen açıklamasını içeren 4096 gen ve 96 koşuldan oluşan veri matrisi kullanılmıştır. Veri seti içerisinde bulunan kayıp değerler için ortalama değer alınmıştır. Bu iki veri matrisine “http://arep.med.harvard.edu/biclustering/” adresinden ulaşılmıştır. Üçüncü olarak farelerin protein etkisini içeren 1080x77 boyutlu veri matrisi kullanılmıştır. Veri seti 72 farenin 77 farklı protein-protein etkileşim skorunu (PPI) her bir fare için 15 farklı ölçümle elde edilen değerlerini göstermektedir. Böylece veri matrisi toplamda 1080 satır eleman değerine sahip olmuştur. Bu veri setine ise “http://archive.ics.uci.edu/ml/datasets/Mice+Protein+Expression#” adresinden erişim sağlanmıştır (Zhu ve diğerleri, 2017). Kullanılan gerçek veri matrisleri maya, insan ve fare verisi olarak isimlendirilmiştir.

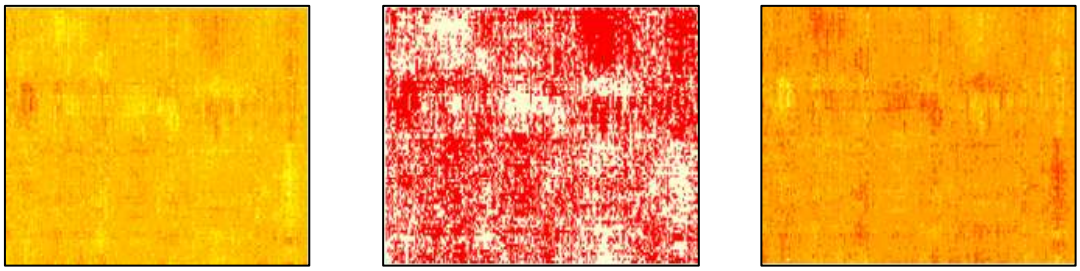
CC, Bimax, Plaid, Quest, Spectral ve Xmotif algoritmaları ile ikili kümeler bulmak için her bir gerçek veri setine ayrı ayrı uygulanmış ve sonuçları değerlendirme ölçülerine göre karşılaştırılmıştır.

Bimax ve Xmotif algoritmalarının uygulanması için maya, insan ve fare verilerine dönüşüm uygulanmıştır. Bimax algoritması için her üç veri seti iki değerli veri matrisine dönüştürülmüştür. Bu dönüşüm uygulanırken eşik değeri olarak her verinin ortalaması alınmıştır. Maya verisi için ortalama değer 225.89, insan verisi için ortalama değer 1.22 ve fare verisi için ortalama değer 0.67 eşik değeri için kullanılan sayı değerleridir. Xmotif algoritmasının uygulanması için üç veri seti de kesikli değerler alacak şekilde dönüştürülmüştür. Maya, insan ve fare veri matrislerinin ısı grafikleri Şekil 4.25’de gösterilmiştir.



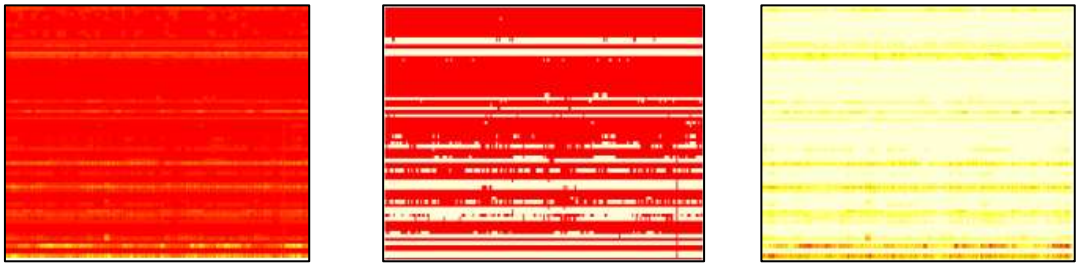
(a) (b) (c)

Şekil 4.25. Maya verisi için ısı grafikleri (a) orijinal maya verisi (b) iki değerli maya verisi (c) kesikli değerli maya verisi



(a) (b) (c)

Şekil 4.26. İnsan verisi için ısı grafikleri (a) orijinal insan verisi (b) iki değerli insan verisi (c) kesikli değerli insan verisi



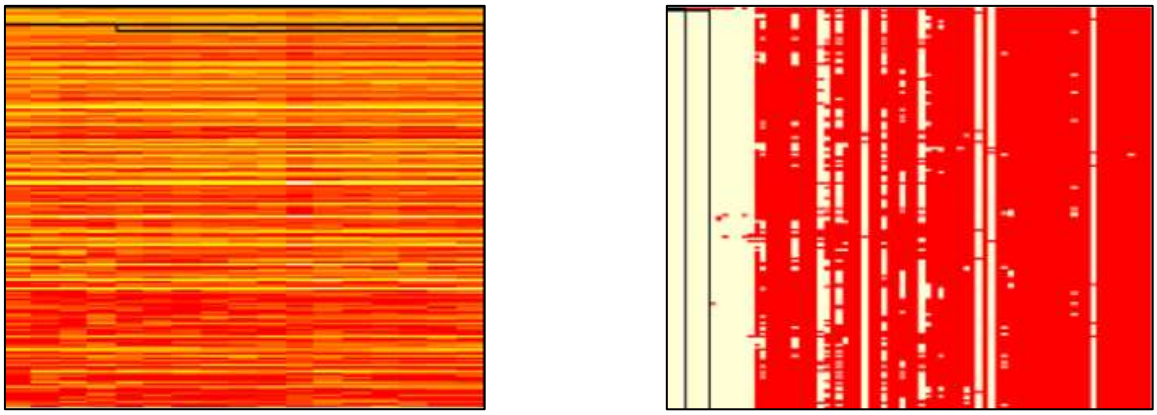
(a) (b) (c)

Şekil 4.27. Fare verisi için ısı grafikleri (a) orijinal fare verisi (b) iki değerli fare verisi (c) kesikli değerli fare verisi

4.2.1. Gerçek veri setlerinden elde edilen ikili kümeler

Maya verisi için elde edilen ikili kümeler

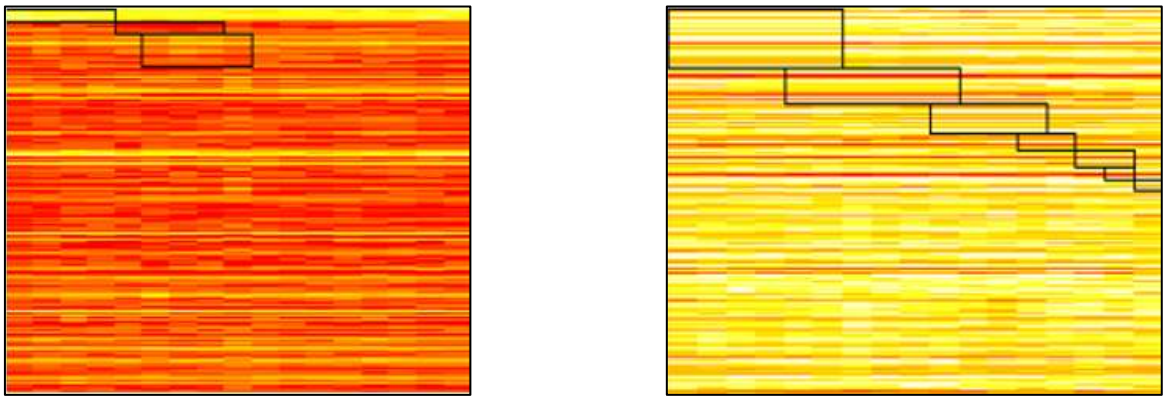
Maya verisi için yapılan analizler sonucunda CC, Bimax, Plaid ve Xmotif algoritmaları ile ikili kümeler elde edilmiştir. Ancak Quest ve Spectral algoritmalarından ikili küme elde edilememiştir. Dört farklı algoritmadan elde edilen ikili kümeler aşağıda gösterilmiştir.



Şekil 4.28. Maya verisi için CC ve Bimax algoritmalarından elde edilen ısı grafikleri

CC algoritmasındaki parametre değerlerinden sadece delta (δ) parametresi değiştirilmiştir. Bu parametre değeri veri setinin birinci çeyreklik değeri 139 olarak ayarlanmıştır (Cheng ve Church, 2000). Maksimum ikili küme sayısı bulma parametresi varsayılan değeri 100 olarak alındığı için maya verisinde maksimum sayıda ikili küme elde edilmiştir. Bulunan 100 ikili küme içerisinde en büyük boyutta (116×17) elde edilen ikili küme değerlendirme ölçüleri hesaplanırken CC algoritması için referans olarak alınmıştır.

Bimax algoritmasından elde edilen 100 ikili kümenin boyutları birbirine çok benzer çıkmıştır. Referans olarak alınan ilk ikili küme 1080×3 boyutludur. Bimax algoritması uygulanırken iki değerli dönüştürülmüş maya verisi kullanıldığı için elde edilen ikili kümelerin eleman değerlerinin tümü 1'dir.



Şekil 4.29. Maya verisi için Plaid ve Xmotif algoritmalarından elde edilen ısı grafikleri

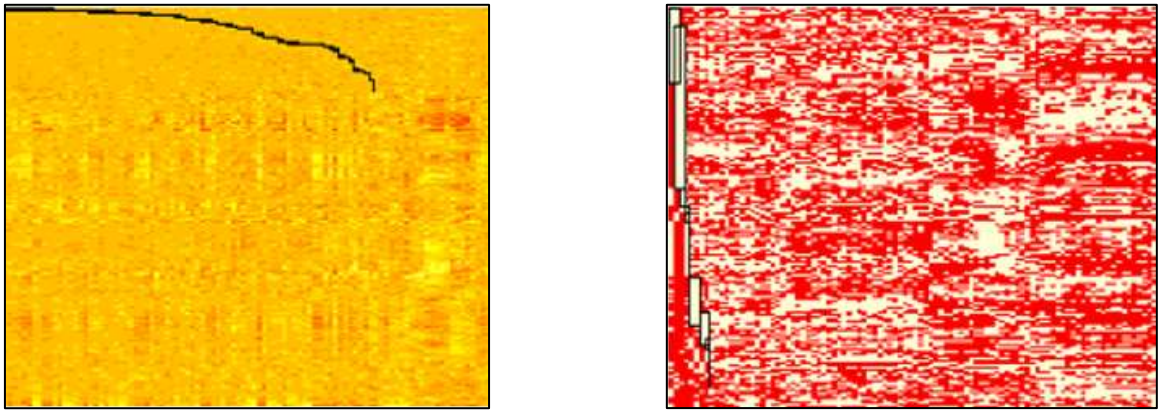
Plaid algoritmasından üç ikili küme elde edilmiştir. Bu ikili kümeler satır elemanları ile örtüşme sağlamamıştır. Ancak ikinci ve üçüncü ikili kümenin sütun elemanlarında örtüşme

sağlanmıştır. Elde edilen üç ikili kümeden en büyük boyutlu (254×4) olan referans ikili küme olarak değerlendirme ölçülerinde kullanılmıştır.

Xmotif algoritmasında ise sütun elemanlarından bazıları örtüşecek şekilde on farklı ikili küme elde edilmiştir. Burada en fazla satır elemanına sahip 423×6 boyutlu ikili küme referans olarak değerlendirme ölçülerinde kullanılmıştır.

İnsan verisi için elde edilen ikili kümeler

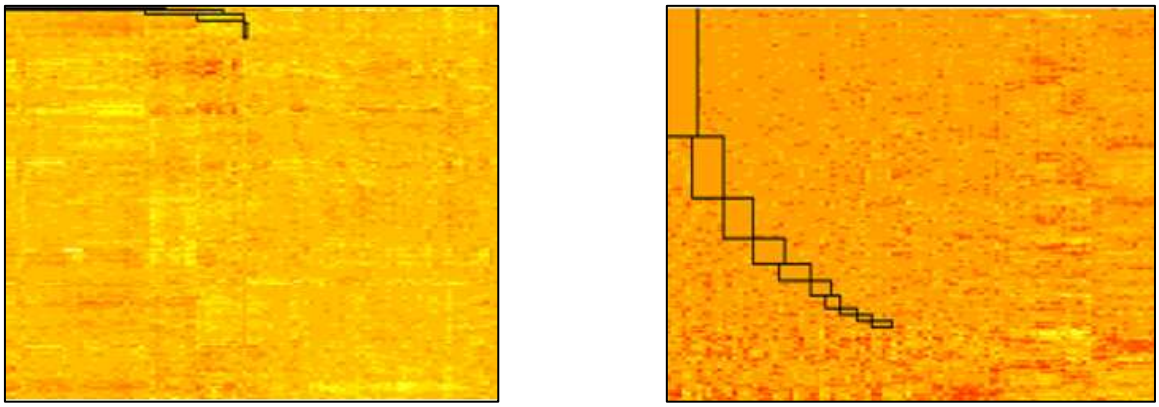
İnsan verisi için elde edilen ikili kümeler maya verisinde olduğu gibi CC, Bimax, Plaid ve Xmotif algoritmalarından elde edilmiştir. Dört farklı algoritmadan elde edilen ikili kümeler aşağıda gösterilmiştir.



Şekil 4.30. İnsan verisi için CC ve Bimax algoritmalarından elde edilen ısı grafikleri

CC ve Bimax algoritması ile elde edilen ikili kümelerin boyutları ısı grafiğinde de görüldüğü gibi veri setine göre oldukça düşük çıkmıştır. Bütün ikili kümelerin boyutları birbirine çok yakındır. Diğer algoritmalar ile karşılaştırması yapılacak ikili küme 14×10 boyutludur.

Bimax algoritmasından elde edilen 100 ikili kümeden 1608 satır elemanı ile oluşturulan ikili küme en yüksek satır elemanına sahiptir. Bulunan ikili kümelerin tümünde sütun elemanları 2 veya 3 değişkenden oluşmuştur.



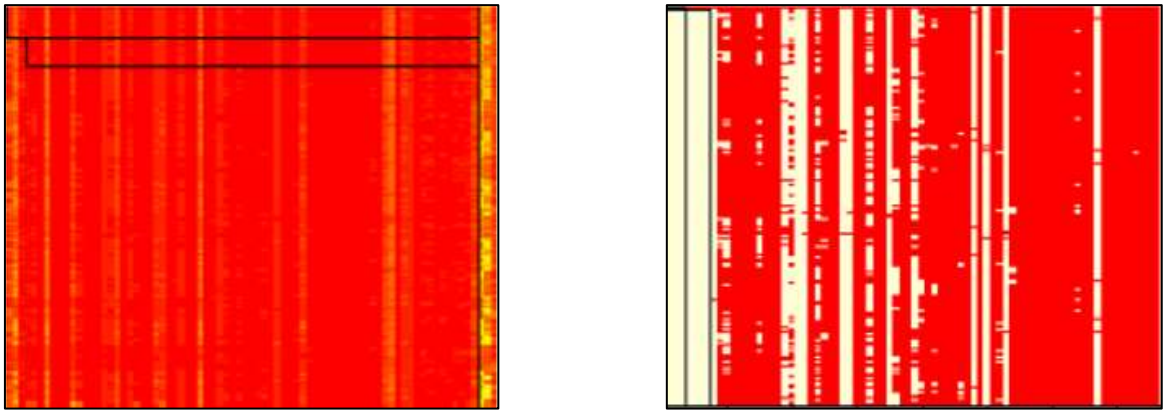
Şekil 4.31. İnsan verisi için Plaid ve Xmotif algoritmalarından elde edilen ısı grafikleri

Plaid algoritmasında altı tane ikili küme elde edilmiştir. Bu ikili kümelerin satır eleman sayıları sütun elemanlarına göre oldukça düşüktür. En fazla satır elemanına sahip 21×31 boyutlu ikili küme değerlendirme ölçülerine göre diğer algoritmalardan elde edilen ikili kümeler ile karşılaştırılmıştır.

Xmotif algoritması ile elde edilen on ikili kümenin sütun elemanları tümünde eşit ve 6 çıkmıştır. En fazla satır elemanına sahip 1318×6 boyutlu ikili küme tüm veri setindeki satır elemanlarının yaklaşık %30'unu oluşturmuştur.

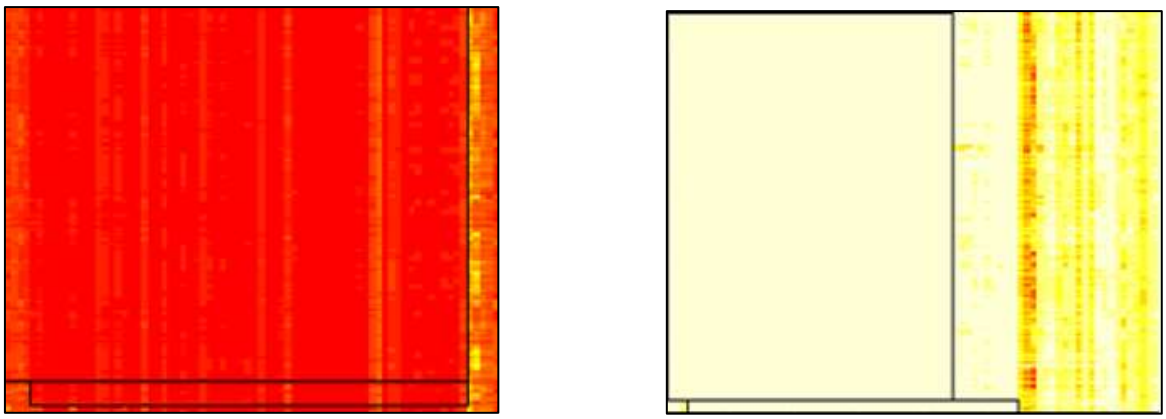
Fare verisi için elde edilen ikili kümeler

Beş farklı ikili küme algoritmalarından fare verisi için ikili kümeler elde edilmiştir. Sadece Plaid algoritması ile ikili küme elde edilememiştir. Bulunan ikili kümeler ve ısı grafikleri aşağıda verilmiştir.



Şekil 4.32. Fare verisi için CC ve Bimax algoritmalarından elde edilen ısı grafikleri

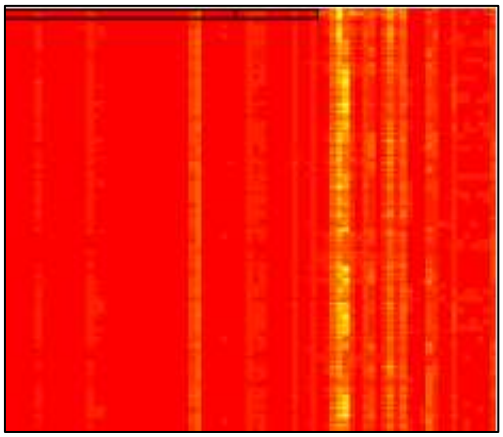
Fare verisinden CC algoritması ile 15 ikili küme Bimax algoritması ile 3 ikili küme elde edilmiştir. CC algoritmasından elde edilen ikili kümelerin sütun eleman sayıları Bimax algoritmasından elde edilen ikili kümelere göre oldukça fazladır. Satır elemanlarında ise durum tam tersidir. Fare verisinde bulunan tüm satır elemanları Bimax algoritması ile elde edilen ikili kümeler içerisinde bulunmaktadır.



Şekil 4.33. Fare verisi için Quest ve Xmotif algoritmalarından elde edilen ısı grafikleri

Maya ve insan verilerinde Quest algoritması ile ikili küme elde edilemezken, fare verisinden iki tane ikili küme elde edilmiştir. Bu ikili kümler neredeyse tüm veriyi kapsayacak şekilde oluşmuştur. 996 satır elemanı ve 72 sütun elemanına sahip ikili küme veri setinin yaklaşık %80'ini oluşturmaktadır.

Xmotif algoritmasından elde edilen ikili kümeler veri setinin tüm satır elemanlarını kapsamaktadır. Birinci ikili kümenin satır eleman sayısı 1050, ikinci ikili kümenin satır eleman sayısı ise 30'dur. Birinci ve ikinci ikili kümenin sütun eleman sayıları ise sırasıyla 44 ve 51'dir.



Şekil 4.34. Fare verisi için Spectral algoritmasından elde edilen ısı grafiği

Quest algoritmasında olduğu gibi üç veri setinden sadece fare verisinde Spectral algoritması ile ikili küme elde edilmiştir. Elde edilen ikili kümelerin boyutları 22×36 ve 22×13'tür. Isı grafiğinde de görüldüğü gibi ikili kümelerin tüm satır elemanları birbiri ile örtüşmüştür.

4.2.2. Gerçek veri setlerinden elde edilen ikili kümelerin değerlendirme ölçülerine göre karşılaştırılması

Maya, insan ve fare verilerinden farklı algoritmalarla elde edilen ikili kümelerin karşılaştırılması 10 farklı değerlendirme ölçüsüne göre yapılacaktır. Bimax ve Xmotif algoritmasından elde edilen ikili kümelerin elemanları tek değerli olduğu için varyans ölçüsü ile belirlenecek değerlendirme ölçüleri bulunamamıştır. Bu algoritmaların karşılaştırması sadece Chia Karuturi ikili küme uygunluk skoruna göre değerlendirilecektir. Üç veri seti için uygulanan algoritmalardan elde edilen ikili kümelerin değerlendirme ölçüleri Çizelge 4.6'da gösterilmiştir.

Çizelge 4.6. Gerçek veri setleri kullanılarak farklı algoritmalardan elde edilen ikili kümelerin değerlendirme ölçüleri

		VAR	MSR	U I	SMSR	CKSB	ACV	SCS	ASR	SBM	VE
Maya Verisi	CC	7 101 481	141.8739	0.1380	0.0018	0.1028	0.4181	0.9754	0.9601	0.9650	51.4699
	Bimax	0.0000	0.0000	-	0.0000	1.4293	-	-	-	-	0.0000
	Plaid	2 851 610	177.2982	0.9896	0.0011	0.1737	0.1557	0.9610	0.2806	0.9510	62.5176
	Xmotif	0.0000	0.0000	-	0.0000	0.2315	-	-	-	-	0.0000
İnsan Verisi	CC	23 206	96.6423	0.9090	2 731	-0.9861	0.0193	0.9982	0.1788	0.9080	8.6709
	Bimax	0.0000	0.0000	-	0.0000	1.4889	-	-	-	-	0.0000
	Plaid	2 439 716	2113.2030	0.9716	0.3835	-0.4582	0.1356	0.9956	0.2606	0.9470	45.7867
	Xmotif	0.0000	0.0000	-	0.0000	-0.0266	-	-	-	-	0.0000
Fare Verisi	CC	1 827	0.0099	0.8911	0.0616	-2.6688	0.2598	0.9968	0.7930	0.9970	0.0653
	Bimax	0.0000	0.0000	-	0.0000	1.4293	-	-	-	-	0.0000
	Quest	17 859	0.0209	0.0721	0.0917	-2.7468	0.0673	0.9990	0.1735	0.9980	0.0916
	Spectral	115	0.0044	0.9993	0.0131	-2.3400	0.1022	0.9535	0.9774	0.9530	0.0375
	Xmotif	0.0000	0.0000	-	0.0000	0.1966	-	-	-	-	0.0000

Maya verisi kullanılarak CC, Bimax, Plaid ve Xmotif algoritmalarından elde edilen ikili kümelerin dördü için bulunan CKSB skoru karşılaştırıldığında en yüksek değeri Bimax almaktadır. Diğer değerlendirme ölçülerine göre CC ve Plaid algoritmalarından elde edilen ikili kümeler karşılaştırılırsa, CC algoritması ile elde edilen ikili kümenin ACV, SCS, ASR ve SBM skorları Plaid algoritması ile elde edilen ikili kümeye göre daha yüksektir. Bu skorlara göre CC algoritması ile elde edilen ikili kümenin Plaid algoritmasından elde edilen ikili kümeye göre daha anlamlı çıktığı söylenebilir. Ayrıca VE skorunun da daha düşük çıkması yine CC algoritmasıyla elde edilen ikili kümenin daha anlamlı olduğunu gösterir. CC algoritmasıyla elde edilen ikili kümenin sadece uygunluk indeksi (UI) düşük çıkmıştır.

İnsan verisi kullanılarak elde edilen ikili kümeler maya verisinde olduğu gibi CC, Bimax, Plaid ve Xmotif algoritmaları ile elde edilen ikili kümelerdir. Bimax ve Xmotif algoritmasından elde edilen ikili kümelerden CKSB skoruna göre Bimax daha iyi sonuç vermiştir. CC algoritmasıyla elde edilen ikili kümenin sadece SCS, SBM ve VE skorları Plaid algoritmasıyla elde edilen ikili kümeye daha iyi sonuçlar vermiştir. Ayrıca ikili kümelerin varyans ölçülerine bakıldığında CC algoritmasıyla elde edilen ikili kümenin Plaid algoritmasıyla elde edilen ikili kümeye göre oldukça düşük çıktığı görülmektedir. Varyans ölçüsü (VAR) ile anlamlı bir karşılaştırma yapılması için ikili kümelerin

boyutlarının birbirine eşit olması gerekir. Çünkü ikili kümedeki eleman sayısının artmasıyla varyans ölçüsü artacaktır. Böylece eleman sayısı daha fazla olan ikili kümenin az olana göre varyans ölçüsü daha yüksek çıkacaktır.

Fare verisinde Bimax ve Xmotif algoritmaları ile elde edilen ikili kümelerin CKSB skoruna göre karşılaştırılması yapıldığında, maya ve insan verisinde olduğu gibi Bimax sonuçlarının daha anlamlı olduğu söylenebilir. Ayrıca CC, Quest ve Spectral algoritması ile elde edilen ikili kümelerin varyans ölçüsü en düşük olan Spectral algoritması ile elde edilen ikili kümedir. MSR, SMSR, CKSB, SCS ve SBM ölçülerine göre Spectral algoritması ile elde edilen ikili küme CC ve Quest algoritması ile elde edilen ikili kümelere göre daha kötü sonuçlar elde etmiştir.

Genel olarak maya verisi için CC algoritması, insan verisi için Plaid algoritması ve fare verisi için CC ve Plaid algoritmaları daha iyi performansa sahiptir. Üç veri seti için de Bimax algoritmasının Xmotif'e göre daha iyi olduğu söylenebilir.

5. SONUÇ VE ÖNERİLER

Gen örüntü (pattern) yapılarının ortaya çıkartılmasında genlerin ve koşulların (örneklerin) aynı anda kümelenmesi son yıllarda artan bir öneme sahiptir. Geleneksel yöntemler bu durumlar için yetersiz kaldığı için ikili kümeleme algoritmaları geliştirilmiştir. Farklı veri yapıları için farklı algoritmalar önerildiği için literatürde çok sayıda ikili kümeleme algoritması mevcuttur. Bunlara; CC, FLOC, Bimax, biMINE, CMonkey, CTWC, OP-cluster, Plaid, QUBIC, Quest, ISA, FABIA, SAMBA, Xmotif ve Spectral algoritmaları örnek olarak verilebilir. Bu çalışma kapsamında bu algoritmalardan yaygın olarak kullanılan CC, Bimax, Plaid, Quest, Spectral ve Xmotif ikili kümeleme algoritmaları ele alınmıştır. Bu algoritmalardan CC, Plaid, Quest ve Spectral algoritmaları sürekli değişkene sahip veri matrisleri için, Xmotif algoritması kesikli veya kategorik değerli veri matrisleri için ve Bimax algoritması iki değer alan veri matrisleri için geliştirilmiştir. Bununla birlikte, gen verileri sürekli değerler alan veri matrislerinden oluştuğu için uygun dönüşümlerle bu algoritmaların hepsi uygulanabilmektedir.

İkili kümeleme algoritmalarının performanslarının karşılaştırılmasına ilişkin çalışmalar ise sınırlıdır. Literatürdeki çalışmalar genellikle bir ya da birkaç değerlendirme ölçüsü bakımından bazı algoritmaların karşılaştırılmasına ilişkindir. Bu çalışmalara ilişkin bilgiler tez çalışmasının birinci bölümünde verilmiştir. Bununla birlikte ikili kümelerin değerlendirilmesi için birçok değerlendirme ölçüsü mevcuttur. Bu değerlendirme ölçüleri varyans ölçüsü (VAR), ortalama karesel artık skoru (MSR), uygunluk indeksi (UI), ölçeklenen ortalama karesel artık skoru (SMSR), Chia ve Karuturi ikili küme skoru (CKSB), ortalama korelasyon ölçüsü (ACV), alt matris korelasyon ölçüsü (SCS), ortalama Spearman korelasyon değeri (ASR), Spearman ikili küme ölçüsü (SBM) ve sanal hata (VE)'dir. Bu tez çalışmasında ikili kümeleme algoritmaları bu değerlendirme ölçüleri kullanılarak daha geniş bir şekilde karşılaştırılmıştır. Buna ilaveten performans karşılaştırılmasına ilişkin sonuçlar ısı grafikleri ile de gösterilmiştir. Böylece ikili kümeleme algoritmalarının görsel açıdan da karşılaştırılması sağlanmıştır.

Bu tez çalışmasının uygulama kısmı iki bölümden oluşmaktadır. Birinci bölüm simülasyon çalışması, ikinci bölüm ise gerçek veri uygulamasıdır. Algoritmaların karşılaştırılmasına ilişkin simülasyon tasarımı şu şekilde yapılmıştır: İlk olarak bir R kodu ile dört farklı

senaryo için yapay veriler üretilmiştir. Senaryo 1, ikili kümeler içerisinde örtüşmenin ve aykırı değerlerin olmadığı durumdur. Senaryo 2, ikili kümeler içerisinde örtüşmenin olduğu ancak aykırı değerlerin olmadığı durumdur. Senaryo 3, ikili kümeler içerisinde örtüşmenin olmadığı ancak aykırı değerlerin olduğu durumdur. Senaryo 4, ikili kümeler içerisinde örtüşme ve aykırı değerlerin olduğu durumdur. Bu senaryoların her biri için yukarıda verilen algoritmalar on farklı değerlendirme ölçüsüne göre karşılaştırılmıştır. Karşılaştırmalar ayrıca ısı grafikleri kullanılarak da yapılmıştır. Gerçek veri seti için yapılan uygulamada ise maya verisi, lenf hücrelerinin gen ifadesini içeren insan verisi ve protein-protein etkileşim skorlarını içeren fare verisi kullanılmıştır. Her bir gerçek veri seti için belirtilen algoritmalardan ikili kümeler elde edilmiş ve elde edilen ikili kümelerin performansını ölçmek için on farklı değerlendirme ölçüsü kullanılmıştır. Çalışmada yer alan tüm R kodları Ek.1 ve Ek.2'de verilmiştir.

Yapay veri setinin ilk senaryosunda örtüşme ve aykırı değerlerin olmadığı durumda yapılan analizler doğrultusunda CC, Bimax ve Plaid algoritmaları ile elde edilen ikili kümelerin Quest ve Spectral algoritmaları ile elde edilen ikili kümelere göre daha iyi sonuçlar verdiği görülmüştür. Eren ve diğerleri (2000) tarafından yapılan çalışma incelendiğinde CC ve Plaid algoritmalarının bazı diğer algoritmalardan daha kötü sonuç verdiği görülse de Spectral, Bimax ve Xmotif algoritmalarına göre daha iyi sonuç verdiği görülmektedir. Algoritmaları tek bir değerlendirme ölçüsüne (küme skoruna) göre karşılaştırmasına rağmen ilk senaryodaki çıkan sonucu desteklemektedir.

Örtüşmenin olduğu ancak aykırı değerlerin olmadığı durumdaki ikinci senaryoda CC, Plaid ve Bimax algoritması ile elde edilen ikili kümelerin değerlendirme ölçüleri birbirine çok yakın değerler alarak diğer algoritmalar ile elde edilen ikili kümelere göre daha iyi sonuçlar verdiği görülmüştür. Algoritmaların elde ettiği sonuçlardaki değerlendirme ölçüleri tek bir ikili kümenin skoruna göre değerlendirildiği için aslında örtüşme durumuna göre karşılaştırmak doğru değildir. Elde edilen diğer ikili kümeler de dikkate alındığında Bimax algoritması ile elde edilen ikili kümeler diğer algoritmalardan elde edilen ikili kümelere göre ikinci senaryo için beklenen en yakın sonucu elde etmiştir. Prelic ve diğerleri (2006), CC ve Xmotif algoritmalarının ikili kümeler arasında örtüşme olduğu durumda iyi bir performans göstermediğini belirterek bu durumu desteklemektedir. Bimax algoritması ile elde edilen ikili kümeler dışındaki diğer tüm algoritmalarla elde edilen ikili

kümelerde örtüşmenin olduğu kısımdaki satır ve sütun elemanları ikili kümeler içerisine dâhil edilememiştir.

Örtüşmenin olmadığı ancak aykırı değerlerin olduğu durumdaki üçüncü senaryoda UI, SCS, ASR ve SBM değerlendirme ölçülerine göre CC, Plaid ve Quest algoritmaları ile elde edilen ikili kümeler öne çıkmıştır. Ancak CC algoritması veri matrisi içerisindeki elemanlar arasında Öklit uzaklık ölçüsünü kullandığı için aykırı değerleri ikili kümeler içerisinde bulundurmayarak küme dışına atmaktadır. Bulunan ikili kümelerin değerlendirme ölçüleri küme içerisinde aykırı değerlerin olmadığı skorlar olarak belirlenmiştir. Sadece varyans ölçüsüne göre karşılaştırma yapılırsa Bimax algoritmasından elde edilen ikili kümeler diğer algoritmalarla elde edilen ikili kümelere göre daha iyi sonuç verdiği söylenebilir. Ayrıca Zhao ve diğerleri (2012) yaptıkları çalışmalarında, algoritmaların eşleşme skoruna göre karşılaştırmasında Bimax algoritmasının Spectral ve Plaid algoritmalarından daha iyi performans gösterdiğini belirterek bu durumu desteklemektedir.

Örtüşme ve aykırı değerin aynı anda bulunduğu dördüncü senaryoda ise CC algoritması ile elde edilen ikili kümenin diğer algoritmalar ile elde edilen ikili kümelere göre daha iyi çıktığı görülmüştür. Ancak tüm algoritmalardan elde edilen ikili kümeler içerisinde aykırı değerler dâhil olmamıştır. Ayrıca ikili kümelerin performansları karşılaştırılırken elde edilen değerlendirme ölçüleri örtüşme durumuna bakmadan sonuçlar vermiştir. Böylece ikili kümeleme algoritmalarının karşılaştırılmaları ilk senaryodakine benzer bir şekilde yapılmıştır.

İkili kümeleme algoritmalarından Bimax ve Xmotif algoritması ile elde edilen ikili kümelerin dört farklı senaryoda da anlamlı ikili kümeler oluşturduğu görülmüştür. Ancak bu algoritmaların uygulanabilmesi için veri matrisine dönüşüm uygulanmış ve aykırı değerlerin bulunduğu senaryolarda veri yapısı aykırı değerlerden etkilenmediği için diğer senaryolar ile benzer sonuçlar vermiştir.

Gerçek veri setine uygulanan ikili kümeleme algoritmalarından maya ve insan verisi için CC algoritması birçok değerlendirme ölçüsüne göre diğer algoritmalardan elde edilen ikili kümelere göre daha anlamlıdır. Bhattacharya ve diğerleri (2012) yaptıkları çalışmada maya verileri üzerinde uyguladıkları algoritmaları karşılaştırmış ve en kötü performans olarak

delta (CC) algoritması olarak bulmuştur. Ancak algoritmaları sadece güvenilirlik skoru P-değerine göre karşılaştırmıştır. Aynı veri üzerinde çıkan bu farklı sonuçlar, aslında değerlendirme ölçülerinin ikili kümeleri karşılaştırırken ne kadar önemli olduğunu göstermektedir.

Fare verisinde ise Bimax algoritmasıyla elde edilen ikili küme diğer algoritmalardan elde edilen ikili kümelere göre daha anlamlı sonuçlar elde etmiştir. CC ve Spectral algoritmalarıyla elde edilen ikili kümelerin uygunluk indeksleri diğer algoritmalara göre daha iyi sonuç verse de tek bir ölçüye göre bu algoritmaları karşılaştırmak anlamlı olmayacaktır. Ayrıca Prelic ve diğerleri (2006) tarafından yapılan çalışmada gen eşleşme skoruna göre Bimax algoritmasının daha iyi performans gösterdiği ve Bozdağ ve diğerleri (2010) tarafından yapılan çalışmada CC algoritmasının daha kötü performans gösterdiği bu durumu desteklemektedir.

Sonuç olarak, CC ve Bimax algoritması yapay veri setleri içerisinde dört farklı senaryoda diğer algoritmalara göre daha iyi sonuçlar vermiştir. Ayrıca gerçek veri setlerinden maya ve insan verisi için de diğer algoritmalar ile karşılaştırıldığında daha iyi sonuçlar verdiği görülmüştür. Plaid algoritması ise yapay veri setlerinden ilk üç senaryoda CC algoritması ile birlikte diğer algoritmalara göre daha iyi sonuç verdiği görülmüştür. Örtüşmenin ve aykırı değerlerin aynı anda bulunduğu dördüncü senaryoda ve gerçek veri setlerinde ise Plaid algoritması diğer algoritmalara göre daha kötü performans göstermiştir. Quest algoritması sadece yapay veri setlerinden örtüşmenin ve aykırı değerlerin ayrı ayrı olduğu senaryolarda CC ve Plaid algoritması ile beraber iyi sonuçlar vermiştir. Spectral algoritması yapay veri setlerinde diğer algoritmalara göre iyi sonuçlar vermese de sadece gerçek veri setlerinden fare verisi için Bimax algoritmasıyla birlikte iyi sonuç vermiştir. Bimax algoritması yapay veri setindeki dört farklı senaryoda Xmotif algoritmasıyla beraber anlamlı ikili kümeler elde etmiş ancak gerçek veri setlerinde diğer algoritmalara göre daha kötü performans göstermiştir. Genel olarak yapay veri setlerindeki dört farklı senaryo için ortak olarak CC, Bimax ve Xmotif algoritmalarının diğer algoritmalara göre daha iyi sonuçlar verdiği söylenebilir. Gerçek veri setlerinde ise CC ve Bimax algoritmaları diğer algoritmalara göre daha iyi sonuçlar vermiştir. Bununla birlikte elde edilen ikili kümeleme sonuçlarının gen verileri açısından kuramsal (biyolojik) olarak uygunluğunun da uzman görüşleri doğrultusunda karşılaştırılması ve tartışılması gerekir.

KAYNAKLAR

Aguilar-Ruiz, J. S. (2005). Shifting and scaling patterns from gene expression data. *Bioinformatics*, 21(20), 3840-3845.

Ayadi, W., Elloumi, M. and Hao, J. K. (2009). A biclustering algorithm based on a Bicluster Enumeration Tree: application to DNA microarray data. *BioData Mining*, 2(1), 9.

Bhattacharya, S., Go, D., Krenitsky, D. L., Huyck, H. L., Solleti, S. K., Lunger, V. A. and Pryhuber, G. S. (2012). Genome-wide transcriptional profiling reveals connective tissue mast cell accumulation in bronchopulmonary dysplasia. *American Journal of Respiratory and Critical Care Medicine*, 186(4), 349-358.

Bozdağ, D., Kumar, A. S. and Çatalyürek, U. V. (2010, August). *Comparative analysis of biclustering algorithms.* In Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology,265-274.

Busygin, S., Prokopyev, O. and Pardalos, P. M. (2008). Biclustering in data mining. *Computers and Operations Research*, 35(9), 2964-2987.

Cheng, Y. and Church G. M. (2000). Biclustering of expression data. *International Conference on Intelligent Systems for Molecular Biology*, 8, 93-103.

Chia, B. K. H. and Karuturi, R. K. M. (2010). Differential co-expression framework to quantify goodness of biclusters and compare biclustering algorithms. *Algorithms for Molecular Biology*, 5(1), 23.

Das, S. and Idicula, S. M. (2011). Comparative Advantages of Novel Algorithms Using MSR Threshold and MSR Difference Threshold for Biclustering Gene Expression Data. *In Software Tools and Algorithms for Biological Systems*,123-134.

Eren, K., Deveci, M., Küçüktunç, O. and Çatalyürek, Ü. V. (2012). A comparative analysis of biclustering algorithms for gene expression data. *Briefings in Bioinformatics*, 14(3), 279-292.

Flores, J. L., Inza, I., Larrañaga, P. and Calvo, B. (2013). A new measure for gene expression biclustering based on non-parametric correlation. *Computer Methods and Programs in Biomedicine*, 112(3), 367-397.

Hartigan, John A. (1972). Direct clustering of a data matrix. *Journal of the American Statistical Association*, 67(337), 123-129.

Kaiser, S. (2011). *Biclustering: methods, software and application*. Doctoral dissertation, Ludwig Maximilian University, Münih.

Kluger, Y., Basri, R., Chang, J. T. and Gerstein, M. (2003). Spectral biclustering of microarray data: coclustering genes and conditions. *Genome Research*, 13(4), 703-716.

Lazzeroni, L. and Owen, A. (2002). Plaid models for gene expression data. *Statistica Sinica*, 61-86.

Li, L., Guo, Y., Wu, W., Shi, Y., Cheng, J. and Tao, S. (2012). A comparison and evaluation of five biclustering algorithms by quantifying goodness of biclusters for gene expression data. *BioData Mining*, 5(1), 8.

Mirkin, B. (1996). Mathematical classification and clustering: From how to what and why. In Classification, data analysis, and data highways. *Springer*, 172-181.

Mukhopadhyay, A., Maulik, U., & Bandyopadhyay, S. (2009). A novel coherence measure for discovering scaling biclusters from gene expression data. *Journal of Bioinformatics and Computational Biology*, 7(05), 853-868.

Murali, T. M. and Kasif, S. (2003). Extracting conserved gene expression motifs from gene expression data. *Pacific Symposium on Biocomputing*. 8, 77-88.

Padilha, V. A. and Campello, R. J. (2017). A systematic comparative evaluation of biclustering techniques. *BMC bioinformatics*, 18(1), 55.

Prelic, A., Bleuler, S., Zimmermann, P. Wil, A., Buhlmann, P., Gruissem, W., Hennig, L., Thiele, L. and Zitzler, E. (2006). A Systematic Comparison and Evaluation of Biclustering Methods for Gene Expression Data Bioinformatics. *Oxford Univ Press*, 22, 1122-1129.

Pontes, B., Girldez, R. and Aguilar-Ruiz, J. S. (2015). Quality measures for gene expression biclusters. *PloS one*, 10(3), e0115497.

Rodriguez-Baena, D. S., Perez-Pulido, A. J. and Aguilar− Ruiz, J. S. (2011). A biclustering algorithm for extracting bit-patterns from binary datasets. *Bioinformatics*, 27(19), 2738-2745.

Teng, L. and Chan, L. (2008). Discovering biclusters by iteratively sorting with weighted correlation coefficient in gene expression data. *Journal of Signal Processing Systems*, 50(3), 267-280.

Turner, H., Trevor B. and Wojtek K. (2005). Improved biclustering of microarray data demonstrated through systematic performance tests. *Computational Statistics and Data Analysis* 48(2), 235-254.

Verma, N. K., Meena, S., Bajpai, S., Singh, A., Nagrare, A. and Cui, Y. (2010, December). *A comparison of biclustering algorithms*. In Systems in Medicine and Biology, 2010 International Conference, 90-97.

Yang, W. H., Dai, D. Q. and Yan, H. (2011). Finding correlated biclusters from gene expression data. *IEEE Transactions on Knowledge and Data Engineering*, 23(4), 568-584.

Yip, K. Y., Cheung, D. W. and Ng, M. K. (2004). Harp: A practical projected clustering algorithm. *IEEE Transactions on Knowledge and Data Engineering*, 16(11), 1387-1397.

Zhu, X., Qiu, J., Xie, M. and Wang, J. (2017). A multi-objective biclustering algorithm based on fuzzy mathematics. *Neurocomputing*, 253, 177-182.

Zhao, H., Wee-Chung Liew, A., Z Wang, D. and Yan, H. (2012). Biclustering analysis for pattern discovery: current techniques, comparative studies and applications. *Current Bioinformatics*, 7(1), 43-55.

EKLER

EK-1. Yapay veri matrislerini oluşturmada kullanılan R program kodları

```
say <- matrix(runif(1000000,0,1),1000,1000)
A <- matrix(0,1000,1000)
for (i in 1:1000)
     for(j in 1:1000) {
          if(say[i,j] < 0.5)          A[i,j]<-rnorm(1,0,1)
            else if(say[i,j] < 0.8)   A[i,j]<-rnorm(1,1,2)
                  else                A[i,j]<-rnorm(1,2,2)
}
A[1:500,1:500]       <- rnorm(500,0,1)* rnorm(500,0,1)+ rnorm(500,0,1)
A[501:800,501:800]  <- rnorm(300,0,1)* rnorm(300,0,1)+ rnorm(300,0,1)
A[801:1000,801:1000]<- rnorm(200,0,1)* rnorm(200,0,1)+ rnorm(200,0,1)
```

Şekil 1.1. Senaryo 1 için üretilen veri matrisinin elemanlarını karıştırmada kullanılan program kodu

```
A2 <- A
A2[1:500,1:500]    <- rnorm(500,0,1)* rnorm(500,0,1)+ rnorm(500,0,1)
A2[401:700,401:700]<- rnorm(300,0,1)* rnorm(300,0,1)+ rnorm(300,0,1)
A2[601:800,601:800]<- rnorm(200,0,1)* rnorm(200,0,1)+ rnorm(200,0,1)
```

Şekil 1.2. Senaryo 2 için üretilen veri matrisinin elemanlarını karıştırmada kullanılan program kodu

```
A3 <- A
A3[1:500,1:500]        <- rnorm(500,0,1)* rnorm(500,0,1)+
                          rnorm(500,0,1)
A3[501:800,501:800]    <- rnorm(300,0,1)* rnorm(300,0,1)+
                          rnorm(300,0,1)
A3[801:1000,801:1000]  <- rnorm(200,0,1)* rnorm(200,0,1)+
                          rnorm(200,0,1)
A3[101:120,101:120]    <- rnorm(400,10,3)
A3[601:620,601:620]    <- rnorm(400,15,3)
A3[901:920,901:920]    <- rnorm(400,20,3)
```

Şekil 1.3. Senaryo 3 için üretilen veri matrisinin elemanlarını karıştırmada kullanılan program kodu

EK-1. (devam) Yapay veri matrislerini oluşturmada kullanılan R program kodları

```
A4 <- A
A4[1:500,1:500]      <- rnorm(500,0,1)* rnorm(500,0,1)+
                        rnorm(500,0,1)
A4[401:700,401:700]  <- rnorm(300,0,1)* rnorm(300,0,1)+
                        rnorm(300,0,1)
A4[601:800,601:800]  <- rnorm(200,0,1)* rnorm(200,0,1)+
                        rnorm(200,0,1)
A4[101:120,101:120]  <- rnorm(400,10,3)
A4[501:520,501:520]  <- rnorm(400,15,3)
A4[701:720,701:720]  <- rnorm(400,20,3)
```

Şekil 1.4. Senaryo 4 için üretilen veri matrisinin elemanlarını karıştırmada kullanılan program kodu

Algoritmalar yapay veri setlerine uygulanmadan önce veri matrisi elemanları rastgele olarak karıştırılmıştır. Örnek olarak birinci senaryoda veri matrisi elemanlarını karıştırma işleminde kullanılan program kodu Şekil 1.1'de gösterilmiştir.

```
mixA<-A
ras<-c(1:1000)
i<-sample(ras,1000)
j<-sample(ras,1000)
for (k in i[1:1000]) for (l in j[1:1000]) { mixA[,c(k,l)]<-
mixA[,c(l,k)] }
i<-sample(ras,1000)
j<-sample(ras,1000)
for (k in i[1:1000]) for (l in j[1:1000]) { mixA[c(k,l),]<-
mixA[c(l,k),] }
```

Şekil 1.5. Senaryo 1 için üretilen veri matrisinin elemanlarını karıştırmada kullanılan program kodu

EK-2. Yapay veri matrisine uygulanan algoritmalardan elde edilen ikili kümelerin değerlendirme ölçülerini hesaplayan R program kodları

Her bir algoritma ile elde edilen ikili kümelere uygulanan değerlendirme ölçüleri fonksiyonları kullanılmıştır. Örnek olarak CC algoritmasıyla elde edilen ikili kümelere uygulanan değerlendirme ölçüleri fonksiyonları sırasıyla aşağıda gösterilmiştir.

```
ccbc1<-bicluster(mixA,ccmixA,1)
ccbc1<-matrix(unlist(ccbc1),ncol=500,byrow = FALSE)
ortccbc1<-mean(ccbc1)
varccbc1<-matrix(0,400,500)
for (i in 1:400) for (j in 1:500) { varccbc1[i,j]<-((ccbc1[i,j]-
ortccbc1)^2)}
varccbc1<-sum(varccbc1)
```

Şekil 2.1. VAR ölçüsünü hesaplayan fonksiyonu gösteren program kodu

```
satccbc1<-matrix(0,400,1)
for (i in 1:400) { satccbc1[i,1]<-mean(ccbc1[i,1:500])}
sutccbc1<-matrix(0,1,500)
for (i in 1:500) { sutccbc1[1,i]<-mean(ccbc1[1:400,i])}
msrccbc1<-matrix(0,400,500)
for (i in 1:400) for (j in 1:500) { msrccbc1[i,j]<-((ccbc1[i,j]-
satccbc1[i,1]-sutccbc1[1,j]+ortccbc1)^2)}
msrccbc1<-mean(msrccbc1)
```

Şekil 2.2. MSR ölçüsünü hesaplayan fonksiyonu gösteren program kodu

```
uiccbc1<-matrix(0,400,1)
uiccbc1r<-matrix(0,400,400)
for (i in 1:400) for (j in 1:400) { uiccbc1r[i,j]<-((ccbc1[i,j]-
sutccbc1[1,j])^2)}
sutuiccbc1<-matrix(0,1,400)
for (i in 1:400) { sutuiccbc1[1,i]<-sum(uiccbc1r[1:400,i])}
for (i in 1:400) { uiccbc1[i,1]<-(1-(sutuiccbc1[1,i]/varccbc1))}
uiccbc1<-mean(uiccbc1)
```

Şekil 2.3. UI ölçüsünü hesaplayan fonksiyonu gösteren program kodu

EK-2. (devam) Yapay veri matrisine uygulanan algoritmalardan elde edilen ikili kümelerin değerlendirme ölçülerini hesaplayan R program kodları

```
ortccbc1<-mean(ccbc1)
satccbc1<-matrix(0,400,1)
for (i in 1:400) { satccbc1[i,1]<-mean(ccbc1[i,1:500])}
sutccbc1<-matrix(0,1,500)
for (i in 1:500) { sutccbc1[1,i]<-mean(ccbc1[1:400,i])}
smsrccbc1<-matrix(0,400,500)
for (i in 1:400) for (j in 1:500)
{smsrccbc1[i,j]<-(((satccbc1[i,1]*sutccbc1[1,j])-
(ccbc1[i,j]*ortccbc1))^2)/(satccbc1[i,1]^2*sutccbc1[1,j]^2)}
smsrccbc1<-mean(smsrccbc1)
```

Şekil 2.4. SMSR ölçüsünü hesaplayan fonksiyonu gösteren program kodu

```
ccbc1t<-matrix(ccbc1,400,500,byrow = T)
rccbc1sat<-matrix(0,400,500)
rccbc1sat<-cor(ccbc1t,method = c("pearson"))
acccbc1sat<-((sum(rccbc1sat)-400)/2)
rccbc1sut<-matrix(0,400,500)
rccbc1sut<-cor(ccbc1,method = c("pearson"))
acccbc1sut<-((sum(rccbc1sut)-500)/2)
acvccbc1sat<-(acccbc1sat-400)/(400^2-400)
acvccbc1sut<-(abs(acccbc1sut)-500)/(500^2-500)
acvccbc1<-if(acvccbc1sat>acvccbc1sut) acvccbc1sat else acvccbc1sut
```

Şekil 2.5. ACV ölçüsünü hesaplayan fonksiyonu gösteren program kodu

```
sccbc1sat<-((sum(rccbc1sat)-400)/2)
sccbc1sut<-((sum(rccbc1sut)-500)/2)
scsccbc1sat<-(1-(acvccbc1sat/400))
scsccbc1sut<-(1-(acvccbc1sut/500))
scsccbc1<-if(scsccbc1sat<scsccbc1sut) scsccbc1sat else scsccbc1sut
```

Şekil 2.6. SCS ölçüsünü hesaplayan fonksiyonu gösteren program kodu

EK-2. (devam) Yapay veri matrisine uygulanan algoritmalardan elde edilen ikili kümelerin değerlendirme ölçülerini hesaplayan R program kodları

```
ccbc1t<-matrix(ccbc1,400,500,byrow = T)
rccbc1sat<-matrix(0,400,500)
rccbc1sat<-cor(ccbc1t,method = c("spearman"))
acccbc1sat<-((sum(rccbc1sat)-400)/2)
rccbc1sut<-matrix(0,400,500)
rccbc1sut<-cor(ccbc1,method = c("spearman"))
acccbc1sut<-((sum(rccbc1sut)-500)/2)
asrccbc1sat<-(acccbc1sat-400)/(400^2-400)
asrccbc1sut<-(abs(acccbc1sut)-500)/(500^2-500)
asrccbc1<-if(asrccbc1sat>asrccbc1sut) 2*asrccbc1sat else 2*asrccbc1sut
```

Şekil 2.7. ASR ölçüsünü hesaplayan fonksiyonu gösteren program kodu

```
sbmccbc1sat<-scsccbc1sat
sbmccbc1sut<-scsccbc1sut
sbmccbc1<-1*sbmccbc1sat*1*sbmccbc1sut
```

Şekil 2.8. SBM ölçüsünü hesaplayan fonksiyonu gösteren program kodu

```
sutccbc1<-matrix(0,1,500)
for (i in 1:500) { sutccbc1[1,i]<-mean(ccbc1[1:500,i])}
veccbc1<-matrix(0,400,500)
for (i in 1:400) for (j in 1:500) { veccbc1[i,j]<-(abs(ccbc1[i,j]-
sutccbc1[1,j]))}
veccbc1<-mean(veccbc1)
```

Şekil 2.9. VE ölçüsünü hesaplayan fonksiyonu gösteren program kodu

ÖZGEÇMİŞ

Kişisel Bilgiler

Soyadı, adı	: KOCATÜRK, Ahmet
Uyruğu	: T.C.
Doğum tarihi ve yeri	: 05.07.1989, Malatya
Medeni hali	: Evli
Telefon	: 0 (543) 353 63 93
e-mail	: ahmetkocaturk@gazi.edu.tr

Eğitim

Derece	Eğitim Birimi	Mezuniyet tarihi
Yüksek lisans	Gazi Üniversitesi /İstatistik	Devam Ediyor
Lisans	Yıldız Teknik Üniversitesi /İstatistik	2013
Lise	Malatya Turgut Özal Anadolu Lisesi	2007

İş Deneyimi

Yıl	Yer	Görev
2014-Halen	Gazi Üniversitesi	Araştırma Görevlisi
2013-2014	Erzurum Teknik Üniversitesi	Araştırma Görevlisi

Yabancı Dil

İngilizce

Yayınlar

Altunkaynak, B. ve Kocatürk, A. (2016, Haziran). *Veri Madenciliği ile Daire Fiyatlarına Etki Eden Değişkenlerin Belirlenmesi: Ankara İli Örneği*. 17. Uluslararası Ekonometri, İstatistik ve Yöneylem Araştırması Sempozyumunda Sunuldu, Sivas.

Kocatürk, A., Altunkaynak, B. ve Örkcü, H.H. (2017, Ekim). *Türkiye'de Sektörlerin Ölümlü Kaza Türlerine Göre Kümelenmesi: İkili Kümeleme Yöntemi*. 18. Uluslararası Ekonometri, İstatistik ve Yöneylem Araştırması Sempozyumunda Sunuldu, Trabzon.

Özkan Aksu, E., Küpeli, M. ve Kocatürk, A. (2017, Ekim). *Veri Zarflama Analizi Etkinlik Skorlarına Dayalı Karar Ağacı Sınıflandırması: Ar-Ge Örneği*. 18. Uluslararası Ekonometri, İstatistik ve Yöneylem Araştırması Sempozyumunda Sunuldu, Trabzon.

Kocatürk, A., Altunkaynak, B., Örkcü, H.H. ve Arslan, R. (2017, Ekim). *Suç Verilerinin Analizinde Kullanılan Veri Madenciliği Yöntemleri ve Kaçakçılık Suçlarıyla İlgili Bir Uygulama.* 2. Uluslararası Akdeniz Bilim ve Mühendislik Sempozyumunda sunuldu, Adana.

Kocatürk, A., Bodur, S. and Gül, H.H. (2017, December). *European Union Countries and Turkey's Waste Management Performance Analysis with Malmquist Total Factor Productivity Index.* Paper presented at the 10th International Statistics Congress, Ankara.

Hobiler

Futbol, Masa Tenisi, Yüzme.

GAZİ GELECEKTİR...