ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH

DOI: 10.5336/biostatic.2018-60153

Gen Açıklama Verileri İçin İkili Kümeleme Algoritmalarının Karşılaştırılması ve Uygulanması

Comparison and Application of Biclustering Algorithms for Gene Expression Data

Ahmet KOCATÜRK,[a]
Bülent ALTUNKAYNAK[a]

[a]İstatistik Bölümü,
Gazi Üniversitesi Fen Fakültesi,
Ankara

Received: 13.02.2018
Received in revised form: 21.05.2018
Accepted: 22.05.2018
Available Online: 07.09.2018

Correspondence:
Ahmet KOCATÜRK
Gazi Üniversitesi Fen Fakültesi,
İstatistik Bölümü, Ankara,
TÜRKİYE/TURKEY
ahmetkocaturk@gazi.edu.tr

ÖZET Amaç: Gen açıklama verilerinde benzer ifade yapılarına göre gen gruplarını belirlemek hastalık teşhisleri için oldukça önemlidir. Gen açıklama verileri satırları genlerden sütunları ise koşullar veya örneklemlerden oluşan büyük bir veri matrisidir. Bu veriler için satır ve sütunların aynı anda kümelenmesini sağlayan ikili kümeleme yöntemleri son yıllarda yaygın olarak kullanılmaktadır. Literatürde farklı gen açıklama verileri için birçok ikili kümeleme algoritmaları önerilmiştir. Ancak bu algoritmaların performanslarını karşılaştırmaya yönelik çalışmalar sınırlıdır. Bu çalışmanın amacı, gerçek veri setlerine uygulanan farklı ikili küme algoritmalarının performanslarını karşılaştırmaktır. **Gereç ve Yöntemler:** İyi bilinen üç gerçek veri kümesi üzerinden CC, Bimax, Plaid, Quest, Spectral and Xmotif ikili kümeleme algoritmaları karşılaştırılmıştır. Bu veri kümeleri; 2884 x 17 boyutlu maya verisi, 4096 x 96 boyutlu lenf verisi ve 1080 x 77 boyutlu fare verisidir. Tüm analizler R Project v3.4.4 programı ile yapılmış elde edilen sonuçlar farklı değerlendirme ölçüleri ve ısı grafikleri kullanılarak karşılaştırılmıştır. **Bulgular:** Isı grafiklerine göre maya ve insan verisi için Bimax ve Xmotif, fare verisi için Quest ve Xmotif algoritmaları en büyük hacme sahip ikili kümeleri elde etmişlerdir. Maya ve insan verisinde değerlendirme ölçülerinin çoğunda CC ve Bimax algoritması en iyi sonucu vermiştir. Fare verisinde ise Spectral algoritması en iyi sonucu vermiştir. Genel olarak bakıldığında CC, Bimax ve Spectral algoritmalarının Plaid, Quest ve Xmotif algoritmalarına göre daha iyi sonuçlar verdiği söylenebilir. **Sonuç:** Bu çalışmada, gen açıklama verilerinin yapıları farklı olduğundan ve çalışmada kullanılan değerlendirme ölçülerine göre algoritmaların performansları farklılık göstermiştir.

Anahtar Kelimeler: İkili kümeleme; gen açıklama; ısı grafiği; değerlendirme ölçüleri

ABSTRACT Objective: Identifying gene groups according to similar expression patterns in gene expression data is crucial for disease diagnosis. Gene expression data are large data matrices which rows are composed of genes and columns are composed of conditions or samples. Biclustering methods for clustering simultaneously rows and columns for this data have been widely used in recent years. Several biclustering algorithms are proposed for different gene expression data in the literature. However, comparative studies the performance of algorithms are limited. The purpose of this work is to compare the performance of different biclustering algorithms applied to real datasets. **Material and Methods**: CC, Bimax, Plaid, Quest, Spectral and Xmotif algorithms have been compared with three well known real data sets. These data sets are; 2884 x 17 size yeast data, 4096 x 96 size lymph data and 1080 x 77 size mice data. All analyzes were performed with the R Project v3.4.4 program and the results obtained were compared using different evaluation measures and heatmaps. **Results:** According to the heatmaps, Bimax and Xmotif for yeast and human data, Quest and Xmotif for the mice data have obtained biclusters with the largest size. Most of the evaluation measures in yeast and human data give the best results in the CC and Bimax. In mice data, Spectral gives the best result. In general, CC, Bimax and Spectral give better results than Plaid, Quest and Xmotif. **Conclusion:** In this study, the gene expression data are different and the performance of the algorithms differed according to the evaluation measures used in the study.

Keywords: Biclustering; gen expression; heatmap; evaluation measures

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 zorlaştırmaktadır. Bu nedenle literatürde ikili kümeleme için birçok sezgisel algoritma önerilmiştir. Bunlardan bazıları Cheng ve Church (2000) tarafından önerilen CC algoritması, Prelic ve ark. (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 ve ark. (2003) tarafından önerilen Spectral algoritması olarak verilebilir.[1-5] Bununla birlikte, literatürde bu algoritmaların karşılaştırılmasına ilişkin çalışmalar sınırlıdır.

Prelic ve ark. (2006), ikili kümeler içeren yapay gen açıklama verileri üretmişler ve CC, Bimax, OPSM, SAMBA, Xmotif ve ISA ikili kümeleme algoritmalarını karşılaştırmışlardır. Gerçek veri üzerinden yaptıkları karşılaştırmada ISA, SAMBA ve OPSM algoritmalarının daha iyi sonuçlar verdiğini ifade etmişlerdir.[2] 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 kümesi kullanmışlardır. Bu üç algoritmanın diğerlerine göre daha iyi sonuçlar verdiğini göstermişlerdir.[6] Bhattacharya ve ark. (2012) farklı gen açıklama verileri için kümeleme ve ikili kümeleme algoritmalarını 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.[7] Li ve ark. (2012), Bimax, FABIA, ISA, QUBIC ve SAMBA ikili kümeleme algoritmalarını iki farklı veri setini, GDS1620 ve Pathway, kullanarak karşılaştırmışlardır. GDS1620 verisi için ISA algoritması en iyi sonucu verirken Pathway verisi için en iyi skorlar Bimax algoritmasından elde edilmiştir.[8] Zhao ve ark. (2012), ikili kümeleme algoritmalarını karşılaştırmış ve farklı senaryolar için Bayesçi ikili kümeleme algoritmasının iyi performans gösterdiği sonucuna ulaşmışlardır.[9] Eren ve ark. (2012) ile Padilha ve Campello (2017) de farklı veri setleri için farklı sonuçlar elde etmişlerdir.[10,11] Yapılan çalışmaların özeti Tablo 1'de gösterilmiştir.

Bu çalışmada amaç, ikili kümeleme algoritmalarının daha geniş bir şekilde karşılaştırılmasını yapmaktır. Literatürde karşılaştırma için kullanılan algoritmaların yapısı ikili küme türlerine göre farklılık göstermektedir. Çalışmanın amacına uygun olarak kullanılan CC, Bimax, Plaid, Quest, Spectral ve Xmotif ikili

TABLO 1: İkili küme algoritmalarının karşılaştırılmasına ilişkin çalışmalar.

Yıllar	Yazarlar	Algoritmalar	Kullanılan Veri	Sonuç
2006	Prelic ve ark.	CC, Bimax, OPSM, SAMBA, Xmotif ve ISA	Yapay ve Gerçek	OPSM, SAMBA ve ISA
2011	Das ve Idicula	MSRT, MSRDT ve ISIMSRDT CC, SEBI, FLOC ve DBF	Gerçek	MSRT, MSRDT ve ISIMSRDT
2012	Bhattacharya ve ark.	delta (CC), OPSM, SAMBA, ISA, Bimax ve bioNMF	Gerçek	SAMBA
2012	Li ve ark.	Bimax, FABIA, ISA, QUBIC ve SAMBA	Gerçek	ISA ve Bimax
2012	Zhao ve ark.	CC, ISA, OPSM, Bimax, BBC ve MSBE	Yapay ve Gerçek	BBC
2012	Eren ve ark.	BBC, Bimax, CC, COALESCE,CPB, FABIA, ISA, OPSM, Plaid, QUBIC, Spectral ve Xmotif	Yapay ve Gerçek	Her bir senaryo için farklı sonuçlar elde edilmiştir.
2017	Padilha ve Campello			

küme algoritmaları varyans ölçüsü (VAR), ortalama karesel artık skoru (MSR), uygunluk indeksi (RI), 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 üç farklı gerçek veri kullanılarak karşılaştırılmıştır.

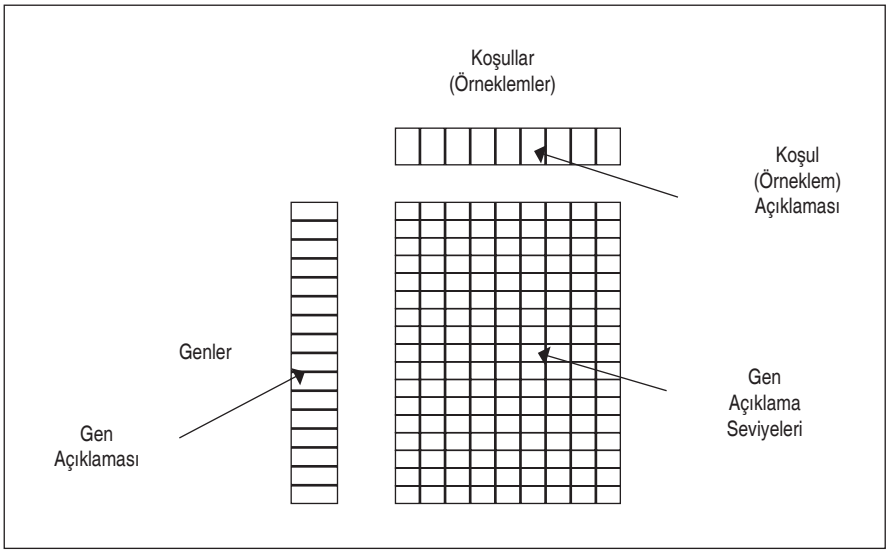
GEREÇ VE YÖNTEMLER

İKİLİ KÜMELEME YÖNTEMLERİ

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.[12] Gen açıklama veri matrisleri birçok koşullardan (örneklemden) elde edilen binlerce genin açıklamasını içeren çok boyutlu verilerdir. Bir gen açıklama verisi Şekil 1'deki gibi gösterilebilir.

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 kümelemek gen örüntüleri ortaya çıkarmada daha önemli hale gelmiştir. Bu nedenle ikili kümeleme algoritmaları geliştirilmiştir. Bir ikili küme yapısı eşitlik (1) ile ifade edilebilir.

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



ŞEKİL 1: Gen açıklama verisi.

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 eşitlik (2), (3) ve (4) ile bulunabilir.

$$b_{Ij} = \frac{1}{|I|}\sum_{i=1}^{|I|} b_{ij} \quad (2) \qquad b_{iJ} = \frac{1}{|J|}\sum_{j=1}^{|J|} b_{ij} \quad (3) \qquad b_{IJ} = \frac{1}{|I||J|}\sum_{i=1}^{|I|}\sum_{j=1}^{|J|} b_{ij} \quad (4)$$

Burada, b_{Ij} j inci sütunun ortalaması, b_{iJ} i inci satırın ortalaması ve b_{IJ} genel ortalamayı göstermektedir. $|I|$ ve $|J|$ değerleri ise sırasıyla satır ve sütun sayılarını belirtir.

İkili Küme Türleri

Gen verilerinin içerisinde aldığı değerler bakımından farklılaşan 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 (coherent) oluşan bir ikili küme yapısı da mümkündür. Bu ikili küme türleri aşağıda kısaca 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.

Toplamsal ve çarpımsal ikili küme türleri

Aguilar-Ruiz (2005), gen verileri içerisindeki ikili kümeler için sabit ikili kümelerden farklı yapılar önermiştir.[13] 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. 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 coherent ikili küme modeli $(b_{ij} = \pi_i \times \alpha_j + \beta_j)$ oluşturulur.

Farklı yapılardaki ikili küme örnekleri Şekil 2'de gösterilmiştir.

Gen açıklama verileri farklı yapılarda ikili küme türleri barındırabileceği için çok sayıda ikili küme algoritmaları geliştirilmiştir. En yaygın kullanılan CC, Bimax, Plaid, Quest, Spectral ve Xmotifs ikili kümeleme algoritmaları aşağıda anlatılmıştır.

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)

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

(d)

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

(e)

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

(f)

ŞEKİL 2: İkili küme yapıları (a) sabit (b) sabit satırlar (c) sabit sütunlar (d) toplamsal (e) çarpımsal (f) coherent

CC Algoritması

Cheng ve Church (2000)'a göre; bir ikili kümenin anlamlı olması, ikili kümenin eşitlik (5) ile gösterilen Ortalama Karesel Artık Değeri (H)'nin minimum olmasına bağlıdır. İkili kümedeki b_{ij} elemanı için artık değer $b_{ij} - b_{iJ} - b_{Ij} + b_{IJ}$ formülü ile hesaplanmaktadır.[1]

$$H(I,J) = \frac{1}{|I||J|}\sum_{i\in I, j\in J}(b_{ij} - b_{iJ} - b_{Ij} + b_{IJ})^2 \quad (5)$$

Bir ikili kümenin uygunluğu, küme içerisindeki elemanların birbirine olan benzerliğini ile belirlenir. H değerinin küçük olması küme elemanlarının benzerliğinin yüksek olduğunu, H değerinin büyük olması ise küme elemanların benzerliğinin düşük olduğunu ifade eder. CC algoritmasında, en büyük kabul edilebilir ortalama kare artığı sabiti, δ belirlenerek H değeri ile karşılaştırılır. H değeri δ sabitinden küçük olduğunda elde edilen ikili küme δ-ikili küme (δ-bicluster) olarak ifade edilir. CC algoritmasının temel amacı, minimum H değerine sahip maksimum büyüklükte ikili kümeler bulmaktır. Algoritmanın adımları aşağıdaki gibidir:

Girdiler:

B : Veri matrisi

δ *:* En büyük kabul edilebilir ortalama kare artığı skoru, $\delta \geq 0$.

α *:* Çoklu satır silme için eşik değer, $\alpha > 0$.

Adımlar:

1. $A_{IJ} = B$.
2. b_{iJ}, b_{Ij} ve b_{IJ} değerlerini hesapla.
3. $H(I,J)$ değerini hesapla.
4. Eğer $H(I,J) \leq \delta$ ise $B = A_{IJ}$ matrisini ikili küme olarak al aksi halde sonraki adıma git.
5. Aşağıdaki koşulu sağlayan $i \in I$ satırlarını sil.

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

6. b_{iJ}, b_{Ij}, b_{IJ} ve $H(I,J)$ değerlerini hesapla.
7. Aşağıdaki koşulu sağlayan $j \in J$ sütunlarını sil.

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

8. Eğer tüm satır ve sütun silme işlemi bittiyse sonlandır aksi halde A_{IJ} matrisini B'den çıkarıp geriye kalan matris için adım 1'e dön.

Bimax Algoritması

Gen açıklama verilerinde belli bir eşik değere göre iki değer alan veri matrislerinin oluşturulmasıyla Bimax yöntemi kullanılabilmektedir. İki değer alan veri matrislerinde 1'lerden oluşan alt matrislerin aranmasına dayalı olan Bimax algoritması hızlı ve basit bir yaklaşımdır.[2,14] Bu yöntemin diğer algoritmalara göre önemli bir avantajı aykırı değerlerden etkilenmemesidir. Algoritmanın işleyişi aşağıdaki gibi verilebilir.

Girdi:

A : M×N boyutlu veri matrisi

Adımlar:

1. *A* 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.

Plaid Algoritması

Bu model, veri matrisinin temel köşegenleri boyunca ikili kümeler oluşturur. Bu algoritma veri matrislerindeki katmanların toplamını modellemektedir ve hataların minimum yapılmasına dayalıdır.[15] Plaid model algoritmasında satır ve sütunları bir araya getirdikten sonra büyük veri matrisi içerisinde dikdörtgensel bölge oluşturularak bir alt matris şeklinde kümeleme yapılır. Her küme kendi içerisinde birbirine yakın değerlere sahip olur. Cebirsel olarak bloklar içerisindeki genler şu şekilde ifade edilebilir:

$$Y_{ij} = \mu_0 + \sum_{k=1}^{K} \mu_k r_{ik} c_{jk} \quad (6)$$

Eşitlik 6'da μ_0 arkaplan rengi ve μ_k ise k bloğunda bulunan rengi temsil eder. Eğer k bloğunda i. gen mevcutsa r_{ik} değeri 1'e eşit olur, aksi takdirde 0 olur. Benzer şekilde j. örnek k bloğu içerisinde mevcutsa c_{jk} değeri 1'e eşit olur, aksi takdirde 0 olur.[3] Algoritmanın adımları aşağıda verilmiştir:

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 sıralı 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.

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

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

Xmotif Algoritması

Xmotif adı verilen benzer satır ve sütunları kümelemek için Murali ve Kasif (2003) tarafından önerilmiştir. En geniş kümeleri bulmak için tekrarlayıcı bir algoritma yapısı vardır. Algoritma en geniş kümeyi bulduktan sonra tüm veri setinden bu kümeyi çıkararak geriye kalan veri matrisinden işlemleri tekrarlar.[4] Bu algoritmayla birbiriyle örtüşen kümler tespit edilebilir. Algoritmanın adımları aşağıda verilmiştir:

Girdi:

A : 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* örneğinde *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 *Cij* axn boyutlu örnekten daha küçük ise (C_{ij}, G_{ij}) çiftini çıkar.

6. $|G_{ij}|$ maksimum olana kadar *(C , G)* motifini çalıştır.

Quest Algoritması

Veri matrisindeki değerlerin sıralama ölçme düzeyinde olduğu durumda kullanılmaktadır.[4] Bu nedenle gen verilerine uygulamadan önce veri matrisinde uygun dönüşümün yapılması gerekir. Algoritma, önceden belirlenmiş bir parametreyle (d) oluşturulan değer aralığındaki benzer değerleri arar *(Kaiser S. Biclustering: methods, software and application. Doctoral dissertation. Münih: Ludwig Maximilian University; 2011. p.157).* Algoritmanın adımları Xmotif algoritmasıyla aynıdır.

Spectral Algoritması

Kluger ve ark. (2003) tarafından geliştirilen bu algoritma gen ve koşulların bir alt kümesini dama tahtası yapısına benzer olarak tanımlamaya çalışır.[5] Bu algoritma tek değer ayrıştırmasına dayalıdır. Spectral ikili kümelemede, normalleştirilmiş gen açıklama verisinin tek değer ayrıştırması eşitlik (7) ve (8) ile gösterilen iki farklı özvektör probleminin çözümü ile sağlanabilir.

$$A^T A\omega = \lambda\omega \quad (7) \qquad AA^T z = \lambda z \quad (8)$$

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 aşağıdaki gibi verilebilir.

Girdi:

A : NxM boyutlu veri matrisi

Adımlar:

1. Bağımsız satır ve sütunların ölçeklenmesi (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.

İKİLİ KÜME DEĞERLENDİRME ÖLÇÜLERİ

İkili kümeleme algoritmalarından 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 verilmiştir.

Varyans Ölçüsü

Hartigan (1972) tarafından ikili kümeleri değerlendirme ölçüsü olarak küme içi varyans kullanılmıştır.[16] İkili kümelerin varyanslarının toplamına dayalı olan bu ölçü eşitlik (9) ile gösterilmiştir.

$$VAR(B) = \sum_{i=1}^{|I|} \sum_{j=1}^{|J|} (b_{ij} - b_{IJ})^2 \quad (9)$$

Bu ölçü, sabit elemanlı ikili kümelerin tespit edilmesinde etkili olmaktadır.[17]

Ortalama Karesel Artık Skoru

İkili kümelerin kalitesini ölçmek için kullanılan ortalama karesel artık skoru (MSR) yaygın kullanılan bir yaklaşımdır.[1] Bu ölçü eşitlik (10) ile gösterilmiştir.

$$MSR(B) = \frac{1}{|I||J|} \sum_{i=1}^{|I|} \sum_{j=1}^{|J|} (b_{ij} - b_{iJ} - b_{Ij} + b_{IJ})^2 \quad (10)$$

Bu ölçü ikili kümede yer alan satır ve sütun değerlerinin tutarlılığını incelemektedir. Bu değer küçüldükçe ikili kümelemenin başarısı artar. Bu ölçü veri matrisindeki değerlerin ölçeklendirilmesine aşırı duyarlıdır ve bu durumda ikili kümenin değerlendirmesinde yetersiz olduğu görülmüştür.[13]

Uygunluk İndeksi

Yip ve ark. (2004) tarafından önerilen uygunluk indeksi (RI) ölçüsü ikili kümenin sütunlarının uygunluğunu inceler. Sütun elemanlarının uygunluklarının toplamına dayalı olan bu ölçü özellikle sabit değerli ikili kümelerde iyi sonuçlar vermektedir.[18] Her bir sütun değeri için uygunluk indeksi eşitlik (11) ile gösterilmiştir.

$$R_{Ij} = 1 - \frac{\sigma_{Ij}^2}{\sigma_j^2} \quad (11)$$

Burada σ_{Ij}^2 (yerel varyans) ikili kümedeki sütun değerlerinin varyansı ve σ_j^2 (genel varyans) tüm veri setindeki sütun değerlerinin varyansıdır.

ÖLÇEKLENEN ORTALAMA KARESEL ARTIK SKORU

Mukhopadhyay ve ark. (2009) tarafından geliştirilen ölçeklenen ortalama karesel artık skoru (SMSR) çarpımsal ikili kümeler için uygun bir değerlendirme ölçüsüdür.[19] MSR'nin çarpımsal ikili kümeler için ortaya çıkan dezavantajını gidermek için geliştirilmiştir. Bu ölçü eşitlik (12) ile gösterilmiştir.

$$SMSR(B) = \frac{1}{|I||J|}\sum_{i\in I, j\in J}\frac{(b_{ij}\times b_{Ij} - b_{ij}\times b_{IJ})^2}{b_{ij}^2 \times b_{Ij}^2} \quad (12)$$

Bu ölçü toplamsal ikili küme türlerinde etkin değildir.

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.[20] 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 (13) ve (14) ile hesaplanır.

$$T_k(b) = \frac{1}{|I(b)|}\sum_{i=1, i\in b}^{I(b)} {b_{iJk}}^2 - \frac{MSR_k(b)}{J_k(b)} \quad (13)$$

$$B_k(b) = \frac{1}{|J_k(b)|}\sum_{j=1, j\in b}^{J_k(b)} {b_{Ijk}}^2 - \frac{MSR_k(b)}{I(b)} \quad (14)$$

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. $MSR_k(b)$ değeri, ikili kümedeki ve ikili küme dışındaki ortalama karesel artık skorunu 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*) eşitlik (15) ile hesaplanır.

$$CKSB(b) = LOG\left(\frac{max(T_1(b)+\alpha, B_1(b)+\alpha}{max(T_2(b)+\alpha, B_2(b)+\alpha}\right) \quad (15)$$

Burada α değeri 0 ile 1 arasında belirlenen bir ölçek faktörüdür. *CKSB* skoru ne kadar yüksekse ikili kümeler o kadar uygundur.

Korelasyon Tabanlı Ölçüler

Korelasyon katsayısı geleneksel kümeleme yöntemlerinde olduğu gibi farklı gen açıklama verilerinin bulunduğu analiz türlerinde de yoğun olarak kullanılmaktadır. Gen ifade değerlerinin bulunduğu 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.[17]

Ortalama korelasyon ölçüsü

i_1 ve i_2 ikili küme içerisindeki herhangi iki satırı göstermek üzere bu katsayı eşitlik (16) ile hesaplanır.

$$r(i_1,i_2)=\frac{\sum_{j=1}^{|J|}(b_{i_1j}-b_{i_1J})(b_{i_2j}-b_{i_2J})}{\sqrt{\sum_{j=1}^{|J|}(b_{i_1j}-b_{i_1J})^2\sum_{j=1}^{|J|}(b_{i_2j}-b_{i_2J})^2}} \quad (16)$$

Burada b_{i_1j} ve b_{i_2j} değerleri, j sütununda bulunan i_1 ve i_2 satır eleman değerlerini, b_{i_1J} ve b_{i_2J} değerleri ise i_1 ve i_2 satır elemanlarının ortalamasını göstermektedir. Bu katsayı her bir satır (gen) çifti için hesaplanır ve eşitlik (17) verilen ortalama korelasyon ölçüsü elde edilir.

$$AC(B)=\frac{\sum_{i_1=1}^{|I|-1}\sum_{i_2=i_1+1}^{|I|}r(i_1,i_2)}{\binom{|I|}{2}} \quad (17)$$

Burada korelasyon değeri sadece satırlar arasındaki ilişkiyi ölçtüğü için sütunlar arasındaki ilişkiyi dikkate almamaktadır. Bu problemi ortadan kaldırmak için Yang ve ark. (2011) ve Teng ve Chan (2008), çalışmalarında sütunlar için de korelasyon değerini kullanmıştır.[21,22]

Bir başka ortalama korelasyon ölçüsü Teng ve Chan (2008) tarafından önerilen eşitlik (18) ile gösterilmiştir.

$$ACV(B)=max\left\{\frac{\sum_{i_1=1}^{|I|}\sum_{i_2=1}^{|I|}|r(i_1,i_2)|-|I|}{|I|^2-|I|},\frac{\sum_{j_1=1}^{|J|}\sum_{j_2=1}^{|J|}|r(j_1,j_2)|-|J|}{|J|^2-|J|}\right\} \quad (18)$$

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.

Alt matris korelasyon skoru

Yang ve ark. (2011), hem satırlar hem de sütunlar üzerinden bir değerlendirme ölçüsü geliştirmek amacıyla korelasyon katsayısını kullanmıştır.[21] Satırlar ve sütunlar için hesaplanan korelasyon katsayıları sırasıyla eşitlik (19) ve (20)'deki gibi tanımlanır.

$$S_{satır}=min_{i_1\in I}(S_{i_1J}),\quad S_{i_1J}=1-\frac{\sum_{i_2\neq i_1,i_2\in I}|r(x_{i_1J},x_{i_2J})|}{|I|-1} \quad (19)$$

$$S_{sütun}=min_{j_1\in J}(S_{Ij_1}),\quad S_{Ij_1}=1-\frac{\sum_{j_2\neq j_1,j_2\in J}\left|r(x_{Ij_1},x_{Ij_2})\right|}{|J|-1} \quad (20)$$

Burada $r(x_{i_1J},x_{i_2J})$ satırlar arasındaki, $r(x_{Ij_1},x_{Ij_2})$ 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 (*SCS*) elde edilir. Yang ve ark. (2011), belirli bir eşik değeri parametresi δ (delta) belirleyerek $S(I,J)<\delta$ koşulunu sağlayan ikili kümelere δ-corbicluster ismini vermiştir.[21]

Ortalama Spearman korelasyon değeri

Ayadi ve ark. (2009) tarafından önerilen bu ölçü iki değişkene atanan sıra sayıları arasındaki korelasyonu kullanır.[23] Bu değer *PCC*'nin aksine bir dağılım varsayımı gerektirmemektedir. Bununla beraber *PCC*'ye göre daha az hassas olduğu açıktır. Ortalama Spearman korelasyon değeri (*ASR*) eşitlik (21)'de verilmiştir.

$$ASR(B) = 2 \times max\left\{\frac{\sum_{i_1 \in I}\sum_{i_2 \geq i_1+1, i_2 \in I} r_s(i_1,i_2)}{|I|(|I|-1)}, \frac{\sum_{j_1 \in J}\sum_{j_2 \geq j_1+1, j_2 \in J} r_s(j_1,j_2)}{|J|(|J|-1)}\right\} \quad (21)$$

Burada $r_s(i_1, i_2)$ değeri iki satır arasındaki, $r_s(j_1, j_2)$ değeri iki sütun arasındaki Spearman korelasyonunu göstermektedir.

Spearman ikili küme ölçüsü

Flores ve ark. (2013) tarafından önerilen bu ölçü, toplamsal ve çarpımsal ikili kümelerin değerlendirilmesinde kullanılır.[24] Bu ölçümün hesaplanması iki aşamadan oluşur. Birinci aşama, Spearman korelasyon katsayılarının hesaplanması, ikinci aşama ise bu katsayıların kullanılmasıyla eşitlik (22)'deki değerin elde edilmesidir.

$$SBM(B_{ij}) = \alpha(B_{ij}) \times r^G_{B_{IJ}} \times \beta(B_{ij}) \times r^C_{B_{IJ}} \quad (22)$$

Burada $r^G_{B_{IJ}}$ ve $r^C_{B_{IJ}}$ değerleri ikili küme içerisindeki satırlarda ve sütunlarda gözlenen eğilimlerin ifadesini özetler. Bu değerler eşitlik (23) ve (24)'deki gibi hesaplanır.

$$r^G_{B_{IJ}} = \frac{2}{|I|\times(|I|-1)}\sum_{i_1=1}^{|I|}\sum_{i_2=i_1+1}^{|I|}|r_s(i_1,i_2)| \quad (23)$$

$$r^C_{B_{IJ}} = \frac{2}{|J|\times(|J|-1)}\sum_{j_1=1}^{|J|}\sum_{j_2=j_1+1}^{|J|}|r_s(j_1,j_2)| \quad (24)$$

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

Eşitlik (22)'de gösterilen $\alpha(B_{ij})$ ve $\beta(B_{ij})$ değerleri güvenilirlik katsayılarını gösterir. Bu değerler hem satırlardaki hem de sütunlardaki gözlenen eğilimlerin etkisini ağırlıklandırır.

STANDARTLAŞTIRILMIŞ ÖLÇÜLER

İkili kümelemede etkili olan önemli bir nokta veri matrisi içerisindeki değerlerin aralığıdır. 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 (25)'de verildiği gibi yapılır.

$$\hat{b}_{ij} = \frac{b_{ij}-\mu_{g_i}}{\sigma_{g_i}}, \quad 1 \leq i \leq |I|, \quad 1 \leq j \leq |J| \quad (25)$$

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

Standartlaştırılmış değerlere dayalı olarak önerilen ölçülerden biri sanal hata (*VE*) değerlendirme ölçüsüdür. *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 (26)'daki gibi bulunur.

$$VE(B) = \frac{1}{|I||J|}\sum_{i=1}^{|I|}\sum_{j=1}^{|J|}|\hat{b}_{ij} - b_{Ij}| \quad (26)$$

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.

BULGULAR

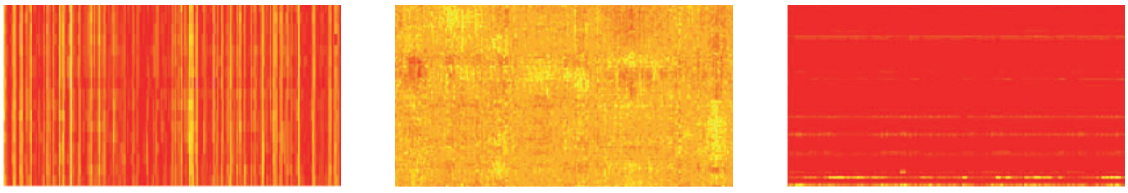
Bu bölümde üç farklı veri seti kullanılarak algoritmalardan elde edilen ikili kümelerin karşılaştırılması yapılmıştır. İlk olarak 2884 gen ve 17 koşuldan oluşan "Saccharomyces cerevisiae" maya veri matrisi kullanılmıştır. İkinci olarak insan lenf hücrelerinin gen ifadesini içeren 4096 gen ve 96 koşuldan 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 *1080×77* boyutlu veri matrisi kullanılmıştır. Veri setin 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.[25] 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 karşılaştırılmıştır.

Bimax ve Xmotif algoritmalarının uygulanması için maya, insan ve fare verilerine dönüşüm uygulanmıştır. Bimax algoritması için dönüşüm uygulanırken eşik değeri olarak her verinin ortalaması alınmıştır. Xmotif algoritmasının uygulanması için üç veri seti de kesikli değerler alacak şekilde dönüştürülmüştür. Maya, insan ve fare veri matrislerinin ısı grafikleri Şekil 3'de gösterilmiştir.

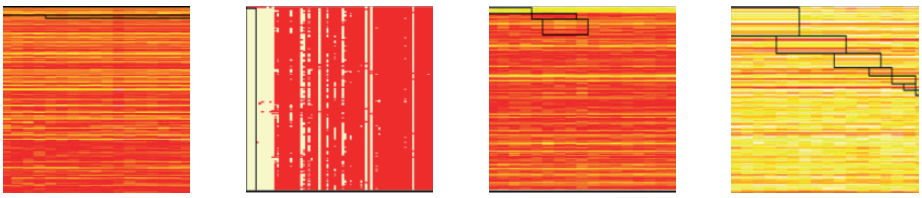
Isı grafiğindeki renklendirme matris içerisindeki sayıların değerine göre belirlenmiştir. En küçük değerdeki sayılar koyu renkten (kırmızıdan) en büyük değerdeki sayılar açık renge (beyaza) göre renk geçişini göstermektedir. Maya verisi için 0 ile 595 arasındaki sayıların değerlerine göre, insan verisi için -749 ile 999 arasındaki sayıların değerlerine göre ve fare verisi için -0.06 ile 8.48 arasındaki sayıların değerlerine göre renk geçişi olmuştur. Fare verisindeki değerler birbirine çok yakın olduğu için maya ve insan verisine göre ısı grafiğinde daha koyu renkte olduğu görülmektedir.

Maya verisi için CC algoritması ile temel ikili küme *116×17* boyutunda elde edilmiştir. Bimax için ikili küme boyutu *1080×3*, Plaid için *254×4* ve Xmotif için *423×6* dır.



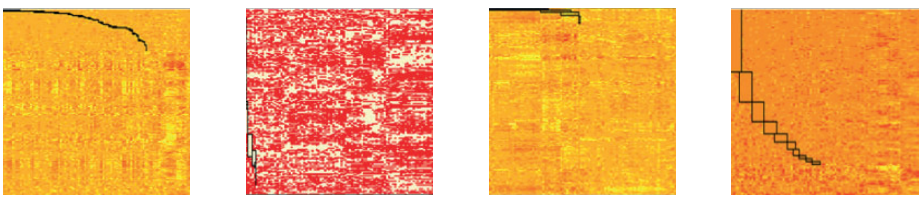
(a) (b) (c)

ŞEKİL 3: Orijinal veri için ısı grafikleri (a) maya verisi (b) insan verisi (c) fare verisi



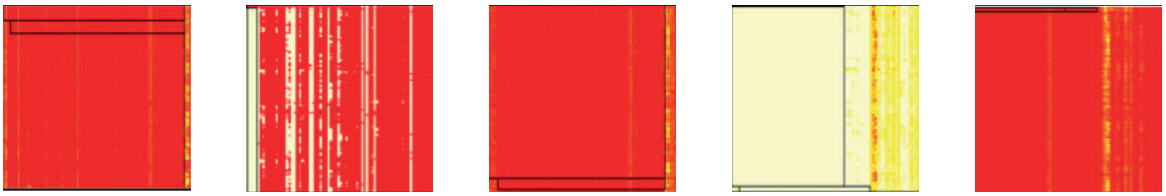
(a) (b) (c) (d)

ŞEKİL 4: Maya verisi için elde edilen ısı grafikleri (a) CC (b) Bimax (c) Plaid (d) Xmotif



(a) (b) (c) (d)

ŞEKİL 5: İnsan verisi için elde edilen ısı grafikleri (a) CC (b) Bimax (c) Plaid (d) Xmotif



ŞEKİL 6: Fare verisi için elde edilen ısı grafikleri (a) CC (b) Bimax (c) Quest (d) Xmotif (e) Spectral

İnsan verisi için temel ikili kümeler CC algoritması ile *14×10* boyutunda, Bimax için *1608×3* boyutunda, Plaid için *21×31* ve Xmotif için *1318×6* boyutunda elde edilmiştir.

Fare verisi için temel ikili kümeler CC algoritması ile *83×74*, Bimax için *1080×3*, Quest için *996×72*, Xmotif için *1050×44* ve Spectral için *22×36* boyutunda elde edilmiştir.

Isı grafiklerine göre maya ve insan verisi için Bimax ve Xmotif, fare verisi için Quest ve Xmotif algoritmaları en büyük hacme sahip ikili kümeleri elde etmişlerdir. Ancak bu algoritmaların değerlendirme ölçülerine göre performanslarının da incelenmesi gerekir. Bimax ve Xmotif algoritmaları ile elde edilen ikili kümeler tek değerli oldukları için değerlendirme ölçülerinden sadece CKSB elde edilmektedir. Ancak ikili küme elemanlarının referans gösterdiği gerçek veri değeri göz önüne alındığında diğer değerlendirme ölçüleri de bu iki algoritma için de hesaplanır. Maya, insan ve fare verileri için algoritmaların değerlendirme ölçülerine göre karşılaştırılması Tablo 2'de verilmiştir.

Sonuçlar incelendiğinde maya verisi için CC, Bimax, Plaid ve Xmotif algoritmalarından elde edilen ikili kümeler için bulunan *VAR, CKSB* ve *ASR* değerlendirme ölçülerine göre en iyi performansı Bimax göstermiştir. *MSR, RI, SCS, SBM* ve *VE* değerlendirme ölçülerine göre en iyi performansı CC göstermiştir. *SMSR* ölçüsüne göre Plaid ve *ACV* ölçüsüne göre Xmotif en iyi performansı göstermiştir. Maya verisinde CC algoritması 5 ve Bimax algoritması 3 değerlendirme ölçüsüne göre en iyi performansı göstermiştir. İkili kümeleme algoritmaları için genel olarak karşılaştırma yapıldığında her bir algoritma için hesap-

TABLO 2: Gerçek veri setleri kullanılarak farklı algoritmalardan elde edilen ikili kümelerin değerlendirme ölçüleri.

Veri		VAR	MSR	RI	SMSR	CKSB	ACV	SCS	ASR	SBM	VE
Maya	CC	7101481	**141.87**	**0.14**	0.01	0.10	0.92	**0.98**	0.96	**0.96**	**51.47**
	Bimax	**1822684**	439.87	0.01	0.01	**1.43**	0.87	0.57	**0.97**	0.86	80.96
	Plaid	2851610	177.29	0.09	**0.00**	0.17	0.95	0.96	0.88	0.95	62.52
	Xmotif	3914500	146.40	0.01	0.01	0.23	**0.98**	0.80	0.77	0.79	93.42
İnsan	CC	**23206**	**96.64**	0.91	0.27	-0.98	0.02	**0.99**	0.17	0.90	**8.67**
	Bimax	4167838	2364.58	0.90	**0.25**	**1.48**	**0.36**	0.94	**0.86**	**0.96**	41.71
	Plaid	2439716	2113.20	**0.97**	0.38	-0.45	0.14	0.98	0.26	0.95	45.78
	Xmotif	7388661	984.47	0.90	0.45	-0.03	0.28	0.95	0.25	0.95	25.68
Fare	CC	1827000	0.02	0.89	0.06	-2.67	0.26	0.98	0.79	0.98	0.06
	Bimax	313467	0.28	0.01	0.03	1.43	0.34	0.92	0.73	0.90	0.02
	Quest	1785900	0.02	0.07	0.09	-2.74	0.07	0.99	0.17	0.99	0.09
	Spectral	115000	0.01	0.99	0.01	-2.34	0.10	0.95	0.98	0.95	0.03
	Xmotif	786629	0.02	0.04	0.10	0.19	0.27	0.98	0.77	0.97	0.03

TABLO 3: Veri setleri için farklı değerlendirme ölçülerine göre en iyi algoritmalar.

Veri	Isı grafiği	VAR	MSR	RI	SMSR	CKSB	ACV	SCS	ASR	SBM	VE
Maya	B ve X	B	CC	CC	P	B	X	CC	B	CC	CC
İnsan	B ve X	CC	CC	P	B	B	B	CC	B	B	CC
Fare	B ve X	S	S	S	S	B	B	Q	S	Q	B

B: Bimax; P: Plaid; Q: Quest; S: Spectral; X: Xmotif

lanmış değerlendirme ölçülerine göre CC ve Bimax algoritmaları diğerlerine göre daha iyi performans göstermiştir.

İnsan verisi için *SMSR, CKSB, ACV, ASR* ve *SBM* skorlarına göre Bimax daha iyi sonuç vermiştir. CC algoritmasıyla elde edilen ikili kümenin *VAR, MSR, SCS* ve *VE* skorları daha iyi sonuçlar vermişken *RI* skoruna göre Plaid algoritması daha iyi sonuç vermiştir. İnsan verisinde Bimax algoritması 5 ve CC algoritması 4 değerlendirme ölçüsüne göre en iyi performansı göstermiştir. İkili kümeleme algoritmalarının tümü için karşılaştırma yapıldığında 10 değerlendirme ölçüsünün 9'unda en iyi performansı gösteren Bimax ve CC algoritmalarıdır.

Fare verisinde *VAR, MSR, RI, SMSR* ve *ASR* skorlarına göre Spectral algoritması daha iyi sonuç vermiştir. Bimax algoritmasıyla elde edilen ikili kümenin *CKSB, ACV* ve *VE* skorları daha iyi sonuçlar vermişken *SCS* ve *SBM* skorlarına göre Quest algoritması daha iyi sonuç vermiştir. Fare verisi için Spectral algoritması 5, Bimax algoritması 3 ve Quest algoritması 2 değerlendirme ölçüsüne göre en iyi performansı göstermiştir. CC ve Xmotif algoritmaları ise değerlendirme ölçülerinin hiçbirinde en iyi performansı gösterememiştir. Beş farklı ikili kümeleme algoritması için hesaplanmış değerlendirme ölçülerine göre en iyi performansı Spectral algoritmasının gösterdiğini söyleyebiliriz.

Genel olarak maya ve insan verisi için CC ve Bimax, fare verisi için Spectral algoritmalarının diğerlerine göre daha iyi olduğu söylenebilir. Hem ısı grafikleri hem de değerlendirme ölçülerinden elde edilen sonuçlar Tablo 3'de özetlenmiştir.

Özet tablo incelendiğinde tüm kriterler dikkate alarak en fazla iyi sonuç veren algoritmanın Bimax olduğu söylenebilir.

SONUÇ

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 kalmaktadır ve bu nedenle ikili kümeleme algoritmaları geliştirilmiştir. Farklı veri yapıları için farklı algoritmalar önerildiği için literatürde birçok ikili kümeleme algoritması mevcuttur. Bu çalışma kapsamında bu algoritmalardan ikili küme türlerine göre uygun olarak kullanılan CC, Bimax, Plaid, Quest, Spectral ve Xmotif ikili kümeleme algoritmaları dikkate 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 sınırlı sayıdadır. Literatürdeki çalışmalar genellikle bir ya da birkaç değerlendirme ölçüsü bakımından bazı algoritmaların karşılaştırılmasına ilişkindir. İkili kümelerin değerlendirilmesi için birçok değerlendirme ölçüsü

mevcuttur. Bu değerlendirme ölçüleri varyans ölçüsü (*VAR*), ortalama karesel artık skoru (*MSR*), uygunluk indeksi (*RI*), 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 çalışmada 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.

Uygulamalar üç gerçek veri seti için yapılmıştır. Bunlar maya verisi, lenf hücrelerinin gen ifadesini içeren insan verisi ve protein-protein etkileşim skorlarını içeren fare verisidir. Gerçek veri setlerinden maya ve insan verisi için ikili kümeleme algoritmalarından CC ve Bimax, fare verisi için Spectral algoritması birçok değerlendirme ölçüsüne göre diğer algoritmalardan elde edilen ikili kümelere göre daha iyi sonuçlar vermiştir. Veri setlerinin yapısı incelendiğinde maya ve insan verisinin değer aralıkları fare verisinin değer aralıklarına göre daha fazladır. Böylece CC ve Bimax algoritmalarının uygulanmasında geniş ölçekli veri setleri, Spectral algoritmasının ise birbirine çok yakın değer alan veri setlerinde uygulanması sonucuna ulaşılmıştır.

Sonuç olarak, bu çalışmada kullanılan gerçek veri setlerinde (maya, insan ve fare) CC, Bimax ve Spectral algoritmaları diğer algoritmalara göre daha iyi sonuçlar vermiştir. Ayrıca kullanılan bu algoritmaların performanslarının simülasyon çalışmalarıyla değerlendirilmesi önerilmektedir. Bununla birlikte elde edilen ikili kümeleme sonuçlarının gen verileri açısından kuramsal (biyolojik) olarak uygunluğunun da uzman görüşleri doğrultusunda karşılaştırılması ve tartışılması gerekir.

Finansal Kaynak

Bu çalışma sırasında, yapılan araştırma konusu ile ilgili doğrudan bağlantısı bulunan herhangi bir ilaç firmasından, tıbbi alet, gereç ve malzeme sağlayan ve/veya üreten bir firma veya herhangi bir ticari firmadan, çalışmanın değerlendirme sürecinde, çalışma ile ilgili verilecek kararı olumsuz etkileyebilecek maddi ve/veya manevi herhangi bir destek alınmamıştır.

Çıkar Çatışması

Bu çalışma ile ilgili olarak yazarların ve/veya aile bireylerinin çıkar çatışması potansiyeli olabilecek bilimsel ve tıbbi komite üyeliği veya üyeleri ile ilişkisi, danışmanlık, bilirkişilik, herhangi bir firmada çalışma durumu, hissedarlık ve benzer durumları yoktur.

Yazar Katkıları

Fikir/Kavram: *Ahmet Kocatürk, Bülent Altunkaynak;* ***Tasarım:*** *Ahmet Kocatürk, Bülent Altunkaynak;* ***Denetleme/Danışmanlık:*** *Ahmet Kocatürk, Bülent Altunkaynak;* ***Veri Toplama Ve/Veya İşleme:*** *Ahmet Kocatürk, Bülent Altunkaynak;* ***Analiz Ve/Veya Yorum:*** *Ahmet Kocatürk, Bülent Altunkaynak;* ***Kaynak Taraması:*** *Ahmet Kocatürk, Bülent Altunkaynak;* ***Makalenin Yazımı:*** *Ahmet Kocatürk, Bülent Altunkaynak;* ***Eleştirel İnceleme:*** *Bülent Altunkaynak*

KAYNAKLAR

1. Cheng Y, Church GM. Biclustering of expression data. Proc Int Conf Intell Syst Mol Biol 2000;8:93-103.
2. Prelić A, Bleuler S, Zimmermann P, Wille A, Bühlmann P, Gruissem W, et al. A systematic comparison and evaluation of biclustering methods for gene expression data. Bioinformatics 2006;22(9):1122-9.
3. Lazzeroni L, Owen A. Plaid models for gene expression data. Statistica Sinica 2002;12(1):61-86.
4. Murali TM, Kasif S. Extracting conserved gene expression motifs from gene expression data. Pac Symp Biocomput 2003;8:77-88.
5. Kluger Y, Basri R, Chang JT, Gerstein M. Spectral biclustering of microarray data: coclustering genes and conditions. Genome Res 2003;13(4):703-16.
6. Das S, Idicula SM. Comparative advantages of novel algorithms using MSR threshold and MSR difference threshold for biclustering gene expression data. In: Arabnia RH, Tran Q, eds. Software Tools and Algorithms for Biological Systems. 1st ed. New York: Springer; 2011. p.123-34.
7. Bhattacharya S, Go D, Krenitsky DL, Huyck HL, Solleti SK, Lunger VA, et al. Genome-wide transcriptional profiling reveals connective tissue mast cell accumulation in bronchopulmonary dysplasia. Am J Respir Crit Care Med 2012;186(4):349-58.

8. Li L, Guo Y, Wu W, Shi Y, Cheng J, Tao S. A comparison and evaluation of five biclustering algorithms by quantifying goodness of biclusters for gene expression data. BioData Min 2012;5(1):8.
9. Zhao H, Wee-Chung Liew A, Wang DZ, Yan H. Biclustering analysis for pattern discovery: current techniques, comparative studies and applications. Current Bioinformatics 2012;7(1):43-55.
10. Eren K, Deveci M, Küçüktunç O, Çatalyürek ÜV. A comparative analysis of biclustering algorithms for gene expression data. Brief Bioinform 2012;14(3):279-92.
11. Padilha VA, Campello RJ. A systematic comparative evaluation of biclustering techniques. BMC Bioinformatics 2017;18(1):55.
12. Mirkin B. Mathematical classification and clustering: from how to what and why. In: Balderjahn I, Mathar R, Schader M, eds. Classification, Data Analysis, and Data Highways. 1st ed. Berlin, Heidelberg: Springer; 1996. p.172-81.
13. Aguilar-Ruiz JS. Shifting and scaling patterns from gene expression data. Bioinformatics 2005;21(20):3840-5.
14. Rodriguez-Baena DS, Perez-Pulido AJ, Aguilar-Ruiz JS. A biclustering algorithm for extracting bit-patterns from binary datasets. Bioinformatics 2011;27(19):2738-45.
15. Turner H, Trevor B, Wojtek K. Improved biclustering of microarray data demonstrated through systematic performance tests. Comput Stat Data Anal 2005;48(2):235-54.
16. Hartigan JA. Direct clustering of a data matrix. J Am Stat Assoc 1972;67(337):123-9.
17. Pontes B, Girldez R, Aguilar-Ruiz JS. Quality measures for gene expression biclusters. PloS One 2015;10(3):e0115497.
18. Yip KY, Cheung DW, Ng MK. Harp: a practical projected clustering algorithm. IEEE Trans Knowl Data Eng 2004;16(11):1387-97.
19. Mukhopadhyay A, Maulik U, Bandyopadhyay S. A novel coherence measure for discovering scaling biclusters from gene expression data. J Bioinform Comput Biol 2009;7(5):853-68.
20. Chia BKH, Karuturi RKM. Differential co-expression framework to quantify goodness of biclusters and compare biclustering algorithms. Algorithms Mol Biol 2010;5(1):23.
21. Yang WH, Dai DQ, Yan H. Finding correlated biclusters from gene expression data. IEEE Trans Knowl Data Eng 2011;23(4):568-84.
22. Teng L, Chan L. Discovering biclusters by iteratively sorting with weighted correlation coefficient in gene expression data. J Signal Process Syst 2008;50(3):267-80.
23. Ayadi W, Elloumi M, Hao JK. A biclustering algorithm based on a bicluster enumeration tree: application to DNA microarray data. BioData Min 2009;2(1):9.
24. Flores JL, Inza I, Larrañaga P, Calvo B. A new measure for gene expression biclustering based on non-parametric correlation. Comput Methods Programs Biomed 2013;112(3):367-97.
25. Zhu X, Qiu J, Xie M, Wang J. A multi-objective biclustering algorithm based on fuzzy mathematics. Neurocomputing 2017;253:177-82.