T.C.
İSTANBUL ÜNİVERSİTESİ-CERRAHPAŞA
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

YÜKSEK LİSANS TEZİ

AKIŞ SİTOMETRİSİ VERİLERİNDE ÖRÜNTÜ TANIMA

Eyyüp YILDIZ

DANIŞMAN
Dr. Öğr. Üyesi Tolga ENSARİ

II. DANIŞMAN
Dr. Öğr. Üyesi Leyla TÜRKER ŞENER

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

İSTANBUL- 2018

Bu çalışma 20.12.2018 Tarihinde aşağıdaki jüri tarafından Bilgisayar Mühendisliği Anabilim Dalı, Bilgisayar Mühendisliği Tezli Yüksek Lisans Programı Yüksek Lisans Tezi olarak kabul edilmiştir.

TEZ JÜRİSİ

Dr.Öğr.Üyesi Tolga ENSARİ
İstanbul Üniversitesi-Cerrahpaşa
Mühendislik Fakültesi

Prof. Dr. Ahmet SERTBAŞ
İstanbul Üniversitesi-Cerrahpaşa
Mühendislik Fakültesi

Dr.Öğr.Üyesi Funda AKAR
Erzincan Binali Yıldırım Üniversitesi
Mühendislik Fakültesi

20.04.2016 tarihli Resmi Gazete'de yayımlanan Lisansüstü Eğitim ve Öğretim Yönetmeliğinin 9/2 ve 22/2 maddeleri gereğince; Bu Lisansüstü teze, İstanbul Üniversitesi-Cerrahpaşa'nın abonesi olduğu intihal yazılım programı kullanılarak Fen Bilimleri Enstitüsü'nün belirlemiş olduğu ölçütlere uygun rapor alınmıştır.

ÖNSÖZ

Yüksek lisans eğitimim ve tez çalışmam boyunca yardımları, yönlendirmeleri ve katkılarıyla beni destekleyen değerli hocalarım ve tez danışmanlarım sayın Dr. Öğr. Üyesi Tolga ENSARİ ve sayın Dr. Öğr. Üyesi Leyla TÜRKER ŞENER'e sonsuz teşekkürlerimi sunuyorum.

Bu süreçte karşıma çıkan her sorunda bana yardımcı olan, çözüm yolları öneren Erzincan Binali Yıldırım Üniversitesi Bilgisayar Mühendisliği bölümünde görev yapan değerli hocalarım sayın Prof. Dr. Ahmet BARAN, sayın Dr. Öğr. Üyesi Funda AKAR, sayın Dr. Öğr. Üyesi Fulya ASLAY ve sayın Dr. Öğr. Üyesi Kamil ORMAN'a ve Mühendislik Fakültesinde beraber çalıştığım kıymetli çalışma arkadaşlarıma çok teşekkür ederim.

Tüm hayatım boyunca yanımda olan çok değerli kardeşlerim ve sevgili aileme desteklerinden dolayı sonsuz minnettarım.

Aralık 2018 Eyyüp YILDIZ

İÇİNDEKİLER

Sayfa No

ŞEKİL LİSTESİ

Sayfa No

TABLO LİSTESİ

Sayfa No

SİMGE VE KISALTMA LİSTESİ

Simgeler **Açıklama**

°	: Derece
$\mathbb{E}$	: Beklenen Değer
$\boldsymbol{\mu}$	: Ortalama
$\boldsymbol{\sigma}$	: Varyans
$\sum$	: Kovaryans
$\boldsymbol{T}$	: Transpoz
$\sim$	: Yaklaşık Değer

Kısaltmalar **Açıklama**

AS	: Akış Sitometrisi
Nm	: Nanometre
GKM	: Gauss Karışım Modeli
BE	: Beklenti Eniyileme
BBK	: Bayes Bilgi Kriteri
DLBCL	: Diffüz Büyük B Hücreli Lenfoma (Diffuse Large B-Cell Lymphoma)
BSO	: Birleşik Sınıflandırma Olabilirlik

ÖZET

YÜKSEK LİSANS TEZİ

AKIŞ SİTOMETRİSİ VERİLERİNDE ÖRÜNTÜ TANIMA

Eyyüp YILDIZ

İstanbul Üniversitesi-Cerrahpaşa

Lisansüstü Eğitim Enstitüsü

Bilgisayar Mühendisliği Anabilim Dalı

Danışman : Dr. Öğr. Üyesi Tolga ENSARİ

II. Danışman : Dr. Öğr. Üyesi Leyla TÜRKER ŞENER

Kan Örnekleri incelenerek yapılan analizlerde Akış Sitometrisi (Flow Cytometry) cihazı sıklıkla kullanılmaktadır. Akış sitometrisi verilerinin analizine hastalık tanısı koyma, hastalığın ilerleme safhasını izleme gibi durumlarda ihtiyaç duyulmaktadır. Fakat bu çok boyutlu verilerin insan eliyle yani manuel olarak analizinin yapılması çeşitli sebeplerden dolayı istenilen seviyede verimle yapılamamaktadır. Bu Tez çalışmasında Akış Sitometrisi verilerinde uzmanlar tarafından manuel olarak yapılan kapılama (gating) işleminin bilgisayar tarafından otomatik olarak yapılabilmesine olanak sağlayan bir algoritma oluşturulması amaçlanmıştır. Oluşturulan algoritma k-ortalamalar (k-means), Gauss Karışım Yöntemi (Gaussian Mixture Method – GKM) kümeleme yöntemleri, Beklenti Eniyileme (Expectation Maximization – BE) algoritması ve Chernoff uzaklık ölçüm yöntemini içermektedir. Tez kapsamında oluşturulan algoritma DLBCL veri kümesi üzerinde iki farklı şekilde test edilmiş olup %87,44 ve %86,06 başarı oranı elde edilmiştir. Sonuçlar var olan yöntemlerle karşılaştırıldığında birçok benzer çalışmaya göre daha yüksek başarı seviyesi elde edilmiştir.

Aralık 2018, 61 sayfa.

Anahtar kelimeler: Akış Sitometrisi, k-ortalamalar, Kümeleme, Gauss Karışım Modeli, Bayes Bilgi Kriteri, Kapılama

SUMMARY

M.Sc. THESIS

PATTERN RECOGNITION in FLOW CYTOMETRY DATA

Eyyüp YILDIZ

İstanbul University-Cerrahpaşa

Institute of Postgraduate Education

Department of Computer Engineering

Supervisor : Assist. Prof. Dr. Tolga ENSARİ

Co-Supervisor : Assist. Prof. Dr. Leyla TÜRKER ŞENER

Flow Cytometry device is frequently used in the analysis of blood samples. Analysis of Flow Cytometry data is needed in cases such as diagnosing disease, monitoring progression of disease. However, the manual analysis of these multi-dimensional data by the hand cannot be performed at the desired level due to various reasons. In this thesis study, it is aimed to create an algorithm which allows the gating process by manual gating to be performed automatically by the experts in the flow cytometry data. The algorithm consists of k-means, Gaussian Mixture Method (GMM) clustering methods, Expectation Maximization (EM) algorithm and Chernoff distance measurement method. The algorithm developed in the scope of the thesis was tested in two different ways on the DLBCL dataset and a success rate of 87.44% and 86.06% was obtained. The results showed a higher level of success was obtained compared to many similar studies.

December 2018, 61 pages.

Keywords: Flow Cytometry, k-means, Clustering, Gaussian Mixture Model, Bayesian Information Criteria, Gating

1. GİRİŞ

Akış Sitometrisi (Flow Cytometry - AS) kanda bulunan hücreler üzerinde ölçümler yaparak değerler üreten bir cihazdır. Bu alanda yapılan çalışmalar cihazın hücrelerden ürettiği değerler üzerinden ilerlemektedir. Sağlık alanında birçok yerde kullanılan bu cihazın ürettiği verilerin analizi uzmanlar tarafından yapılmaktadır. AS ölçüm yaptığı her hücre için 2'den fazla ölçüm sonucu üretir. Bu çok boyutlu verileri uzmanlar analiz ederken görsel sınırlılıktan dolayı sıklıkla bir veya iki, nadir olarak üç boyutlu grafikler kullanır. Bu durum verilerin aynı anda tüm boyutlarıyla analiz edilmesine imkân vermez. Bunun sonucu olarak verileri doğru şekilde analiz etmek için daha fazla zaman, insan gücü, analiz vb. etmenlere ihtiyaç duyulur. Bu durum AS verilerinin analizinde bilgisayar destekli yöntemlerin geliştirilmesine ihtiyaç olduğu sonucunu göstermiştir.

AS verilerinin bilgisayar destekli analizi için yapılan çalışmalar büyük oranda otomatik kapılama (gating) üzerine yoğunlaşmıştır. Kapılama AS cihazının ürettiği verilerden yararlanarak aynı hücre grubuna ait olan hücreleri tespit etmektir. AS verilerinin analizinde bu adım hayati önem taşımaktadır. Çünkü yanlış yapılan kapılama sonucunda yanlış çıkarımlar elde edilir. Sonuç olarak üzerinde çalışılan veriler için doğru sonuçlar elde edilemez. Otomatik kapılama için çeşitli yöntemler geliştirilmiştir. Geliştirilen yöntemler çoğunlukla kümeleme tabanlı yöntemlerdir. Kapılama için literatürde var olan kümeleme algoritmalarından en çok karışım modeli tabanlı yöntemler öne çıkmıştır. Bu yöntemler ilk olarak verileri modellemek için olasılık dağılımı seçmekle çalışmaya başlar. Veriler seçilen bu olasılık dağılımının karışımı ile ifade edilir. Burada karışımdan kasıt veriyi birden fazla sayıda dağılımla ifade etmektir. Karışım modeli olarak ifade edilen verilerden geliştirilen çeşitli kriterler veya yöntemler aracılığıyla aynı kümeye ait olan hücre grupları tespit edilmeye çalışılır. Bu şekilde yapılan çalışmaların uzmanlar tarafından yapılan manuel kapılamalarla karşılaştırıldıklarında başarılı sonuçlar ortaya koydukları gözlemlenmiştir.

Bu tez çalışması AS verilerini otomatik kapılama üzerinedir. Bunun için kümeleme yöntemleri arasında yer alan k-ortalamalar (k-means) ve Gauss karışım modeli (Gaussian mixture model)'ni kullanarak çalışan bir yöntem kullanılmıştır. Yöntem DLBCL AS veri kümesi üzerinde test edilmiştir. İki farklı kriterde test yapılmıştır. İlk olarak verilerin küme sayıları

yönteme giriş parametresi olarak verilmiştir. İkinci testte ise yöntem tamamen otomatik olarak çalışmıştır. Elde edilen sonuçlar yöntemin başarılı bir şekilde çalıştığını ortaya koymuştur.

Tezin 2. bölümünde ilk olarak AS cihazının çalışma prensibi ve manuel kapılamanın nasıl yapıldığı anlatılmıştır. Daha sonra literatürde yapılan çalışmalar kullandıkları yöntem benzerliklerine göre gruplanarak özet halinde sunulmuştur. 3. bölümde ilk olarak kullanılan veri kümeleri açıklanmıştır. Ardından tez kapsamında kullanılan kümeleme yöntemini ayrıntılı biçimde alt başlıklar halinde anlatılmıştır. 4. bölümde yapılan testler ve sonuçları paylaşılmıştır. Son olarak 5. bölümde elde edilen sonuçlar değerlendirilmiş ve gelecek çalışmalar için bilgiler verilmiştir.

2. GENEL KISIMLAR

Bu bölümde ilk olarak Akış Sitometrisi (AS) cihazının çalışma prensibi ve üretilen verilere yapılan ön işlemler anlatılmıştır. Daha sonra Akış Sitometrisi verilerinin laboratuvar ortamında nasıl analiz edildiğine kısaca değinilmiştir. Son olarak bilgisayar bilimleri alanında Akış Sitometrisi kapılama üzerine yapılan çalışmalar incelenmiştir.

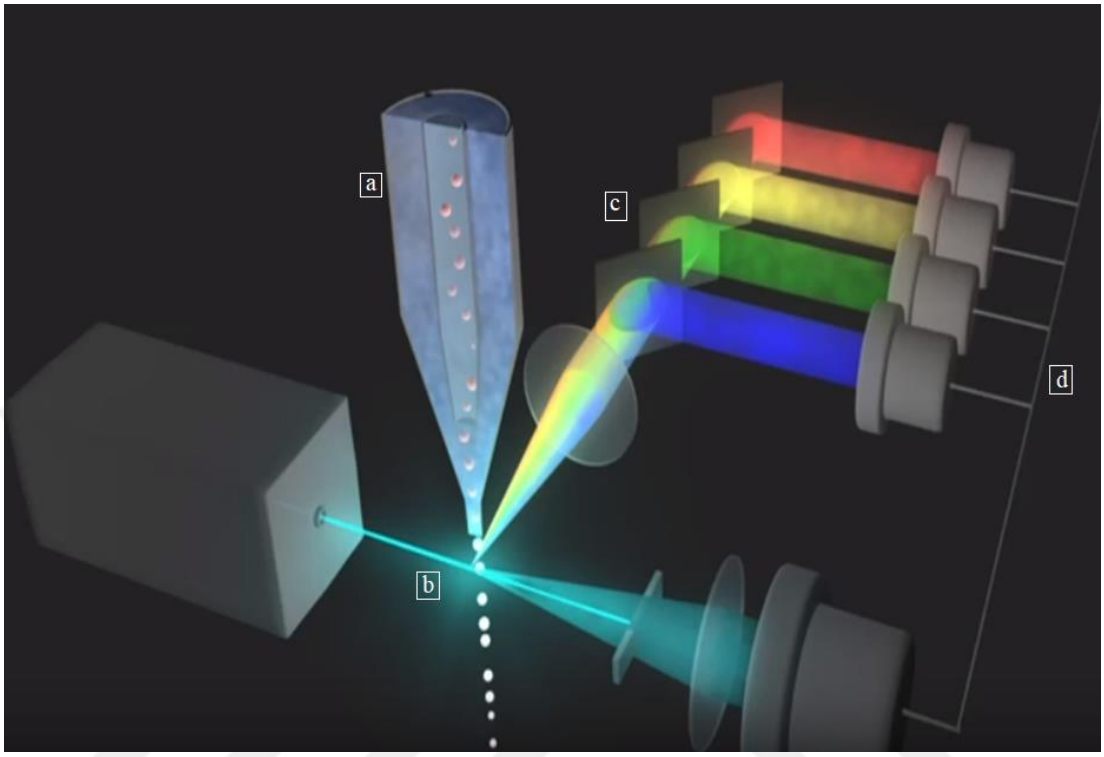
2.1. AKIŞ SİTOMETRİSİ

2.1.1. Akış Sitometrisi Cihazının Çalışma Prensibi

Sitometri hücrelerin fiziksel ve kimyasal özelliklerinin incelenmesi demektir [1]. AS ise özel sıvı içerisinde bir arada bulunan farklı gruptaki hücrelerin her biri hakkında belirli sayıda bilgi üreten bir cihazdır. Amaç cihazın ürettiği veriler sayesinde farklı gruptaki hücreler arasından aynı gruptaki hücreleri tespit etmektir. Daha çok mikroskobu andıran ilk örneği 1934 yılında Andrew Moldavan tarafından oluşturulan cihaz, sonraki yıllarda daha da geliştirilmesiyle birlikte önemi ve kullanım alanları artarak günümüze kadar gelmiştir [2]. Başlangıçta hematoloji laboratuvarlarında kullanılan cihaz günümüzde immünoloji, onkoloji başta olmak üzere organ nakli birimleri, patoloji, biyokimya gibi klinik laboratuvarlarda tanı koymada, hücre alt gruplarının ve hücre döngü safhasının belirlenmesinde vb. sıklıkla kullanılmaktadır [3]. AS cihazının ürettiği bilgilerle; hücrelerin yaklaşık büyüklüğünü, granülaritesini (hücrenin tanecik veya içerisinde bulunan organel yoğunluğu, çeşitliliği) ve alt hücre gruplarını ayırt edici özgün bilgileri elde etmemiz mümkündür.

AS akışkan sıvı sistemi, lazer kaynağı, ayna sistemi, dedektörler ve elektronik sistem olmak üzere beş temel kısımdan oluşmaktadır. İşlem öncelikle hücrelerin özel bir akışkan sıvı olan laminar sıvı içerisinde süspansiyon (hücrelerin sıvı içerisinde erimeden bulunma durumu) hale getirilmesi ile başlar (Şekil 2.1 a). Burada hücreler tek sıra halinde dizilirler. Daha sonra sıvı içerisinde bulunan hücreler tek tek kanal dışına çıkarak lazer ışınına maruz kalırlar (Şekil 2.1 b). Lazer ışınına maruz kalan hücre ışığı farklı açılarda yansıtır. Yansıyan ışınlar aynalar yardımı ile dedektörlere iletilir. Dedektörler bu ışınları elektrik sinyallerine dönüştürür (Şekil 2.1 c). Son olarak elde edilen bu sinyaller bilgisayar sistemi sayesinde sayısal verilere dönüştürülür (Şekil 2.1 d)[1-3]. Bu işlemler tüm hücreler için yapılır. İşlemdeki en önemli nokta lazer ışınına maruz kalan hücrelerin ışığı farklı açılarla yansıtmasıdır. Her hücre türü kendine

özgü özelliklerini gösterecek şekilde ışığı yansıtır ve bu yansımalar sonucu elde edilen verilerden hücre grupları ayırt edilmeye çalışılır.



Şekil 2.1: Akış Sitometrisi Cihazının Yapısı. a) Akışkan sıvı sistemi. b) Lazer kaynağı. c) Ayna sistemi. d) Dedektörler ve elektronik sistem [4]

İşlem sonucunda elde edilen özellikler; İleri Saçılım (Forward Scatter), Yan Saçılım (Side Scatter) ve floresan (fluorescence) özellikler şeklinde adlandırılan üç temel grupta toplanabilir.

2.1.1.1.İleri Saçılım

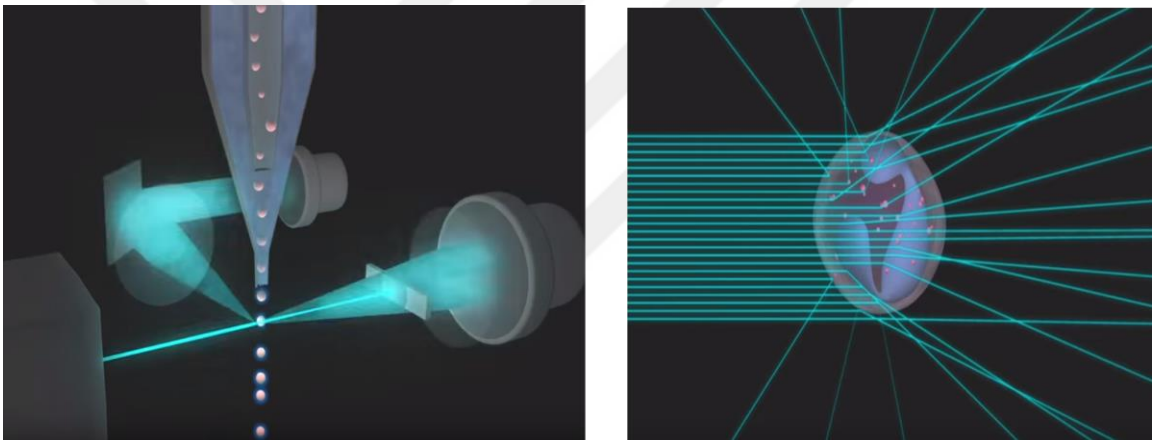
Şekil 2.2 a'da görüldüğü üzere hücre, üzerine gelen ışını ileri ve yan olmak üzere 2 yönde yansıtır. İleri Saçılım (forward scatter) lazer ışının geliş doğrultusu ile aynı yönde ileri doğru yayılan saçılıma denir. En fazla 20° açı ile yayılan ışınları kapsar [3]. Bu durum Şekil 2.2 a'da hücre üzerine düşen ışının düz olarak yansıması olarak gözükmektedir. İleri saçılım hücrenin yaklaşık büyüklüğünü ifade eder. Canlı ve ölü hücreler farklı büyüklüğe sahip olduğu için bunları ayırt etmede bu özellikten yararlanılır [5].

2.1.1.2. Yan Saçılım

Yan Saçılım (side scatter) lazer ışının yaklaşık olarak 90° açı ile yansıması ile oluşan saçılımdır (Şekil 2.2 a). Yan saçılım hücrenin granülaritesi yani içyapısı hakkında bilgi verir [3, 6]. İleri ve yan saçılım özelliklerinin her ikisi de hücrelere özgü ayırt edici özelliklerdir[5].

2.1.1.3.Floresan Özellikler

AS hücreler hakkında büyüklük ve granül yapı haricinde farklı türde verilerde üretir. Oluşturulan deney düzeneğine bağlı olarak bu verilerin sayısı 1'den 20'ye kadar çıkabilir [7]. Bunlar floresan özelliklerden oluşturulan verilerdir. Bu veriler hücre yüzeyine tutunan floresan işaretli antikorların ışını farklı dalga boylarında yansıtması sonucu elde edilir (Şekil 2.2 b).

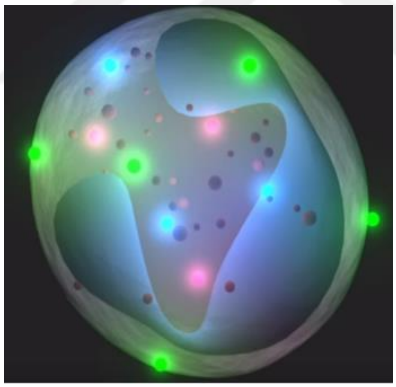


Şekil 2.2: a) İleri ve Yan Saçılım, b) Hücreye Gönderilen Lazer Işının Farklı Açılarla Yansıması [4]

Verileri elde etme işlemi istenilen hücreleri ayırt edici bilgileri verebilecek floresan işaretli antikorların seçimi ile başlar. Floresan, bir madde üzerine düşen ışınların farklı dalga boyunda ışınlara diğer bir deyişle farklı renge dönüştürerek geri yansıması olayı olarak tanımlanır. Antikorlar (marker- cluster of differentiation-CD) ise özelliklerine göre uygun hücre yüzeyine tutunan, hücreleri ayırt etmede başvurulan maddeler olarak düşünülebilir [8]. Örneğin A antikoru X hücresi yüzeyine tutunurken, Y hücresi yüzeyine tutunmaz. Bu sayede A antikorunu bulunduran hücreye X hücresi denilebilir. Antikorlar bu floresan maddeler ile birleştirilir/işaretlenir. Farklı tür antikorlar farklı dalga boyunda ışın yansıtan floresan maddeler ile birleştirilir. Bu sayede farklı tür floresan işaretli antikorlar farklı rengi yansıtırken, aynı tür floresan işaretli antikorlar aynı rengi yansıtır. Örnek olarak FITC-CD3 bir floresan işaretli

yüzey antikorudur. Burada FITC floresan maddesi ışığı 490 nm'de (mavi renk) absorbe eder ve en iyi 525 nm'de (yeşil renk) olmak üzere 575-700 nm aralığında yansıtır. CD3 ise kanda bulunan T hücreleri için yüzey antikorudur [5]. Daha sonra Şekil 2.3'te görüldüğü üzere floresan işaretleri antikorlar hücre yüzeylerine tutunurlar.

Hücre lazer ışınına maruz kaldığında floresan madde lazer ışının önce absorbe eder, floresan maddeye özgü belirli bir eşik değerinden sonra ise bu absorbe ettiği ışını farklı bir dalga boyu aralığında (renkte) yansıtır (Şekil 2.1 d). Yansıyan bu ışın aynalar yardımı ile o rengi toplayan dedektöre iletilir. Dedektöre ne kadar güçlü ışın gelirse o kadar güçlü sinyal üretir. Fazla sinyal oluşması hücrede o antikorun çok olduğu anlamına gelirken az sinyal oluşması o antikorun az bulunduğu anlamına gelir [2, 5]. Floresan maddeler kullanılan lazer ışığına bağlı olarak mümkün olduğunca ayırt edici dalga boylarında seçilmelidirler[9]. Çünkü yakın dalga boyunda ışın yansıtan floresan maddeler seçilirse, ışınları toplayan dedektörler kendi floresan maddeleri haricinde yakın renkte olan diğer floresan maddelerin ışınlarını da toplayabilir. Bu durum hücre özelliğinin durum yanlış şekilde ölçülmesi anlamına gelir.



Şekil 2.3: Hücre Yüzeyinde Bulunan Farklı Renkte Işık Yansıtan Floresan İşaretli Antikorlar [4]

AS verileri elde edildikten sonra doğru şekilde analiz yapabilmek için veriler çeşitli önişlemlerden geçirilir. Bunlar arasından en önemlisi düzeltme (compensation) ve dönüşümdür (transformation) [7].

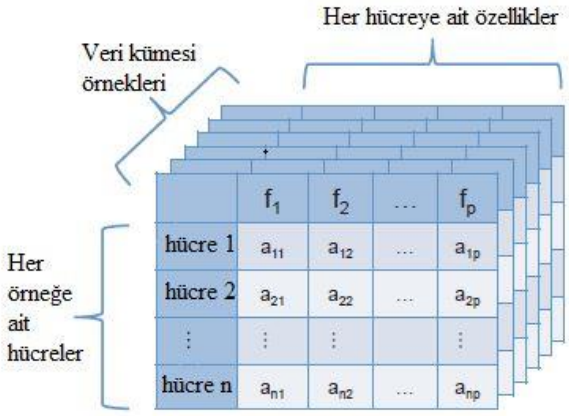
Yukarıda bahsedildiği gibi ileri ve yan saçılım haricinde hücre özelliklerini ölçmek için her ne kadar ayrı dalga boylarına sahip floresan maddeler seçilse de yansıyan ışınların bir kısmı yansıma aralıklarının çakışması yüzünden diğer dedektörler tarafından da toplanabilir. Bu

durumda elde edilecek verilerde belirli oranda yanlışlık oluşur. Bunu ortadan kaldırmak için her floresan dedektörünün elde ettiği sinyal üzerinde düzeltme işlemi yapılır [8]. Bu işlem sonucunda floresan ışınlarını toplayan her dedektör yalnızca kendi dalga boyundaki ışınları toplamış olur. Elde edilen bu bilgilerle farklı hücre türleri ayırt edilmeye çalışılır. Tezde kullanılan veriler üzerinde düzeltme işlemi yapılmış halde olduğu için bu işlem ayrıntılı olarak anlatılmayacaktır. Ayrıntılı bilgi için [10]'a bakılabilir.

AS verileri lineer olarak saklanırlar ve analizden önce üzerlerinde dönüşüm yapılır. Dönüşümün amacı hücre gruplarını daha iyi ayırt edecek uzaya aktarmaktır. Bu amaçla çeşitli dönüşüm yöntemleri önerilmiştir. Başlangıçta logaritmik dönüşüm sıklıkla kullanılmıştır. Fakat düzeltme yapılan AS verilerinin negatif değere sahip olduğu durumlarda logaritmik dönüşüm yetersiz kalmaktadır. Bu sebeple birçok farklı dönüşüm yöntemleri önerilmiştir. Hiperbolik ters sinüs (arcsine) dönüşümü, Box-Cox dönüşümü bunlardan bazılarıdır [7].

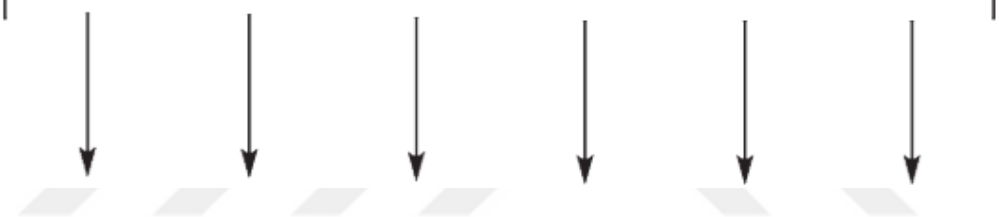
2.1.2. Akış Sitometrisi Verilerinin Manuel Analizi

AS verilerinin analizinde amaç benzer hücre gruplarını tespit ederek istenen özelliklerin ölçülmesidir. Şekil 2.4'te örnek AS veri formatı verilmiştir. Buna göre her AS veri kümesi örneklerden, her örnek ise hücrelerden oluşmaktadır. Her örnekte bulunan hücreler de ileri, yan saçılım ve farklı floresan özelliklerle ifade edilmektedir.



Şekil 2.4: Örnek AS veri kümesi [11]

Hücre	İleri Saçılım	Yan Saçılım	Yeşil fl.	Turuncu fl.	Kırmızı fl.
1	784	1233	10344	476	300
2	700	1145	11657	334	435
3	698	1289	13228	476	436
4	877	990	10453	335	478
5	789	1119	12897	501	512
6	690	998	14987	375	423
7	777	1309	14376	349	584
8	689	1401	13765	360	474
9	2089	3022	543	299	14099
10	786	1322	10367	474	499
11	688	1034	11438	356	375
12	1998	3400	464	487	15833
13	2134	3289	502	503	14998
14	745	1008	13245	499	416
15	300	432	321	321	431
16	876	1204	11498	509	485
17	775	1023	11749	464	458
18	2109	3356	387	375	15684
19	799	1039	12149	399	396



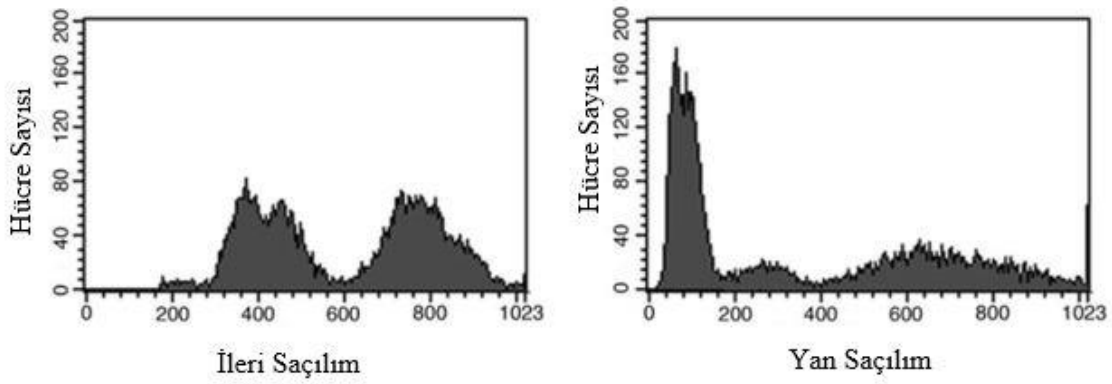
Şekil 2.5: Akış Sitometrisi örnek veri yapısı [2]

Şekil 2.5'te AS cihazının ürettiği örnek bir veri dosyası görülmektedir. Burada her hücre ileri saçılım (forward scatter), yan saçılım (side scatter) ve yeşil, turuncu, kırmızı renk yansıtan floresan sinyaller olmak üzere 5 farklı özellikle ifade edilmiştir. Her hücre 5 boyutlu bir veri gibi düşünülebilir. AS cihazında bir canlı için on binlerce hücre analiz edilmektedir. Şekil 2.5'teki örnekte 10000 hücre olduğu varsayılırsa toplam 10000x5 boyutunda bir veri matrisi ortaya çıkar [2]. Bu kadar büyük boyuttaki verileri de doğru şekilde analiz etmek oldukça zordur. Üç boyut ve üzeri verileri görsel olarak ifade etmek mümkün olmadığından uzmanlar tarafından manuel olarak yapılan AS veri analizinde genellikle histogram grafikleri ile iki boyutlu saçılım grafikleri (scatter plot) kullanılır [2].

2.1.2.1. Histogram Grafiği

Histogram grafikleri verilerin yoğunluğunu gösteren tek boyutlu grafiklerdir. Histogramlarda x ekseni hücrenin analiz edilmek istenen özelliğinin (ileri, yan saçılım veya floresan işaretli

antikor) sayısal değerlerini, y ekseni ise hücrelerin sayısını ifade eder. Şekil 2.6'daki histogram grafiklerine baktığımızda hücrelerin belirli değerlerde yoğunlaşarak tepe noktaları oluşturduğunu görebiliriz. AS verilerinin histogram grafiklerinde ölçülen özelliklere göre tepe sayısı değişebilir.



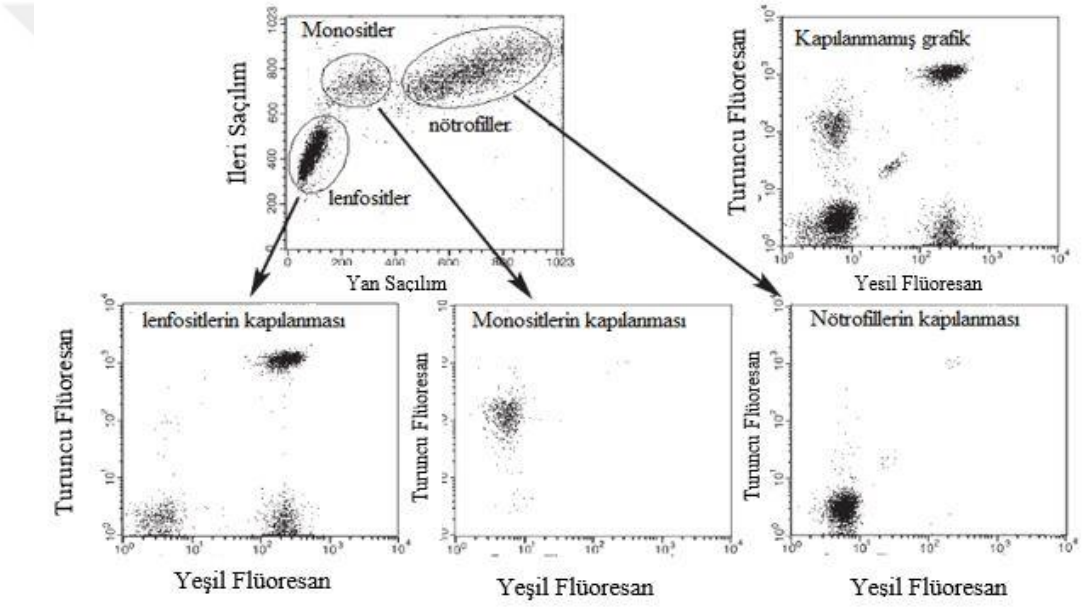
Şekil 2.6: İleri ve Yan Saçılım Değerlerinin Histogram Grafikleri [2]

İleri saçılım özelliği için oluşturulan histogram grafiğinde benzer büyüklükteki hücreler beraber yer alırlar. Bu grafik veride bulunan hücre grubu sayısı hakkında bilgi verir. Yan saçılım grafiği de ileri saçılıma benzer olarak yakın granülariteye sahip hücreler tepe oluştururlar. Eğer ölçülen değerler floresan özellikler ise genellikle 2 büyük tepe oluşur. Büyük sayısal değerli tepe hücrelerin incelenen özelliği gösterdiğine işarettir. Benzer şekilde küçük değerli tepe ise hücrelerin incelenen özelliği göstermediğini ifade eder. Bu durum pozitiflik-negatiflik olarak adlandırılır [5][12]. Buradan yola çıkarak her histogram grafiğinde hücreler çoğunlukla iki kümeye ayrılır.

2.1.2.2. Saçılım Grafiği

Saçılım grafikleri hücrenin istenilen iki özelliğinin iki boyutlu uzayda görünümüdür. Grafiklerde x ekseni hücrenin bir özelliğini (örneğin ileri saçılım), y ekseni ise diğer özelliğini ifade (örneğin yan saçılım) ifade eder. Şekil 2.7 a'da da görüleceği üzere benzer gruptaki hücrelerin değerleri belirli bölgelerde yoğunlaşmışlardır. Uzmanlar kendi bilgi deneyimlerine göre bu grafikleri kullanarak hücre gruplarının sınırlarını manuel olarak belirlerler. Buna *Kapılama* (Gating) denir [5].

Kapılama işlemi ikili özelliklerin saçılım grafiklerinin sırayla analiz edilmesiyle yapılır. Burada genellikle ilk olarak ön-yan saçılım grafiği kullanılır (Şekil 2.7 a). Fakat burada kesinlik yoktur. Genellikle hücreleri en iyi şekilde ayırt ettiği düşünülen grafik seçilmeye çalışılır. Hücre analizini en iyi şekilde yapabilmek için verilerin tüm ikili grafik kombinasyonlarına bakılabilir. Ön-yan saçılım grafiğinde belirlenen hücre grupları için floresan maddelerden elde edilen grafikler kullanılarak kapılamanın doğruluğu kontrol edilebilir, alt hücre grupları belirlenebilir, hastalık teşhisi konulabilir [13]. Şekil 2.7 b grafiği yeşil ve turuncu özellik gösteren floresan maddelerin saçılım grafiğidir. Eğer ilk tüm veri kümesi bu grafik üzerinde incelenecek olursa yanlış şekilde gruplama yapılabilir.



Şekil 2.7: a) x ekseni ön saçılım, y ekseni yan saçılım olan, saçılım grafiği, b) x ekseni yeşil y ekseni turuncu olan grafik, c) a'da elde edilen hücre gruplarının alt grafikleri [2]

Şekil 2.7 c'de görüleceği üzere ön-yan saçılım grafiği üzerinden belirlenen hücre grupları için alt grafikler oluşturulmuştur. Bu alt grafikler incelenerek hücre gurubu alt gruplarına bölünür. Örneğin en soldaki grafik 3 alt gruba sahiptir. Burada her iki floresan işaretli maddeyi gösteren hücreler, her iki floresan maddeyi de yansıttığı için yüksek değerde sinyal oluştururlar. Bu sebeple bu hücreler grafikte sağ üstte yer alırlar. Benzer şekilde yalnızca bir floresan maddeyi gösteren hücreler sağ altta veya sol üstte, her iki floresan maddeyi göstermeyen hücreler ise sol altta yer alır [5, 12]. Buradan yola çıkılarak floresan özelliklere göre oluşturulan saçılım grafikleri en fazla 4 gruba ayrılabilir sonucuna varılabilir. Örnek olarak Şekil 2.7 c'de en

sağdaki grafik incelendiğinde nötrofil hücre grubunun yeşil ve turuncu floresan özelliklerinin ikisini de göstermediği görülür. Bu grafikler incelenerek hücre gruplarının sayısı, her gruptaki hücrelerin sayısı, bu hücrelerin tüm hücrelere oranı vb. gibi birçok bilgi elde edilir.

Temel olarak AS veri analizi 4 ana başlıkta incelenebilir; önişlemler (düzeltme, dönüşüm vb.), benzer hücre kümelerinin bulunması (kapılama), aynı hücre kümelerinin diğer örneklerde bulunan hücre kümeleri ile eşlenmesi ve elde edilen analiz sonuçlarından çıkarımlar yapmak (tanı koyma, aşı geliştirme, hücre döngü seviyesi tespiti vb.) [7]. Tez çalışması kapsamında bu işlemlerden ikincisi olan kapılama işleminin bilgisayar tarafından yapılması araştırılacaktır.

2.2. AKIŞ SİTOMETRİSİ VERİLERİNİN OTOMATİK KAPILANMASI

AS analizinde en önemli kısım benzer hücre kümelerinin belirlenmesidir. Hücre kümelerinin belirlenmesi sonrası elde edilen bilgilere göre analizler yapılır. Bu sebeple hücre kümelerinin doğru şekilde belirlenmesi hayati öneme sahiptir. FCS Express [14], Flowing Software [15] gibi yazılımlar AS verilerinin manuel kapılanması için kullanılmaktadır.

AS verilerinden benzer hücre kümelerinin manuel olarak tespit edilmesi yani kapılanması bölüm 2.1.2' de anlatılmıştır. Bölüm incelendiğinde kapılama işleminin manuel olarak yapılmasının birçok sebepten ötürü verimli olmadığı anlaşılmaktadır. Bunlardan birkaçını açıklamak gerekirse; öncelikle manuel kapılama yapmak uzman kişinin bilgisine bağlıdır. Bu durumda belirlenen hücre kümeleri subjektiftir, kişiden kişiye değişebilir. Bir veri kümesindeki verileri bir uzman 4 kümeye ayırırken başka bir uzman 5 kümeye ayırabilir veya kümelerin sınırları farklı olabilir. Bu durum üzerine yapılan bir çalışma 15 farklı kurumun aynı AS verisi üzerinde çalışmasını incelemiştir ve kurumların sonuçları %15-44 arası farklılık gösterdiğini ortaya koymuştur. Bu farklılığın temel sebebi ise kapılamalardaki değişimin olduğu görülmüştür [16]. İkinci olarak ise manuel kapılama fazla zaman alır. Çünkü her örnekteki hücrelerin grafikler üzerindeki dağılımları farklı olduğundan dolayı her örnek kendi içerisinde kapılanmaktadır. Son olarak AS verilerinin manuel kapılanması cihazın tam olarak verimli kullanılmadığı anlamına gelir. Şekil 2.4'te de görüleceği üzere incelenen örnekte her hücre 5 özellikle ifade edilmektedir. Fakat manuel kapılama da sıralı olarak histogramlar veya saçılım grafikleri, nadir olarak da 3 boyutlu grafikler kullanılmaktadır [17]. Yani mevcut durumdaki çok boyutlu verilerin aynı anda en fazla 3 boyutu kullanılabilmektedir. Bu durum hücrelerin aynı anda tüm özellikleri ile analiz edilemediği anlamına gelir [11, 18, 19]. Bahsedilen bu gibi

sebepler AS verilerini otomatik kapılayan yöntemlerin geliştirilmesine ihtiyaç olduğunu ortaya çıkarmıştır.

AS verilerinin otomatik kapılanması üzerine yapılan çalışmalar gözetimli ve gözetimsiz öğrenme olmak üzere iki kategoriye ayrılabilir. Gözetimli öğrenmede iki grup oluşturan değişkenler kullanılır. Bunlar; bağımsız değişken olarak AS cihazından elde edilen ölçüm sonuçları ile bağımlı değişken olarak hücre türlerini ifade eden etiketler (label) olarak tanımlanır. Burada ölçümler bağımsız değişken, etiketler ise bağımlı değişkenlerdir. Gözetimli öğrenmenin eğitim safhasında bu değişkenlerden yola çıkarak benzer hücre grupları belirlenmeye çalışılır. Bu öğrenme sürecinden sonra, algoritma yeni verileri doğru şekilde kapılamaya çalışır. Gözetimsiz öğrenmede ise gözetimli öğrenmede olduğu gibi bir öğrenme süreci yoktur. Yani bağımlı değişkenler olan etiketler yoktur. Algoritmalar doğrudan veriler üzerinden işlemler yaparak kapılama yapmaya çalışır [16]. Bu tezde uygulanan yöntem gözetimsiz öğrenme sınıfına girdiği için literatür taramasında gözetimli öğrenme ile yapılan çalışmalar göz ardı edilmiştir.

Literatürde AS verilerini gözetimsiz öğrenme yöntemlerini kullanarak kapılama yapan birçok çalışma mevcuttur. AS verilerinin gözetimsiz öğrenme yöntemleri kullanılarak kapılanması üzerine yapılan ilk çalışma 1985 yılında Robert F. Murphy tarafından yayınlanmıştır. Çalışmada kapılama yapmak için K-means/ISODATA yöntemi, hangi ölçüm parametrelerinin kapılamada kullanılacağı, başlangıçta oluşturulan kümelerin belirlenen kritere göre birleştirilerek küme sayısının azaltılması veya ilk kümeleme işlemi için başlangıç pozisyonlarının belirlenmesi gibi çeşitli parametre girişleri ile kontrol edilerek kullanılmıştır. Elde edilen sonuçlar AS verilerinden benzer hücre gruplarının elde edilmesinde kümeleme analizinin yararlı olduğunu göstermiştir [20]. N. Aghaeepour ve diğ. 2011 yılında K-ortalamalar algoritmasına yeni özellikler ekleyerek FlowMeans algoritmasını geliştirdikleri çalışmalarını yayınlamışlardır. Çekirdek Yoğunluk Tahmini (Kernel Density Estimator – ÇYT) ile belirledikleri k başlangıç sayısını mahalanobis uzaklık ölçüm yöntemine göre birleştirerek belirlenen küme sayısına ulaşılmaya çalışılan makalede yöntem hem manuel kapılamayla uyumlu sonuçlar vermesi hem de hızlı çalışmasıyla ön plana çıkmaktadır [21].

M. Boedigheimer ve J. Ferbas 2008 yılında kapılama problemini Gauss Karışım Modelini (Gaussian Mixture Model - GKM), Beklenti Eniyileme (Expectation Maximization - BE) kullanarak çözmeye çalıştıkları bir çalışma yayınlamışlardır. Burada her gauss bileşeni bir

hücre türünü ifade ederken, modeldeki değişkenler ise AS ölçüm sonuçlarına denk gelmektedir. Çalışma sonucunda önerilen yöntemin hücre gruplarının otomatik olarak belirlenmesinde gelişmiş performans sunduğu görülmüştür [17]. Aynı yıl yapılan bir diğer çalışma da ise C. Chan ve diğ. GKM'yi Markov zinciri Monte Karlo yöntemi ile birlikte kullanarak kapılama yaptıkları çalışmada başarılı sonuçlar elde etmişlerdir [22]. Yine 2008 yılında yapılan diğer bir çalışmada K.Lo ve diğ. AS verilerinde bulunabilen elipsoidal olmayan hücre kümelerini de tespit etmek amacıyla GKM'yi genelleştiren bir çalışma yayınlamışlardır. Çalışmalarında T-Karışım Modeli ve GKM yöntemlerini Box-Cox dönüşüm yöntemi ile kullanarak başarılı sonuçlar elde etmişlerdir [23]. G. Finank ve diğ. 2009 yılında, Lo'nun çalışmasını geliştirerek FlowMerge adını verdikleri kapılama yöntemini açıklayan çalışmalarını yayınlamışlardır. Çalışmada T-karışım modeliyle bulunan kümeler entropi tabanlı olarak Bayes Bilgi Kriteri (Bayesian Information Criteria - BBK) ile elde edilen küme sayısına erişene kadar birleştirilerek ilerler. Çalışma belirli bir şekli olmayan hücre gruplarını tespit etmede başarılı olarak önemli sonuçlar ortaya çıkarmıştır [24]. Y. Ge ve diğ. 2012 yılında sırası ile K-ortalamalar ve GKM yöntemlerini beraber kullandıkları bir çalışma yayınlamışlardır. FlowPeaks adı verilen yöntem K-ortalamalar yöntemi ile verileri fazla sayıda kümeye ayırır. Elde edilen GKM ile ifade edildikten sonra dağılımdaki tepe noktaları geliştirilen algoritmadaki kurallara birleştirilerek hücre kümeleri oluşturulur [25].

G. Walther ve diğ. 2009 yılında yayınladıkları makalede kapılama için karışım modelleri benzeri yaklaşımlardan farklı olarak ızgara tabanlı bir algoritma önermişlerdir. Yoğunluk tabanlı birleştirme adını verdikleri algoritmada eşit hacimli ızgaralar veriler üzerine düşürülür ve istatistiksel yöntemlere göre birleştirilmesi gereken komşu ızgaralar birleştirilerek kümeler oluşturulur [26]. Izgara tabanlı (grid based) diğer bir çalışma ise 2010 yılında Y. Qian ve diğ. tarafından yapılmıştır. FLOCK adını verdikleri yöntemde veriler yine ızgara üzerine düşürülür. Komşu ızgaraların hücre yoğunlukları Silhouette sınır değerinin üstünde ise birleştirilerek ilerlenir. FLOCK algoritmasının karışım modellerine göre daha hızlı çalışarak yüksek başarı oranı elde ettiği görülmüştür [27].

A. Cron ve diğ. 2013 yılında kapılamada AS verilerinde bulunabilen çok az sayıda hücre içeren kümeleri de tespit etme problemini de çözmek için bir yöntem önermişlerdir. Karışım modellerinde ilk olarak belirlenmesi gereken bileşen sayısı sorununu Dirichlet Process işlemi ile aşmaya çalışarak kapılama için Hiyerarşik Dirichlet Process Gaussian Mixture Model

(DPGMM) Yöntemini uygulamışlardır [28]. B.E. Köktürk ve B. Karaçalı 2014 yılında yayınladıkları çalışmada öz yinelemeli ikili bölme algoritması adını verdikleri modelden bağımsız sonlu olasılık hesabı yapan BE algoritması kullanılarak kümeleme yapmışlardır. İlk adım olarak önerilen BE algoritmasıyla veriler iki altkümeye ayrılır. Sonra her altküme iki kümeye ayrılarak daha fazla kümeye ayırmanın gereksiz olduğu maliyet şartı sağlanana kadar devam edilir [29].

Aynı yıl yapılan bir diğer çalışmada M. Dündar ve diğ. kümeleme ile örnekler arası küme eşleme problemleri için parametrik olmayan Bayes algoritması olarak ASPIRE adını verdikleri yöntemi anlatan çalışmalarını yayınlamışlardır [30]. 2016 yılında ise K. Johnsson ve diğ. Bayes hiyerarşik model temelli olan BayesFlow yöntemini önermişlerdir. GKM, normal ve ters Wishart dağılım yöntemleri Markov Zinciri Monte Karlo ile birlikte kullanılarak alt kümelere ayrılan veriler Bhattacharya uzaklığı kullanılarak birleştirilir. Bu şekilde belirli şekle sahip olmayan kümeler tespit edilebilir. Bhattacharya uzaklığı hesaplanırken GKM bileşenlerinin ağırlıkları katsayı olarak çarpılır. Çalışmada elde edilen sonuçlar yöntemin başarılı şekilde kümeleme ve kümeler arası eşleme yaptığını ortaya koymuştur [31]. 2018 Yılında Markus Lux ve diğ. FlowLearn adını verdikleri yarı gözetimli bir kapılama algoritması geliştirdikleri çalışmalarını yayınlamışlardır. Az sayıdaki manuel kapılanmış örneklerden yola çıkarak kümelerin yoğunluklarını her boyutta ayıran algoritmayla yapılan kapılamada yüksek başarı elde etmişlerdir [32].

A. Bashashati ve R. R. Brinkman 2009 yılında AS veri analizi aşamalarını anlatan ve bu aşamalarla ilgili çalışmaları derleyen bir derleme makalesi yayınlamışlardır. Çalışmada AS analiz adımları; kalite kontrolü, normalize etme, istenmeyen hücrelerin ayıklanması, gözetimli ve gözetimsiz otomatik kapılama yöntemleri, belirlenen kümeleri tanımlama ve elde edilen verileri yorumlama olmak üzere yedi adımda toplanmıştır. Buna göre yapılan çalışmaların %70'inden fazlası kapılama üzerinedir. Çalışmada kırkın üzerinde makale incelenmiştir [16]. 2010 yılında AS analizi üzerine yapılan çalışmaların daha sistematik ve verimli olması açısından *"The Flow Cytometry: Critical Assessment of Population identification methods (FlowCAP)"* projesi başlatılmış ve AS analizinde var olan problemlerin çözümü için toplu çağrılarda bulunulmuştur. Bu kapsamda N. Aghaeepour ve diğ. 2013 yılında yayınladıkları derleme makalede AS verilerinin otomatik kapılanması ve sınıflandırılması sonuçlarını paylaşmışlardır. Birçok ekibin katıldığı makalede kapılama ve sınıflandırma problemlerine

yüksek başarıyla çalışan yöntemler ortaya konulmuştur [18]. L. M. Weber ve M. D. Robinson 2016 yılında yayınladıkları derleme makalesinde ise güncel çalışmalardan elde edilen sonuçları FlowCAP sonuçları ile karşılaştırarak gelişmeleri göstermişlerdir [33].

3. MALZEME VE YÖNTEM

3.1. VERİ KÜMESİ

Bu tez çalışmasında N. Aghaeepour ve diğ. tarafından 2013 yılında yapılan FlowCAP [18] çalışmasında kullanılan diffüz büyük B hücreli lenfoma (Diffuse large B-cell lymphoma - DLBCL) veri kümesi kullanılmıştır. Veri kümesine http://flowrepository.org/ adresinden veri kümelerine özgü kimlik bilgisi girilerek ulaşılabilir. DLBCL veri kümesi için kimlik bilgisi FR-FCM-ZZYY şeklindedir. Bununla beraber verilerin manuel kapılanma sonuçlarına http://flowcap.flowsite.org/codeanddata/ adresinden ulaşılabilir. Veri kümesine ait özellikler özet halinde Tablo 3.1' de sunulmuştur.

Tablo 3.1: Akış Sitometrisi veri kümeleri özelliklerinin özeti

Veri Kümesi	Örnek Sayısı	Örnekte Bulunan Ortalama Hücre Sayısı	Ölçülen Özellikler
DLBCL	30	5000	İleri saçılım, Yan saçılım, CD3, CD5, CD19

DLBCL veri kümesi, 2003 ve 2008 yılları arasında British Columbia Kanser Kurumunda tedavi edilen hastalardan rastgele seçilen 30 lenf düğüm biyopsisinden elde edilen verilerden oluşmaktadır. Veri kümesindeki örnekler 1856 ile 24654 arasında olmak üzere ortalama 5000 hücre içermektedir. Burada her hücreye ait ön saçılım, yan saçılım, CD3, CD5, CD19 olmak üzere 5 özellik ölçülmüştür [18].

DLBCL veri kümesi kullanıma sunulmadan önce veri kümesi üzerinde üç önişlem yapılmıştır. İlk olarak veri kümesinde kullanılan floresan maddelerin yansıma dalga boylarındaki örtüşmeleri en aza indirgemek için düzeltme yapılmıştır. İkinci olarak verileri görsel olarak uygun şekilde ölçeklemek için verilere lineer dönüşüm uygulanmıştır. Son olarak verilerde bulunan gereksiz hücreler (örneğin ölü hücreler) ön kapılama yapılarak veri kümesinden çıkarılmıştır.

3.2. KÜMELEME ANALİZİ

Bilgisayar bilimlerinde kümeleme; gözlemlerin, verilerin, nesnelerin vb. sahip oldukları örüntüler incelenerek benzer olanların gruplanması olarak tanımlanabilir. Başka bir deyişle kümeleme gözetimsiz öğrenmedir [34]. Kümelemede verilerin sınıfları belli değildir. Sezgisel olarak aynı grupta yer alan verilerin diğer gruplarda yer alan verilere göre daha benzer örüntülere/özelliklere sahip olduğu kabul edilmektedir. Bununla birlikte birçok problemde veriler hakkında belirli miktarda ön bilgiye (verilerin istatistiksel dağılımı gibi) sahip olunmaktadır veya belli kurallar kabul (verilerin aynı kaynaktan üretildiği varsayımı gibi) edilmektedir. Daha sonra elde edilen başarı oranına göre önbilgilerin/kabullerin doğruluğu da değerlendirilir. Kümeleme örüntü tanıma, karar verme problemlerinde, veri madenciliğinde, veri kurtarmada, görüntü işlemede, biyoinformatikte vb. alanlarda sıklıkla kullanılır [34].

Kümeleme işlemleri genellikle şu adımları içerir; Verileri örüntülerle/özelliklerle ifade etme, probleme uygun yakınlık ölçüsü tanımlama, kümeleme veya gruplama, gerekiyorsa veri soyutlama/basitleştirme ve sonuçları değerlendirme. Karşılaşılan her problem için gerekli örüntüleri/özellikler belirleyen kesin kurallar yoktur. Bu sebeple örüntü belirleme aşaması kontrol edilebilir değildir. Bu aşama da problemle uğraşan kişini etkisi büyüktür. Verileri iyi bir şekilde ifade eden örüntülerle çalışmak, daha şekilde basit kümeleme işlemleri yaparak yüksek başarı elde edilir. Tersi durumda ise yüksek başarı elde etmek mümkün olmayabilir veya yüksek başarı elde etmek için oluşturulan kümeleme yöntemi çok karmaşık olabilir. Kümeleme yöntemleri benzerlik üzerine kurulu olduğundan, verileri kümelerken seçilen benzerlik/uzaklık ölçüm metodu hayati önem taşımaktadır. Benzerlik ölçümü için birçok yöntem kullanılmaktadır. Bu yöntemler arasında en çok bilineni Minkowski yönteminin (3.1) özel bir durumu olan Öklid (p=2) yöntemidir. Bununla beraber Mahalanobis benzerlik ölçüm yöntemi de oldukça sıklıkla kullanılmaktadır [34].

$$d_p(x_i, x_j) = \left(\sum_{k=1}^{d} (x_{i,k} - x_{j,k})^p \right)^{\frac{1}{p}} \tag{3.1}$$

Literatür de birçok kümeleme yöntemleri hiyerarşik/bölümlemeli, kesin/yumuşak, deterministik/skolastik gibi çeşitli kriterlere göre gruplanabilir [34]. Hiyerarşik kümeleme yöntemi verileri iç içe birleştirerek veya ayırarak kümeleyen yöntemlerdir. Hiyerarşik

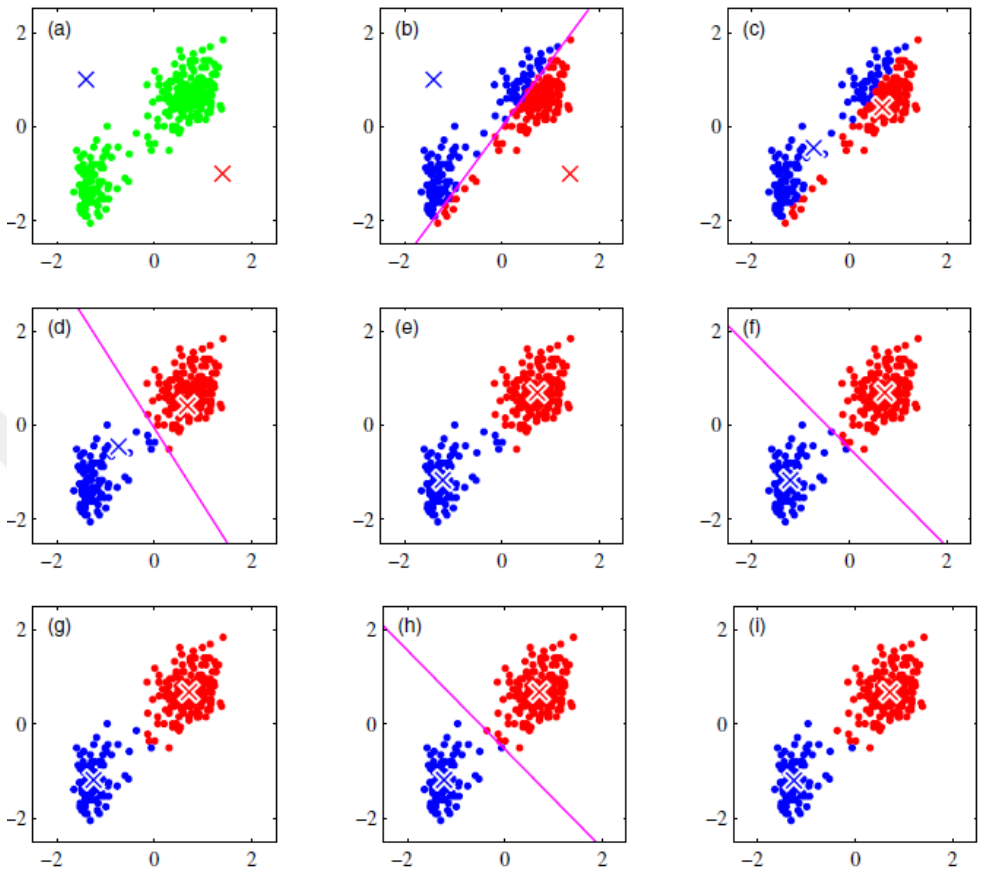
kümeleme temelde iki alt gruba ayrılır; yığınsal (agglomerative) ve bölen (divisive) kümeleme. Yığınsal kümele yönteminde her veri ayrı küme olarak kabul edilir ve her adımla benzer olduğu belirlenen veriler birleştirilerek tüm veriler tek bir küme oluncaya kadar ilerlenir. Bu yöntemde birleştirme mesafeleri ayrı kümelerin belirlenmesinde kullanılır. Bölen kümeleme de ise tüm veriler tek küme şeklinde kabul edilerek başlanır ve her iterasyonda benzer olmayan kümeler ayrılarak devam edilir. Bölümlemeli kümeleme ise sert ve yumuşak kümeleme olarak iki alt gruba ayrılabilir. Sert kümele yönteminde her veri yalnızca bir kümeye aittir. Bu grupta en çok kullanılan yöntem k-ortalamalar (k-means) algoritmasıdır. Yumuşak kümeleme yönteminde ise bir veri ağırlıklandırılmış olarak birden çok kümeye ait olabilir [35]. Bu gruba örnek olarak istatistik dağılımlarından faydalanılarak ortaya çıkarılan sonlu karışım modelleri (Finite Mixture models) örnek verilebilir. Tez kapsamında yapılan çalışmada k-ortalamalar ve GKM yöntemleri kullanıldığı için bu yöntemler ayrıntılı olarak anlatılacaktır.

3.2.1. K-ortalamalar Algoritması

Kümeleme yöntemleri arasına en çok kullanılan yöntem olan k-ortalamalar yöntemi 1957 yılında S. Lloyd tarafından önerilmiştir [36]. Diğer yöntemlere göre uygulanması daha kolay olan k-ortalamalar yönteminde her veri sadece bir kümeye ait olabilir. Algoritmada her küme, küme merkezi ile temsil edilir. Yeni gelen verinin hangi kümeye ait olacağı bu küme merkezlerine olan uzaklığına bağlıdır. Algoritmaya ismini veren k parametresi küme sayısını ifade eden bir tamsayıdır. Bu parametre algoritma çalışmaya başlamadan önce belirlenir ve kümeleme işlemi bitinceye kadar sabit kalır. Elimizde $\{x_1, x_2, x_3, \dots x_n\}$ Olmak üzere N adet D boyutlu veri kümemiz olsun. Bu veriler k adet kümeye ayrılmak isteniyor olsun. Bu durumda $\{u_1, u_2, u_3, \dots u_k\}$ olmak üzere k tane D boyutlu küme merkezi olur. Her verinin ait olduğu küme etiketi $j = 1,2 \dots k$ olmak üzere c_j şeklinde gösterilsin. Buna göre algoritma adımları;

1. Küme sayısını ifade eden K sayısını seç.
2. Rastgele olarak k tane D boyutlu küme merkezi belirle.
3. Küme merkezleri sabitleninceye kadar veya belirli bir eşik değerinin altında yer değiştirinceye kadar 4. ve 5. adımları tekrar et.
4. Her veri için $j_n = \underset{k}{\text{Argmin}} \sqrt[2]{\sum_{i=1}^{D}\left(x_i - u_i^k\right)^2}$
5. Her küme merkezi için $u_k = \{\sum_{i=1}^{N}(j_n == k)\ x_i\}/\{\sum_{i=1}^{N}(j_n == k)\ 1\}$

Algoritma k adet küme merkezinin rastgele olarak belirlenmesiyle başlar. Daha sonra her verinin küme merkezlerine olan uzaklığı uzaklık ölçüm metrikleriyle ölçülür. Uzaklık ölçümü



Şekil 3.1: k-ortalamalar kümeleme döngüsü. a) ilk adım. b) ilk döngü. Her veri için küme merkezlerine olan uzaklık hesaplandı ve veriler yakın oldukları kümelere atandı. c) yeni küme merkezleri hesaplandı. d,e,f,g,h,i) küme merkezleri sabitleninceye kadar döngü devam eder [37].

için genellikle Öklid uzaklığı kullanılır. Bu işlem 4. Adımda gösterilmiştir. Her verinin D boyutta tek tek tüm kümelere olan uzaklığı hesaplanmıştır. Çıkan sonuçlara göre veri en yakın olduğu kümeye atanır. Bu adım sonunda kümelere atanan veriler değişiklik gösterir. 5. Adımda küme merkezleri tekrar hesaplanır. Bu hesaplama kümeye atanan veri değerlerinin toplamının veri sayısına bölünmesiyle bulunur. Bu işlemler küme merkezlerinin değişmesi bitinceye kadar veya belirlenen döngü sayısına ulaşılıncaya kadar tekrar edilir. Şekil 3.1 'de verileri hizalanmış (0 ortalama ve birim standart sapma) Old Faithful veri kümesi ile k=2 için yapılan k-ortalamalar algoritmasının çalışma adımları görülmektedir. İki boyutlu olan bu veri kümesi püskürme ve bekleme zamanlarını ifade eden değişkenlere sahiptir.

3.2.2. Gauss Karışım Modeli

Tez kapsamında kullanılan GKM temelde olasılık teorisini kullanan bir yöntemdir. Bu sebeple GKM modelini açıklamadan önce modelde kullanılan temel olasılık kavramları tanımlanmış ve normal dağılım açıklanmıştır.

3.2.2.1. Olasılık Kavramları

Olasılık gerçekleşmesi istenen olayların verilen şartlara bağlı olarak meydana gelme oranlarıyla ilgilenir. Yapılan deneyler sonucunda ortaya çıkan tüm sonuçlar yerine belirli sonuçlarla ilgilenilebilir. Olasılıkta bunun gibi durumları inceleyebilmek için rastgele değişkenler (random variables) kullanılır. Rastgele değişkenler olasılık örneklem uzayından gerçek sayılar kümesine aktarım yapan basit bir fonksiyon olarak düşünülebilir. Rastgele değişkenler kesikli ve sürekli olmak üzere 2 adettir. Kesikli rastgele değişkenler sayılabilir olasılık değerleri alabilen değişkenlerdir. Sürekli rastgele değişkenler olasılık uzayının tüm değerlerini alabilen değişkenlerdir. Rastgele değişkenler X,Y gibi büyük harflerle gösterilirken aldığı değerler ise x, y gibi küçük harflerle gösterilir. Bir rastgele değişkenin olma olasılığı ise P(X=x) şeklinde gösterilir.

Rastgele değişkenlerin olasılıklarını hesaplamak için olasılık fonksiyonları kullanılır. Bu fonksiyonlar kesikli rastgele değişkenler için olasılık fonksiyonu, sürekli rastgele değişkenler için olasılık yoğunluk fonksiyonu olarak adlandırılır. Verilen fonksiyonların olasılık fonksiyonu olabilmesi için;

1. Tüm rastgele değişkenler için elde edilen olasılık değerleri 0-1 aralığında olmalıdır.
2. Tüm olasılık değerleri toplamı 1'e eşit olmalıdır.

Kesikli rastgele değişkenlerin olasılığını hesaplamak için $\sum$ sembolü kullanılır. Yani olasılık fonksiyonları toplanır. Sürekli rastgele değişkenlerin olasılığını hesaplamak için olasılık yoğunluk fonksiyonunun integrali ($\int$) alınır. Yani olasılık yoğunluk fonksiyonunun altında kalan alan olasılığa eşittir.

3.2.2.2. Normal Dağılım

Normal dağılım veya gauss dağılımı hemen hemen her alanda kullanılmakta olan bir sürekli rastgele değişken olasılık dağılımıdır. Dağılım ilk olarak 1733 yılında Abraham de Moivre

tarafından ortaya çıkarılmıştır. C.F. Gauss 'un bu dağılım üzerinde yaptığı çalışmalar ve geliştirmelerden dolayı gauss dağılımı olarak da bilinir. Tek boyutlu X sürekli rastgele değişkeni normal dağılıma sahipse X~$N(\mu,\sigma^2)$ şeklinde gösterilir. $N(\mu,\sigma^2)$ normal dağılım için olasılık yoğunluk fonksiyonu olup (3.2)'deki gibi tanımlanır.

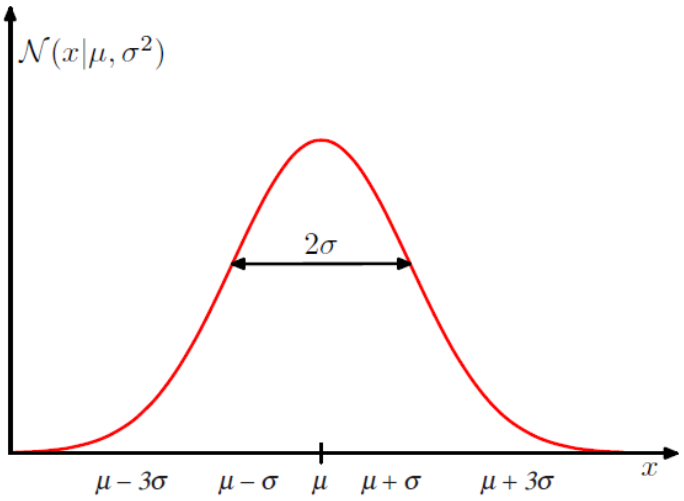
$$N(\mathrm{X} \mid \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} e^{(-1/2\sigma^2)(x-\mu)^2} \tag{3.2}$$

(3.2)'de de görüleceği üzere normal dağılım 2 parametreye sahiptir. Bu parametreler dağılımın ortalamasını temsil eden μ ve varyansını temsil eden σ parametreleridir. Normal dağılımda beklenen değer ortalamaya eşittir ve formülü (3.3)'te gösterilmiştir.

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} N(\mathrm{X} \mid \mu, \sigma^2) x dx = \mu \tag{3.3}$$

Formülü (3.4)'de gösterilen varyans ortalamaya yani beklenen değere bağlı olarak bulunur [37][38].

$$Var(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \sigma^2 \tag{3.4}$$



Şekil 3.2: Tek boyutlu normal dağılım [37, 38]

Şekil 3.2' de tek boyutlu normal dağılım grafiği gözükmektedir. Burada μ verilerin ortalamasını, σ ise varyansını temsil etmektedir. Tek boyutlu normal dağılım çan eğrisine

benzetilirken (Şekil 3.2), iki boyutlu normal dağılım tek tepeli bir dağa benzetilebilir. Normal dağılımda X değişkenine ait gözlemlerin yaklaşık %68'i ortalamadan 1 standart sapma kadar uzaklığın içerisinde, yaklaşık %95'i ortalamadan 2 standart sapma kadar uzaklığın içerisinde, yaklaşık %99'u da ortalamadan 3 standart sapma kadar uzaklığın içerisinde yer alır. Çok boyutlu normal dağılımda varyans kovaryans halini alır. Kovaryans değişkenlerin ilişkileri hakkında bilgi verir. İki değişken birbirlerini olumlu yönde etkiliyorsa değeri pozitif, olumsuz yönde etkiliyorsa değeri negatiftir.

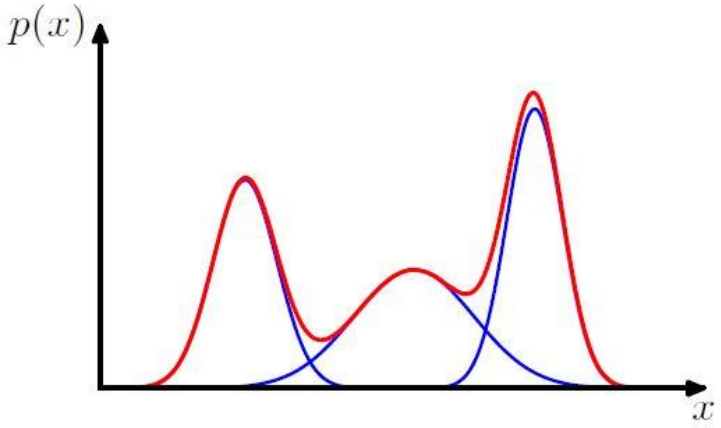
$$N(\mathrm{X} \mid \mu, \sum) = \frac{1}{(2\pi)^{D/2}|\sum|^{1/2}} e^{(-1/2)(x-\mu)^T \sum^{-1}(x-\mu)} \tag{3.5}$$

D boyutlu X rastgele değişkeni için oluşturulan normal dağılımın olasılık yoğunluk fonksiyonu (3.5)'de gösterilmiştir. Çok boyutlu normal dağılımda ortalama bir vektör ile ifade edilir. D boyutlu normal dağılım için μ ortalama vektörü de D boyutludur. $\sum$ ile ifade edilen kovaryans $D \times D$ şeklinde bir matristir ve $|\sum|$ bu matrisin determinantıdır.

3.2.2.3.Gauss karışım Modeli (GKM)

Normal dağılım ile birlikte diğer dağılımlar veri kümelerini modelleme de sıklıkla kullanılırlar. Bu dağılımlar birçok veri kümesini yüksek başarı ile modelleyebilmektedirler. Fakat bununla birlikte tek bir dağılımla ifade edilemeyen veri kümelerini tek bir dağılımla modellemek iyi sonuçlar vermemektedir. Bu problemi çözmek için karışım modellerini önerilmiştir. Karışım modeli bir veri kümesini modellemek için birden çok dağılımı belirli ağırlıklarla beraber kullanarak olasılık yoğunluk fonksiyonu oluşturur [37]. Şekil 3.3'te bir veri kümesinin dağılımı gözükmektedir. Burada x değişken değerlerini, $p(x)$ normal dağılım olasılık yoğunluk fonksiyonunu göstermektedir. Bu veri kümesi üç ayrı normal dağılıma uyan üç ayrı veri kümesi gibi düşünülebilir. Daha sonra bu iç normal dağılım içerdikleri veri sayısıyla orantılı olarak bir ağırlığa sahip olurlar. Böylece veri üç ağırlığa sahip üç normal dağılımın birleşimi şeklinde ifade edilebilir. Bu şekilde karışım modelleri ile daha karmaşık dağılım modelleri elde edilebilir. Gauss Karışım Modeli, karışım modellerinde verilerin normal dağılıma sahip olduğunu kabul eden karışım modelidir.

$$p(x) = \sum_{k=1}^{K} w_k N(x|\mu_k, \sum_k) \tag{3.6}$$



Şekil 3.3: Üç bileşenli iki boyutlu gauss karışım modeli [37]

GKM'nin olasılık yoğunluk fonksiyonu normal dağılımdan türetilmektedir. Çok boyutlu GKM formülü (3.6)'da verilmiştir. Burada $N(x|\mu_k, \sum_k)$ normal dağılımına model bileşeni denir ve bu dağılımın kendi ortalaması ve kovaryansı vardır. K sayısı modelde kaç adet bileşen olduğunu ifade eden bir tamsayıdır ve model çalışmadan önce belirlenir. w_k katsayısı her bileşenin modeldeki oranını ifade eder ve bileşeni oluşturan değişken sayısı ile doğru orantılıdır. Tüm w_k katsayılarının toplamı 1'e eşittir(3.7). Dolayısıyla w_k, $0 \leq w_k \leq 1$ aralığında değer alır [37].

$$\sum_{k=1}^{K} w_k = 1 \tag{3.7}$$

GKM gauss bileşenlerinin ağırlıklı toplamı olarak görülebilir. Modeli oluştururken her bileşene ait w, μ ve $\sum$ parametreleri belirlenir. Yani her olasılık yoğunluk fonksiyonunun parametre ve katsayı değerleri tahmin edilir. Bu parametrelerin bulunmasında en büyük olabilirlik (maximum likelihood) yönteminden yararlanılır. En büyük olabilirlik yöntemi mevcut rastgele değişkenlerin gerçekleşme olasılığının en yüksek olduğu dağılım parametrelerini bulmayı amaçlayan bir kestirim yöntemidir. En büyük olabilirlik yöntemini logaritma fonksiyonu ile birlikte kullanılır. Bunun sebebi logaritma fonksiyonu monoton artan bir fonksiyondur. Bu sebeple logaritma fonksiyonunu en büyük yapan değer aynı zamanda en büyük olabilirlik yöntemini de en büyük yapar [38]. N adet gözlemden oluşan GKM için en büyük olabilirlik logaritma fonksiyonu (3.8)'de görülmektedir.

$$\log N(X|w_k\mu_k,\sum_k) = \sum_{n=1}^{N} \log(\sum_{k=1}^{K} w_k N(x_n|\mu_k,\sum_k)) \quad (3.8)$$

(3.8)'deki denklemi maksimum yapan değeri bulmak için türevi alındığında matematiksel olarak kapalı forma sahip bir çözüme ulaşılamaz. Bu sebeple parametreleri kestirmek için iteratif bir yöntem olan Beklenti Eniyileme algoritması kullanılır.

3.2.2.4. Beklenti Eniyileme (BE) Algoritması

BE algoritması karışım modellerinin parametrelerini bulmak için geliştirilmiş bir algoritmadır. Algoritma iki temel adımdan oluşmaktadır; beklenti adımı ve eniyileme adımı. Beklenti adımında bileşenlerin veriler üzerindeki sorumlulukları (sonsal olasılık) bulunur. Yani x verilerinin k bileşenine ait olma olasılığı hesaplanır. Çünkü Model parametrelerini kestirmek için verilerin hangi sınıflara ait olduğu bilgisine ihtiyaç duyulur. Eniyileme adımında da verilerin bileşenlere ait olma olasılıkları üzerinden w, μ ve $\sum$ parametreleri hesaplanır.

Verilen bir gauss karışım modeli için BE algoritmasının amacı en (3.8)'de verilen en büyük olabilirlik fonksiyonunu w, μ ve $\sum$ parametrelerine göre maksimum yapmaktır. BE algoritması dört temel adımdan oluşur [37]. Bu adımlar;

1. Bileşen sayısı(K) kadar rastgele olarak w_k, μ_k ve $\sum_k$ parametrelerine değer atanır ve (3.8)'de verilen en büyük olabilirlik logaritma fonksiyonunun başlangıç değerini hesaplanır.
2. Beklenti adımı: mevcut parametreler üzerinden bileşenlerin sorumlulukları hesaplanır. Yani bileşenlerin verileri içerme olasılıkları hesaplanır. Tüm veriler için bu işlem yapılır. Bu adımda her verinin karışım bileşenlerine ait olma olasılıkları hesaplanır.

$$p(k|x_n) = \frac{w_k N(x_n|\mu_k,\sum_k)}{\sum_{j=1}^{K} w_j N(x_n|\mu_j,\sum_j)} \quad (3.9)$$

3. Eniyileme adımı: Adım 2'de bulunan bileşenlere ait olma olasılıklarından bileşenlerin yeni parametre değerleri hesaplanır.

$$\mu_k^{yeni} = \frac{\sum_{n=1}^{N}[p(k|x_n)x_n]}{\sum_{n=1}^{N} p(k|x_n)} \quad (3.10)$$

$$\sum_k^{yeni} = \frac{\sum_{n=1}^{N}\left[p(k|x_n)(x_n - \mu_k^{yeni})(x_n - \mu_k^{yeni})^T\right]}{\sum_{n=1}^{N} p(k|x_n)} \quad (3.11)$$

$$w_k^{yeni} = \frac{\sum_{n=1}^{N} p(k|x_n)}{N} \quad (3.12)$$

4. Yeni belirlenen w, μ ve $\sum$ parametre değerlerine göre (3.8)'de verilen en büyük olabilirlik fonksiyonunu hesaplanır. Durma şartı sağlanırsa algoritma çalışmasını bitirir. Durma şartı sağlanmazsa algoritma 2. adıma geri döner.

Adım 3'e hesaplanan $\sum_k^{yeni}$ değeri μ_k^{yeni} değerine bağlı olduğundan önce μ_k^{yeni} değerinin hesaplanması gerekir. Algoritmanın durma şartı olarak önceki adımda hesaplanan en büyük olabilirlik değeri ile şuanda bulunan değer arasındaki fark karşılaştırılır. Bu fark belirlenen bir sınır değerin altında ise karışım modelinin parametreleri belirlenmiş sayılır ve algoritma çalışmasını bitirir. Diğer bir durma kriteri ise iterasyon sayısına sınır koyma olabilir. Böylece algoritma en fazla o sayıya kadar çalışır. Algoritmanın ilk adımında rastgele belirlenen parametre değerleri başarı oranını ve algoritmanın çalışma süresini etkileyen faktörlerdir. En büyük olabilirlik fonksiyonunu maksimum yapan parametrelere uzak seçilen değerler algoritmanın çalışma zamanını artırır. Bununla birlikte algoritma yerel maksimuma ulaşıp mutlak maksimuma varamayabilir.

BE algoritması ile birlikte kullanılan GKM yöntemi k-ortalamalar algoritmasına benzer. Eğer GKM kümeleme için kullanılırsa her bileşen bir sınıfa karşılık gelir. Aradaki fark GKM'de bir veri birden çok bileşene ait olabilir. Bu şekilde bulanık (fuzzy) kümeleme yapılabilir. Karışım modellerinin veri modelleme üzerinde sağladığı bu esneklik birçok alanda kullanılmasını sağlamıştır. Tıp, genetik, endüstri, mühendislik gibi alanlarda var olan birçok problemde karışım modelleri kullanılır. AS verilerini üzerine yapılan çalışmalarda da karışım modellerini kullanmak iyi sonuçlar üretmiştir. Karışım modelleri AS verilerini dağılım değişkenleri olarak kabul eder. Tek bir dağılım tarafından modellenemeyen veriler karışım modelleri sayesinde daha başarılı olarak modellenip kümelenmektedir. Literatür taramasında da bahsedildiği gibi bu yönde yapılan çalışmalar mevcuttur.

3.3. BİLEŞENLERİ BİRLEŞTİRME

AS verilerinin kapılanması üzerine dağılım modelleri bileşenlerini birleştirerek ilerleyen çalışmalar bulunmaktadır [21, 24, 25]. Bu sayede hücre gruplarını doğru olarak bulmayı

amaçlamışlardır. Tüm yöntemlerin ortak amacı aynı kümeyi temsil eden bileşenleri ortaya çıkarmaktır. Birleştirme işlemleri bileşenlerin benzerlik oranları, bileşenler arası mesafeler, entropi değerleri gibi farklı yöntemler üzerinden yapılabilir. Bu tez kapsamında bileşenleri birleştirme yöntemi olarak Chernoff mesafesi kullanılmıştır.

3.3.1. Chernoff Mesafesi

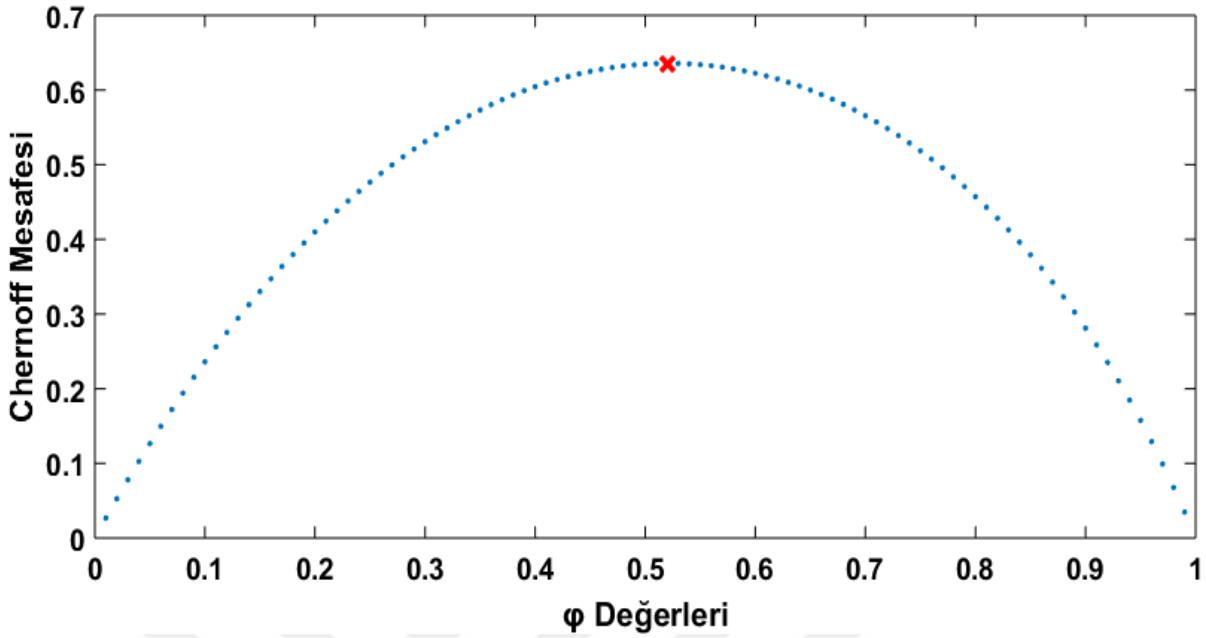
Rastgele değişkenler, olasılık dağılımları gibi olasılıksal nesneler arası mesafeler ölçülürken istatistiksel mesafe ölçüm yöntemleri kullanılır. Veri kümeleri olasılık dağılımları ile ifade edildiğinde veri kümeleri arasındaki mesafe dağılımlar arası mesafe şeklinde yorumlanabilir. Chernoff mesafesi, Patrick-Fischer mesafesi, Mahalanobis mesafesi, KL mesafesi gibi çeşitli mesafe ölçüm yöntemleri mevcuttur. Olasılıksal mesafe ölçümlerini hesaplamak oldukça zordur. Bununla beraber gauss dağılımı gibi bazı dağılımlar için mesafe ölçüm yöntemleri matematiksel olarak ifade edilebilmiştir [39]. Parametreleri $\mu_1, \sum_1$ olan bir P_1 normal dağılımı ile parametreleri $\mu_2, \sum_2$ olan bir P_2 normal dağılım arasındaki Chernoff mesafesi (3.13)'te verilmiştir. $|\sum|$ kovaryansın determinantını ifade eder. Denklemde görülen φ bir sabit sayıdır ve $0 < \varphi < 1$ şartını sağlamalıdır [39]. Optimum φ, değeri $Cher(P_1, P_2)$ sonucunu maksimum yapan değerdir.

$$Cher(P_1, P_2) = {}^{1}/_{2}\,\varphi(1-\varphi)(\mu_1-\mu_2)^T[\varphi\sum_1 + (1-\varphi)\sum_2]^{-1}(\mu_1-\mu_2) + {}^{1}/_{2}\log\left({}^{[\varphi\sum_1 + (1-\varphi)\sum_2]}/_{[|\sum_1|^{\varphi}|\sum_2|^{(1-\varphi)}]}\right) \quad (3.13)$$

Şekil 3.4'te örnek olarak $P_1([2,3],[1,1.5;1.5,3])$, $P_2([0,1],[1,0.5;0.5,2])$ dağılımları arası Chernoff mesafesi $0 < \varphi < 1$ arası φ değerleri için oluşturulan grafik görülmektedir. Mesafeyi maksimum yapan değer kırmızı renkle işaretlenmiştir. Bu değer $\varphi = 0.52$ için elde edilmiştir. Bu şekilde Chernoff mesafesi bulunmuş olur. Buradan da görüleceği gibi maksimum Chernoff mesafesi çoğunlukla φ değer 0.5 civarında bir değer aldığında elde edilir.

Optimum φ değerini daha az hesaplama yaparak bulmak için ikili arama algoritmasına benzer bir yol izlenmiştir. Bu yolun adımları aşağıda verilmiştir.

1. $\varphi = 0.01$ ile $\varphi = 0.99$ değerlerine göre Chernoff mesafesi hesaplanır.
2. Küçük Chernoff mesafesini veren φ değeri yerine adım 1'deki değerlerin ortalaması alınır. Böylece küçük sonucu veren φ değeri elenir.



Şekil 3.4: İki dağılım için farklı φ değerleri için hesaplanan Chernoff mesafeleri

3. Bu döngü iki φ değeri arasındaki fark 0.01 olana kadar devam ettirilir.

3.4.BAYES BİLGİ KRİTERİ

Model tabanlı kümeleme üzerinde çalışılan verileri belirli bir modele uydurma üzerine kuruludur. Bu sebeple kümelemeyi başarılı şekilde yapabilmek için verileri mümkün olan en iyi şekilde ifade eden karışım modeli bulunmaya çalışılır [40]. Karışım modellerinde bileşen sayısı kullanıcı tarafından belirlenen sabit bir değerdir. Bu durum seçilen modelin uygulanmasında kısıtlamalar oluşturabilir. Çünkü verinin kaç adet küme içerdiği birçok durumda belli değildir. Bu sebeple başarı oranını etkileyen bu değeri belirlemeye yönelik yöntemler önerilmiştir. Bayes Bilgi Kriteri (Bayesian Information Criteria - BBK) önerilen yöntemlerden bir tanesidir.

BBK karışım modellerinde bileşen sayısını tahmin etmek için kullanılan bir yöntemdir. Yöntem temelde en büyük olabilirlik metodunu kullanır. Bileşen sayısını belirlemek etmek için BBK değerleri hesaplanacak değerler kümesi $K = \{k_n, k_{n+1}, \cdots, k_m\}$ olacak şekilde verilsin. Bu durumda BBK formülü (3.14)'te görüldüğü gibi yazılır. Karışım modelinin bileşen sayısı BBK formülünde en küçük değeri veren k_n sayısıdır.

$$BBK(K) = \underset{k_n \in K}{\text{Argmin}} \left(-\log\left(p(X|k_n)\right) + \frac{v_{k_n}}{2} \log N \right) \quad (3.14)$$

Burada $\log\left(p(X|k_n)\right)$ ifadesi k_n bileşenli modelin en büyük olabilirlik logaritma değeridir. En büyük olabilirlik değerinin formülü (3.8)'de verilişitir. N değeri modeldeki veri sayısını, v_{k_n} ise k_n bileşenli modelde bulunan serbest parametre (free parameter) sayısını ifade eder [40].

$$v_{k_n} = (k_n \times D_k) - 1 \quad (3.15)$$

D boyutlu k_n bileşenli GKM için v_{k_n} serbest parametre sayısı (3.15)'te gösterildiği gibi bulunur. Denklemde görülen D_k bir gauss bileşeninde bulunan parametre sayısını ifade eder ve aşağıda verilen 3 adım sonucu bulunur [41].

1. Bir gauss bileşeninin kovaryans matrisi $\frac{D \times D - D}{2} + D$ sayısı kadar parametre içerir. Burada $\frac{D \times D - D}{2}$ diyagonal olmayan eleman sayısını, D diyagonal eleman sayısını ifade eder.
2. Gauss bileşeninin ortalama vektörü veri boyutu kadar yani D sayısı kadar parametre içerir.
3. Her gauss bileşeni bileşendeki ağırlığını ifade eden oran (w) parametresine sahiptir.

$$D_k = \frac{D \times D - D}{2} + 2 \times D + 1 \quad (3.16)$$

Buna göre D_k denklem (3.16)'da görüldüğü gibi ifade edilebilir. Buna göre serbest parametre sayısı her bileşendeki parametre sayısının bileşen sayısıyla çarpılması sonucu bulunur. Bu sonuçtan 1 çıkarılmasının sebebi model ağırlıkların toplamı 1'e eşittir. $(k_n - 1)$ tane ağırlık bulunduğu zaman geriye kalan son ağırlık 1'den çıkarılarak bulunabilir [41].

3.5. BİRLEŞİK SINIFLANDIRMA OLABİLİRLİK

Karışım modelindeki bileşen sayısı veriyi en iyi şekilde ifade eden sayı olması gerektiği önceki bölümde belirtilmişti. Fakat karışım modelindeki her bileşen bir kümeyi ifade etmeyebilir. Yani bileşen sayısı ile küme sayısı farklı olabilir. Bölüm 3.4'te anlatılan BBK çoğunlukla modeldeki bileşen sayısını bulmak için kullanılır. Verilerin bulunan bileşen sayısından farklı sayıda kümeye sahip olduğu durumlarda BBK yetersiz kalabilir. Bu durumlarda küme sayısını

belirlemek için Birleşik Sınıflandırma Olabilirlik (Integrated Classification Likelihood - BSO) kullanılabilir. Verilere göre değişmekle birlikte BSO sıklıkla BBK'den daha küçük bileşen değeri seçer. Bu durumda bir küme birden çok normal dağılımla yani GKM ifade edilebilir [40].

BSO BBK'ye oldukça benzemekle birlikte farkı modelin entropisini de hesaplar. Karışımda bulunan kümeler birbirlerinden belirgin bir şekilde ayrıysa modelin entropisi minimuma yani sıfıra yaklaşır. Aksi durumda entropi değeri büyür. Buna göre N adet veriye sahip K bileşenli bir modelin entropi değeri (3.17) uygulanarak bulunur [42]. Denklemdeki $p(k|x_n)$ ifadesi verilerin kümeye ait olma olasılığını ifade eder ve (3.9)'da gösterildiği gibi hesaplanır.

$$ENT = -\sum_{k=1}^{K}\sum_{n=1}^{N} p(k|x_n)\log p(k|x_n) \tag{3.17}$$

$$BSO(K) = \underset{k_n \in K}{\text{Argmin}}\left(-\log\big(p(X|k_n)\big) + \frac{v_{k_n}}{2}\log N + ENT\right) \tag{3.18}$$

Küme sayısını belirlemek etmek için BSO değerleri hesaplanacak değerler kümesi $K = \{k_n, k_{n+1}, \cdots, k_m\}$ olacak şekilde verilsin. Bu durumda BSO formülü (3.18)'te görüldüğü gibi yazılır[42]. Karışım modelinin küme sayısı BSO formülünde en küçük değeri veren k_n sayısıdır.

3.6. DEĞERLENDİRME YÖNTEMİ

Oluşturulan kümeleme yöntemlerinin başarı oranları manuel kapılama ile karşılaştırılarak ölçülmüştür. Manuel kapılama uzmanlar tarafından iki boyutlu saçılım grafiklerini incelenerek yapılmıştır. Bu bağlamda başarı oranı F-skor yöntemi temel alınarak ölçülmüştür [21]. Burada kullanılan F-skor yöntemi kesinlik (precision) ile hassasiyetin (recall) harmonik ortalaması gibi düşünülebilir [18]. F-skor yönteminin formülü (3.19)'te verilmiştir.

$$F = \frac{2 \times (Kesinlik) \times (Hassasiyet)}{(Kesinlik) + (Hassasiyet)} \tag{3.19}$$

Kesinlik ve hassasiyet terimlerini anlamak için birkaç kavramı tanımlamak gerekir. Bunlar; gerçek pozitif (gp), yanlış pozitif (yp) ve yanlış negatif (yn). Gerçek pozitif, gerçekte pozitif değerli olup algoritma tarafından da pozitif olarak değerlendirilme durumudur. Yanlış pozitif, gerçekte negatif değerli olup algoritma tarafından pozitif olarak değerlendirilme durumudur.

Yanlış negatif, gerçekte pozitif olup algoritma tarafından negatif olarak değerlendirilme durumunu ifade eder. Kesinlik ve hassasiyet kavramlarının formülü (3.20)'te verilmiştir.

$$Kesinlik = \frac{gp}{gp + yp} \quad , \quad Hassasiyet = \frac{gp}{gp + yn} \tag{3.20}$$

Kapılama için yapılan değerlendirmede kesinlik bir kümeye doğru şekilde atanan hücre sayısının o kümeye atanan toplam hücre sayısına bölünmesi anlamına gelir. Hassasiyet ise bir kümeye doğru şekilde atanan hücre sayısının o kümeye atanması gereken tüm hücrelerin sayısına bölünmesi anlamına gelir [18].

Algoritma tarafından bir örnek için oluşturulan kümeler $C = \{c_1, c_2, \cdots, c_k\}$ şeklinde, o örneğe ait gerçek kümeler $T = \{t_1, t_2, \cdots, t_m\}$ ise şeklinde ifade edilsin. Buna göre $|c_i|$ ve $|t_j|$, sırasıyla c_i ve t_j kümelerindeki hücre sayısını, $|c_i \cap t_j|$ ise c_i ve t_j kümelerinin her ikisine de ait olan hücre sayısını göstermektedir. c_i ve t_j kümeleri arasındaki kesinlik ve hassasiyet (3.21)'da, F-skor ise (3.22)'de verilmiştir[11].

$$Kesinlik(c_i, t_j) = \frac{|c_i \cap t_j|}{|c_i|} \quad , \quad Hassasiyet(c_i, t_j) = \frac{|c_i \cap t_j|}{|t_j|} \tag{3.21}$$

$$F(c_i, t_j) = \frac{2 \times \left(Kesinlik(c_i, t_j)\right) \times \left(Hassasiyet(c_i, t_j)\right)}{\left(Kesinlik(c_i, t_j)\right) + \left(Hassasiyet(c_i, t_j)\right)} \tag{3.22}$$

Buna göre C ve T arasındaki F-skor formülü (3.23)' de verilmiştir [11].

$$F(C, T) = \sum_{t_j \in T} \frac{t_j}{n} \max_{c_i \in C} \{F(c_i, t_j)\} \tag{3.23}$$

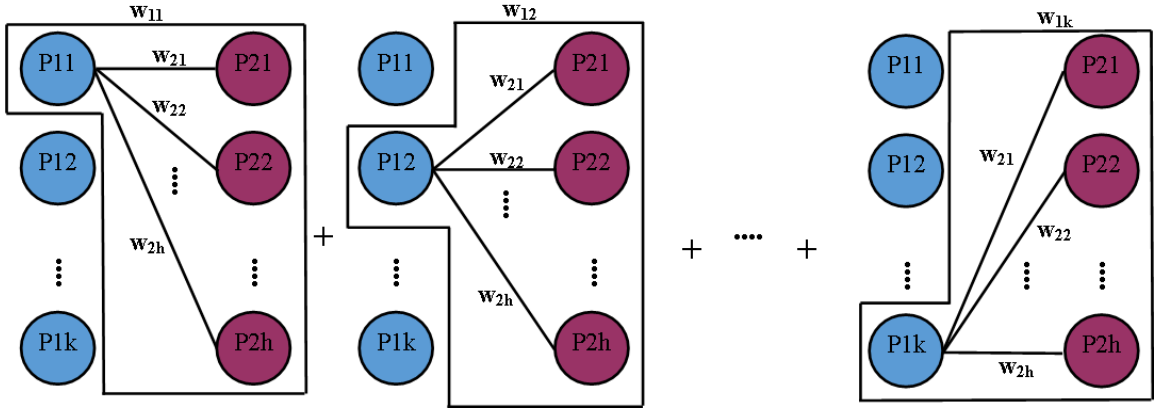
Denklemde n toplam hücre sayısını ifade eder. Yöntem tüm C, T kümeleri için (3.22)'ye göre F-skor değerini hesaplar. Bu değerlerden maksimum olan küme çiftleri eşlenir. Daha sonra her F-skor değeri, t_j kümesinin tüm hücrelere bölümü ile çarpılarak toplanır (3.23). Elde edilen sonuç kapılama başarısını göstermektedir.

3.7. ÖNERİLEN KÜMELEME YÖNTEMİ

Tez çalışmasında seçilen veri kümelerini kümelemek için ağırlıklı GKM birleştirme adını verdiğimiz yöntem kullanılmıştır. Algoritmanın akış diyagramı EKLER bölümünde yer almaktadır. Ağırlıklı GKM birleştirme yöntemi 9 adımdan oluşmaktadır. Bu adımlar;

1. Veri kümesinin bir örneğini kümelemek için seç.
2. BBK ile seçilen örnekteki bileşen sayısını tahmin et.
3. k= 2. adımda tahmin edilen bileşen sayısı olmak üzere, örneği k-ortalamalar algoritması ile kümelere ayır.
4. 3. Adımda bulunan kümeleri çok boyutlu normal dağılım ile ifade et.
5. Örnekteki her küme çiftinin birbirlerine olan uzaklığını denklem (3.24)'de verilen formülle hesapla. Denklem iki karışım dağılımı arasındaki uzaklığı karışımlardaki normal dağılımların ağırlıkları ile çarpılarak toplanması şeklinde ifade edilebilir.

$$Mesafe(P_1, P_2) = \sum_{k}^{P_1} \left(\sum_{h}^{P_2} [Cher(P_{1k}, P_{2h})] \times w_{2h} \right) \times w_{1k} \qquad (3.24)$$



Şekil 3.5: k bileşenli P_1 karışımı ile h bileşenli P_2 karışımı arasındaki uzaklık hesabı

Örneğin k bileşenli P_1 karışım modeli ile h bileşenli P_2 karışım modeli olsun. Buna göre P_1'deki her dağılımın P_2'deki tüm dağılımlara olan uzaklığı ayrı ayrı Chernoff mesafesiyle hesaplanır. Elde edilen sonuç P_2'deki dağılımların ağırlıkları ile çarpılır ve tüm sonuçlar toplanır. Ortaya çıkan bu sonuç P_1 'deki dağılımın ağırlığıyla çarpılır. Bu

işlem P_1'deki tüm dağılımlar için yapılır. Ortaya çıkan sonuçlar toplanarak P_1, P_2 arasındaki nihai uzaklık hesaplanır. Şekil 3.5'te örnek hesaplama gösterilmiştir.

6. 5. adım sonucu oluşan matris içerisinden en küçük değere sahip olan küme çiftlerini birleştirmek için seç.
7. 6. adımda seçilen kümelerle oluşturulacak olan yeni kümeyi GKM ile ifade et. Burada oluşturulan bu yeni kümenin bileşen sayısı birleştirilmek üzere seçilen iki kümenin bileşen sayılarının toplamıdır ve kümenin parametreleri BE algoritmasıyla bulunur.
8. İstenilen küme sayısına ulaşılıp ulaşılmadığını kontrol et. Eğer ulaşıldıysa 9. adıma git, yoksa 5. adıma git. Bu adımda istenilen küme sayısı kullanıcı tarafından yöntem çalışmadan önce girilebilir veya BSO yöntemi ile bulunan değer küme sayısı olarak değerlendirilir.
9. F-skor ile kümeleme işlemi yapılan örneğin başarı oranını bul.

4. BULGULAR

Makine öğrenmesi kapsamında yapılan kümeleme analizinde verilerin kaç adet kümeye sahip olduğunu bulmak hala tam olarak çözülememiş önemli bir problem olarak durmaktadır. Bu durum AS verilerinin kapı/küme sayısını belirlemede de geçerlidir. Küme sayısını olması gerekenden az veya fazla tahmin etmek başarı oranını ve elde edilen sonuçların yorumlanmasını önemli ölçüde etkilemektedir. Bunu göz önünde bulundurarak önceki bölümde ayrıntılı biçimde açıklanan kapılama yöntemi DLBCL veri kümesi üzerinde 2 farklı test senaryosu ile denenmiştir. İlk deney senaryosunda yöntem verilerin kaç adet kümeye sahip olduklarını tahmin etmeye çalışarak kapılama yapmıştır. İkinci deney senaryosunda verilerin kaç adet kümeye sahip oldukları deney başlamadan önce yönteme giriş parametresi olarak verilmiştir. Yapılan deneylerde her bir örnek için başarı oranı F-skor ile değerlendirilmiştir. Veri kümelerinin kapılama başarısı ise F-skor değerlerinin ortalaması alınarak bulunmuştur.

4.1. PARAMETRE GİRİŞİ OLMAYAN KAPILAMA

Bu deneyde önerilen yöntem çalışmadan önce kullanıcıdan bir parametre girişi yapılmadı. Bölüm 3.7' de anlatılan kapılama yönteminin 8. Adımında ki durma kriteri olan küme sayısı BSO yöntemiyle tahmin edilmiştir. Bununla beraber her örneğin kaç adet bileşene sahip olduğu da BBK yöntemiyle tahmin edildi. BBK ve BSO değerlerinin bulunması için $K = \{2, 3, 4, \cdots, 14, 15\}$ değerleri araştırılmıştır. Bir verinin kümelere ayrılabilmesi için en az 2 kümeye sahip olması gerektiğinden alt sınır 2 olarak seçildi. Yapılan deneylerde üst sınırın 15'ten büyük alınması bu veri kümesi için başarı oranını önemsiz sayılabilecek oranda etkilediği görülmüştür. Ayrıca daha büyük değerler için işlem yapılması çalışma zamanını da artırmaktadır. Bu sebeplerden dolayı için üst sınır 15 seçilmiştir. Veri kümesindeki her örnek için verilen K kümesindeki her değer için 100 kez yapılan deneme yapılmıştır. Karışım modelleri verileri rastgele bir yerden başlayarak kümelediği için yaklaşık olarak ortalama değerleri elde etmek için 100 kez tekrar edilmiştir.

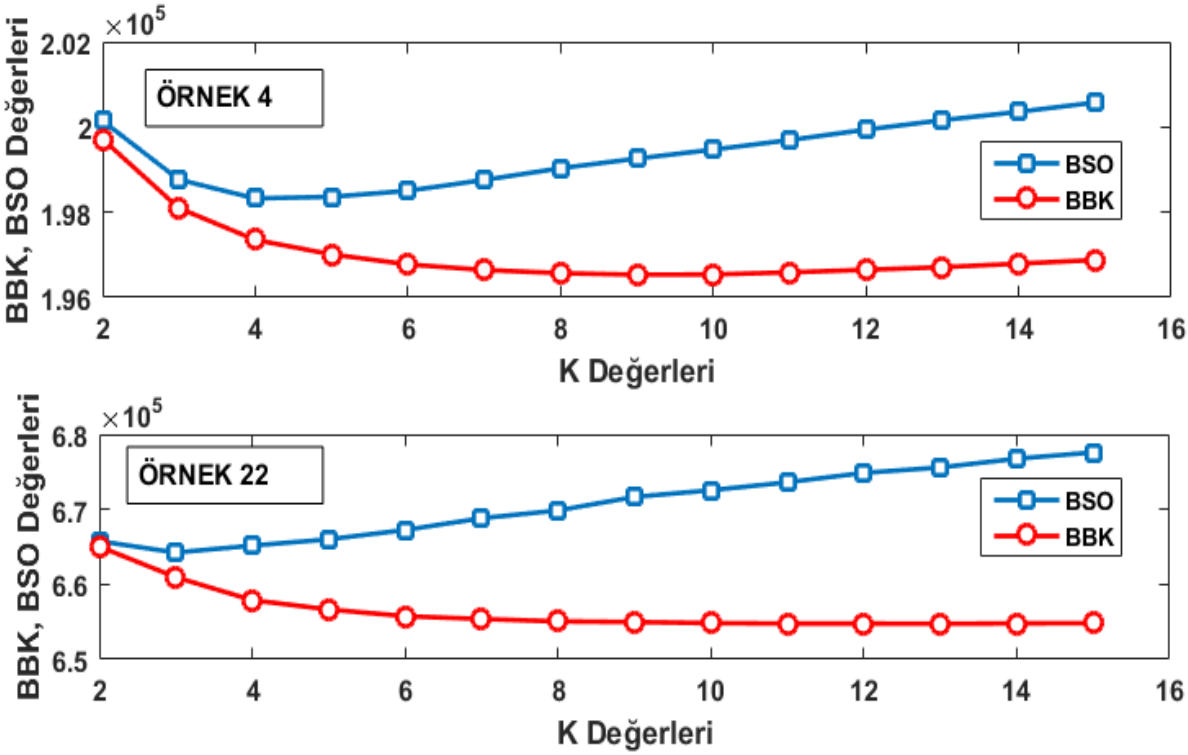
Deneyler sonucunda bulunan ortalama bileşen ve küme sayıları tablo 4.1'de verilmiştir. Ortalama sonuçlara göre başlangıçtaki bileşen sayısını veren BBK'nin çoğunlukla 12-15 arası değerler aldığı görülmektedir. Bu sonuç örneklerin fazla sayıda küçük kümeye sahip olduğu şeklinde yorumlanabilir. BSO kriteri örneklerdeki ortalama küme sayılarını genellikle 2-5 arası olarak tahmin etmiştir. Tablo 4.1'de kırmızı işaretli olarak gösterilen 9 örnekteki küme

sayılarını BSO kriteri doğru tahmin etmiştir. Mavi renkle gösterilen 14 örnekte BSO kriteri küme sayısını olması gerekenden az sayıda tahmin etmiştir. Bu durum BSO kriterinin küçük sayıda hücre içeren kümeleri es geçebileceğini şeklinde yorumlanabilir. Siyah renkle gösterilen 7 örnekte ise BSO kriteri olması gerekenden daha fazla sayıda küme tahmini yapmıştır. 30 örnekten 21'inde küme sayısı tam olarak tahmin edilemese bile iki örnek *(14 ve 24)* hariç küme sayıları birer eksik veya fazla tahmin edilmiştir. Bölüm başında da belirtildiği gibi verilerdeki küme sayısını bulmak hala tam olarak çözülememiş bir problemdir. Bu durum göz önüne alındığında BSO kriterinin DLBCL verilerinin küme değerlerini başarılı şekilde tahmin ettiği söylenebilir.

Tablo 4.1: DLBCL veri kümesi için BBK, BSO ve gerçek küme sayıları

Örnek	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
BBK	8	8	8	9	6	9	8	12	7	7	7	13	12	15	15
BSO	4	2	2	4	3	3	3	3	3	2	3	5	3	7	3
Gerçek Küme Sayısı	3	3	3	4	4	3	4	3	3	3	4	5	5	4	4
Örnek	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
BBK	12	15	14	9	15	13	13	14	15	15	15	15	12	14	14
BSO	3	5	3	2	3	4	3	6	7	5	5	5	2	4	5
Gerçek Küme Sayısı	4	5	4	4	4	4	4	5	5	4	4	4	3	4	5

Şekil 4.1, 4. ve 22. örnek için ortalama BBK ve BSO değerlerini göstermektedir. 4. örnek için çizilen üstteki grafik incelendiğinde BBK kriterinin k=8'den itibaren neredeyse sıfır eğimle devam ettiği görülmektedir. Fakat küçük farkla k=9'da BBK kriteri minimum değer üretmiştir. Buna göre 4. örnek için başlangıç değeri/bileşen sayısı 9 olarak seçilmiştir. 22. örnek için BBK k=8'den itibaren yatay olarak devam etmiştir ve k=13'te minimum olmuştur. Yine Şekil 4.1'te verilen BSO değerleri incelendiğinde 4. örnek için kriter k=4'te 22. örnek için k=3'te minimum olmuştur. Buna göre 4. örnek için küme sayısı tam olarak bulunmuşken 22. örnek için gerçek küme sayısından bir eksik bulunmuştur.



Şekil 4.1: Örnek 4 ve 22 için BBK ve BSO grafiği

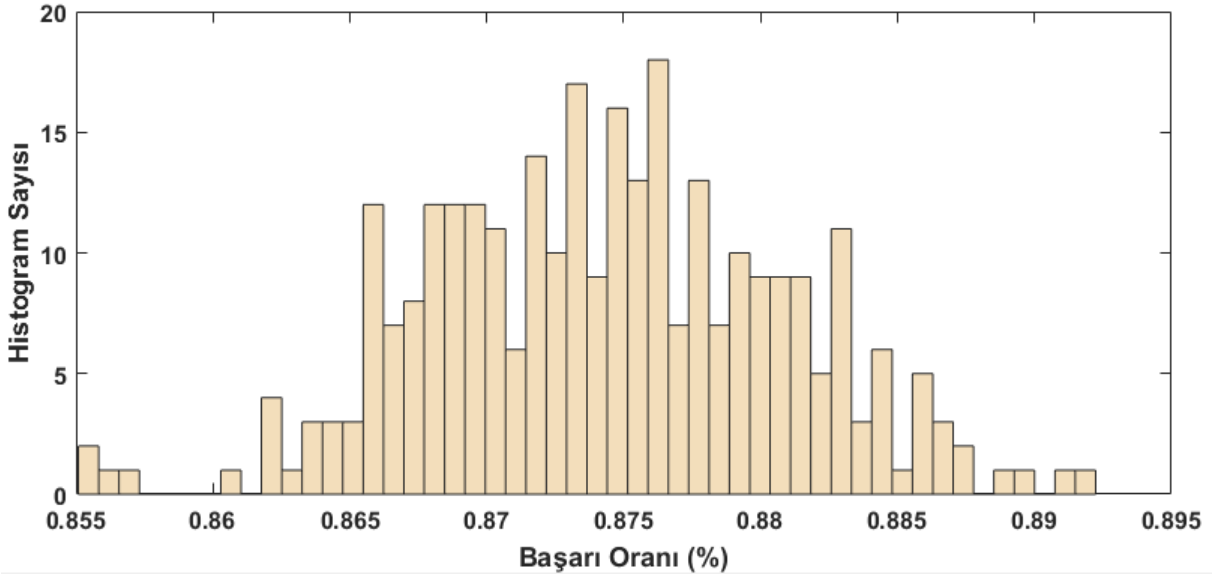
Belirlenen başlangıç (bileşen sayısı) ve bitiş (küme sayısı) değerleri ile DLBCL veri kümesine ait 30 örnek 300 kez bölüm 3.7'de anlatılan yöntemle kapılanmıştır. K-ortalamalar algoritması kümelemeye rastgele bir pozisyondan başladığı için ortalama başarı oranını yakalamak için 300 tekrar yapılmıştır. Her tekrarda 30 örneğin başarı oranının ortalaması alınmıştır.

Tablo 4.2 elde edilen sonuçlara göre en yüksek, en düşük ve ortalama başarı oranını göstermektedir. Buna örnekler ortalama **%87.44** oranında başarıyla kümelenmiştir. En yüksek ve en düşük başarı oranı arasında yaklaşık %4 fark olması kümeleme yönteminin belirli bir seviyede kararlı olarak çalıştığını göstermektedir. Tablo 4.3'te yapılan 300 deneyde elde edilen başarı oranlarının histogram grafiği görülmektedir. Buna göre deney sonuçlarının büyük oranda %87 etrafında toplandığı görülür.

Tablo 4.2: DLBCL veri kümesi için parametre girişsiz 300 deneyin ortalama başarı oranı

Deney Sayısı	**En Yüksek Başarı Oranı**	**En düşük Başarı Oranı**	**Ortalama Başarı Oranı**
300	%89.22	%85.51	%87.44

Tablo 4.3: DLBCL veri kümesi için parametre girişsiz yapılan 300 deneyin histogram grafiği



Tablo 4.4: DLBCL veri kümesi parametre girişsiz kapılama süreleri

Her örnek için ortalama kapılama süresi	~ 10,1 saniye
30 örnek için ortalama kapılama süresi	~ 305,17saniye

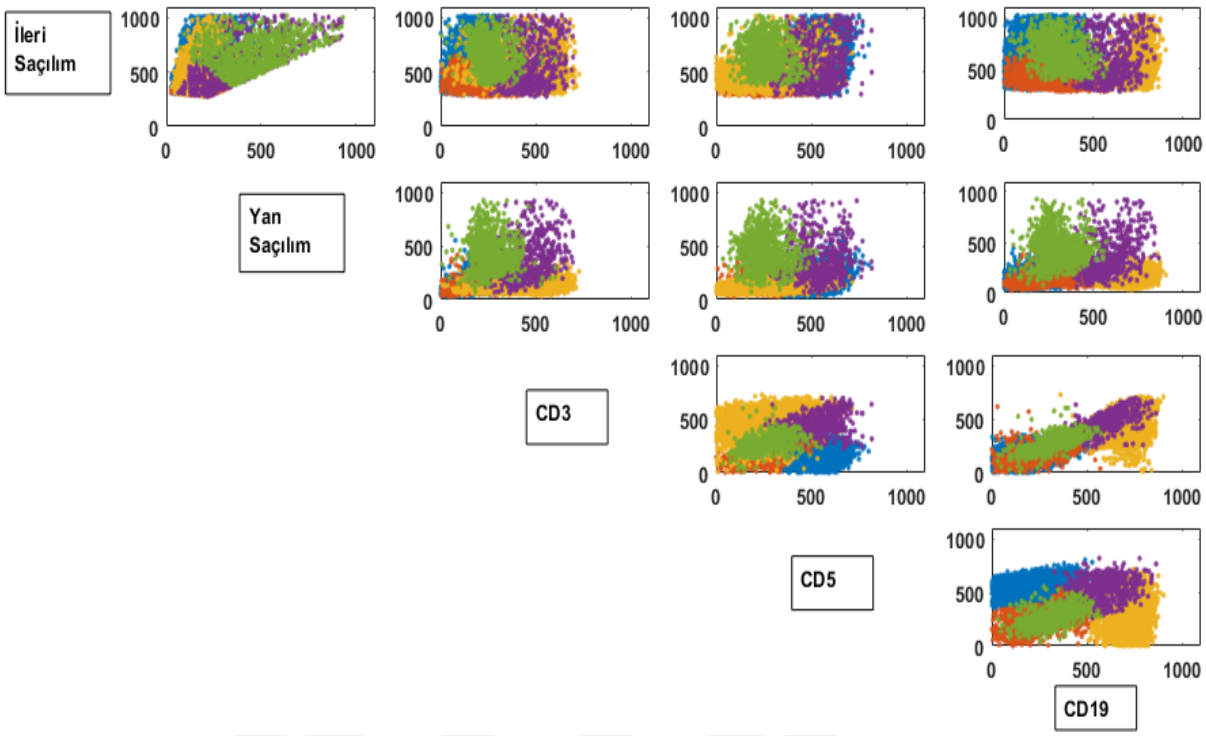
Tablo 4.4 örneklerin kapılama sürelerini göstermektedir. Yapılan 300 denemeden elde edilen ölçüm sonuçlarına göre DLBCL veri kümesinde bulunan 30 örneği bir kez kapılamak yaklaşık olarak 305,17 saniye sürmektedir. Buna göre her örneği kapılamak yaklaşık olarak 10,1 saniye sürmektedir. manuel kapılama süreleri göz önüne alındığında otomatik kapılamanın oldukça hızlı olduğu görülmektedir.

Tablo 4.5'te her örnek için elde edilen en yüksek, en düşük ve ortalama başarı oranları verilmiştir. Sonuçlara göre 17 örnek ortalamanın üzerinde başarı ile kapılanmıştır. %80'in

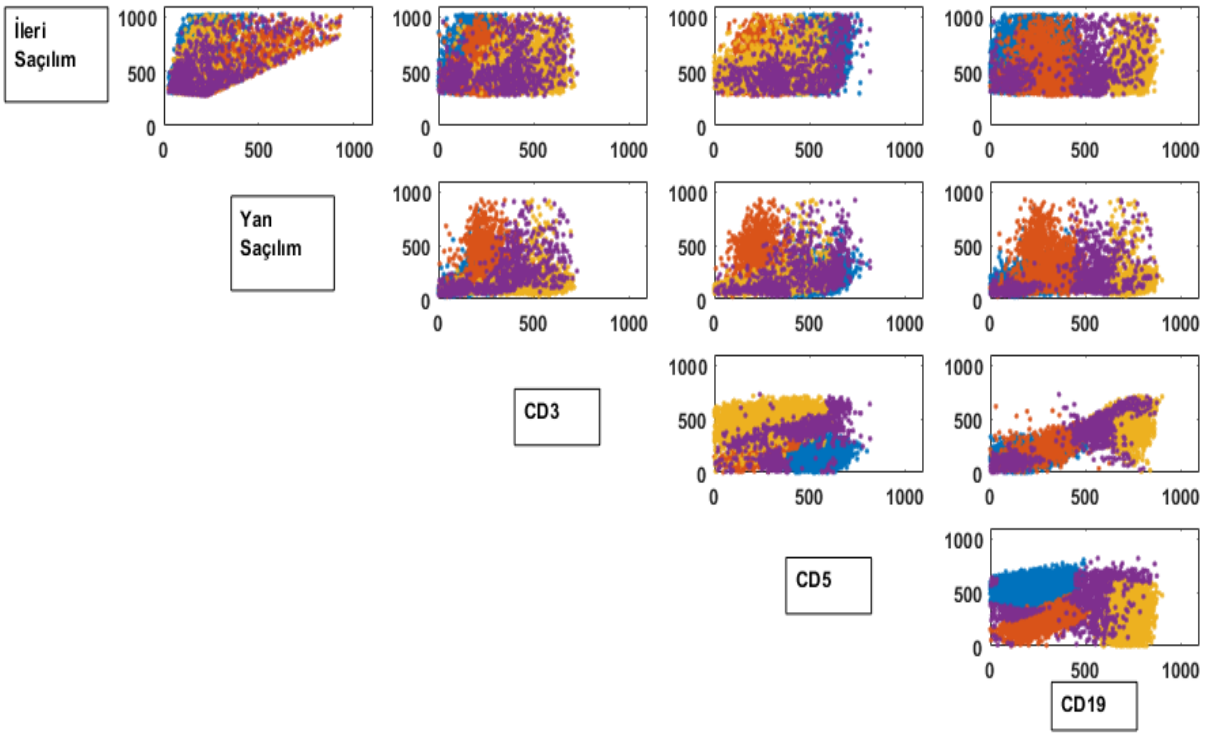
altında başarıyla kapılanan 6 örnek vardır. 2. örnek en yüksek başarıyla kapılanan, 17. örnek en düşük başarıyla kapılanan örneklerdir. Yüksek başarı oranına rağmen 2. örnek için tahmin edilen küme sayısı gerçek küme sayısından bir eksiktir. 17.örnekte tahmin edilen küme sayısı gerçek küme sayısıyla eşittir. Buna rağmen başarı oranının düşük olmasının sebebi verilerin anlamlı dağılım göstermediği olarak görülebilir. Sonuç olarak küme sayısının doğru tahmini kadar verileri doğru olarak da ifade etmek oldukça önemlidir.

Tablo 4.5: DLBCL veri kümesi örnekleri için parametre girişsiz 300 deneyin kapılama sonuçları

Örnek / ***Başarı Oranı***	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
En Yüksek(%)	90.7	98.64	98.41	78.7	92.44	98.84	84.5	92.6	95.31	92.55
En Düşük(%)	81.31	98.37	88.94	71.64	84.94	77.16	72.96	68.44	91.95	89.35
Ortalama(%)	86.20	98.44	95.08	74.2	90.88	96.84	78.1	86.13	93.28	92.42
Örnek / ***Başarı Oranı***	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
En Yüksek(%)	79.9	91.21	72.63	85.21	91.98	97.87	76.21	91.2	94.25	95.57
En Düşük(%)	73.71	79.64	64.23	60.25	79.49	85.1	60.41	70.25	84.29	84.31
Ortalama(%)	78.17	87.56	70.35	76.88	88.91	96.62	67.12	87.05	94.1	94.59
Örnek / ***Başarı Oranı***	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
En Yüksek(%)	94.54	97.99	93.19	84.49	93.35	92.31	90.37	98.92	95.96	96.12
En Düşük(%)	72.46	89.56	69.96	71.89	74.49	79.72	76.92	95.93	77.1	72.27
Ortalama(%)	86.69	96.13	84.18	80.19	88.27	90.5	88.63	98.84	91.2	85.34



Şekil 4.2: Örnek 25 için önerilen kümeleme yöntemiyle bulunan ortalama kapılama (küme) sonuçları



Şekil 4.3: Örnek 25 için manuel kapılama (küme) sonuçları

Şekil 4.2 ve 4.3 örnek 25 için sırasıyla yöntemin ve manuel kapılama sonuçlarını göstermektedir. şekillerde veriler tüm özelliklerin saçılım grafikleri şeklinde verilmiştir. Her özellik düşeyde ve yatayda denk geldiği grafiğin yine denk geldiği eksenini ifade etmektedir. Örnek 25 için tahmin edilen küme sayısı 5 iken gerçek küme sayısı 4'tür. Örnekte aynı kümeyi ifade eden hücreler aynı renkle gösterilmiştir. Buna göre yöntem sonucu elde edilen kapılardan yeşil renkle gösterilen hücre grubu turuncu ve mor renkle gösterilen hücre grubundan elde edildiği görülmektedir. İki küme ifade edilen hücre gruplarının üç gruba bölünmesi başarı oranını düşürmüştür. Bununla beraber sarı ve mavi ile gösterilen hücre grupları başarı ile kapılanmıştır.

4.2. PARAMETRE GİRİŞİ OLAN KAPILAMA

Parametre girişi olmayan kapılama deneyinde örneklerin küme sayıları manuel kapılama sonucu oluşturulan kapı sayıları olarak alınmıştır. Başka bir deyişle BSO'nun bulduğu değer yerine gerçek küme sayıları bitiş değeri olarak kullanılmıştır. Deney Tablo 4.1'de verilen BBK kriterinin tahmin ettiği bileşen sayısıyla kapılamaya başlayıp manuel kapılamada bulunan doğru küme sayısı ile bitecek şekilde yapılmıştır. Bu deneyin amacı doğru küme sayılarıyla yöntem denendiğinde elde edilen başarı oranının ilk deneyde bulunan başarı oranıyla karşılaştırmaktır.

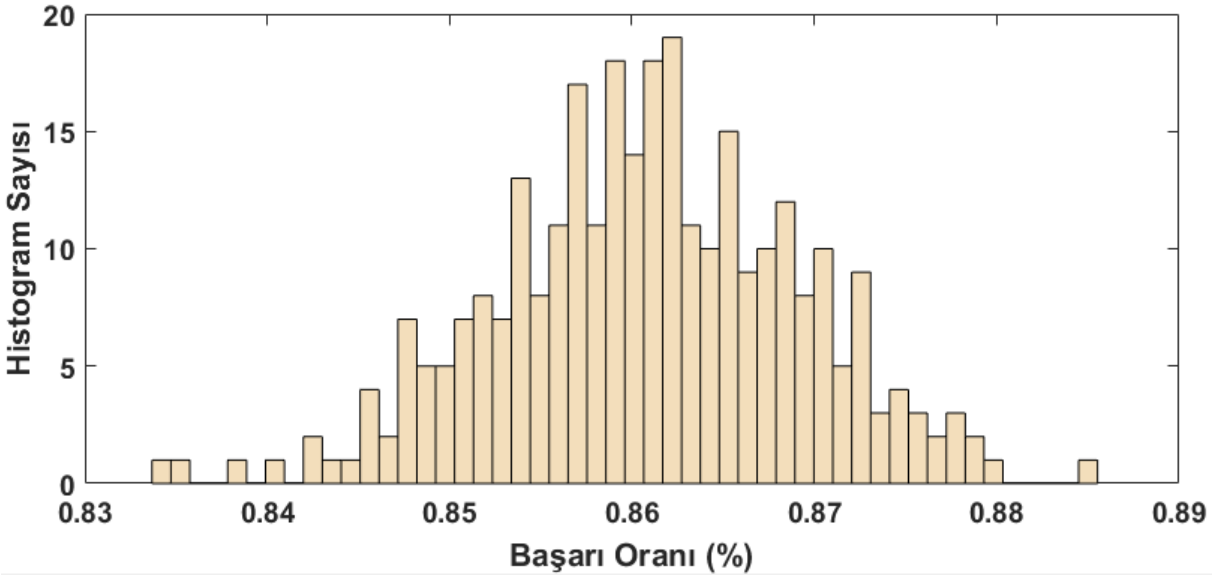
300 kez tekrarlanan deney sonucunda DLBCL için elde edilen kapılama başarı oranları Tablo 4.6'da verilmiştir. Buna göre ortalama başarı oranı **%86.06** olarak bulunmuştur. Doğru küme sayıları kullanılmasına rağmen bu sonuç ilk deneyde bulunan başarı oranından daha düşüktür. İlk deneyde kullanılan BSO kriteri 14 örnekte küme sayısını gerçek küme sayısından daha az tahmin etmişti. Böyle durumlarda genellikle az sayıda hücre içeren bir küme fazla sayıda hücre içeren bir kümeyle birleştirilir. Yöntem bu az sayıdaki kümeyi yanlış BSO kriteri yüzünden ayırt edemez ve F-skor hesaplanırken fazla sayıda hücre içeren grubun toplam doğruluk oranına etkisi fazla olduğundan başarı oranı daha çok olarak bulunur. Bölümün en başında da belirtildiği gibi verilerdeki küme sayısını tahmin etme çok önemli bir problemdir. BSO kriteri dağılım karışımlar için küme sayısını tahmin etme de kullanılsa da daha iyi yöntemlerin veya çözümlerin geliştirilmesine ihtiyaç vardır. Bu üzerinde çalışılması gereken bir araştırma alanıdır.

Tablo 4.6: DLBCL veri kümesi için parametre girişli 300 deneyin ortalama başarı oranı

Deney Sayısı	**En Yüksek Başarı Oranı**	**En düşük Başarı Oranı**	**Ortalama Başarı Oranı**
300	%88.55	%83.37	%86.06

Tablo 4.7'de yapılan 300 deneyin başarı oranlarının histogram grafiği gözükmektedir. şekilden de görüleceği üzere başarı oranı %86 etrafında yoğunlaşmıştır. Bu grafikteki sonuçların tablo 4.3'te verilen sonuçlara göre daha fazla normal dağılıma benzer olduğu görülmektedir. Bu yöntemin doğru küme sayılarıyla daha tutarlı çalıştığı şeklinde yorumlanabilir.

Tablo 4.7: DLBCL veri kümesi için parametre girişli yapılan 300 deneyin histogram grafiği



Tablo 4.8: DLBCL veri kümesi parametre girişli kapılama süreleri

Her örnek için ortalama kapılama süresi	~ 2,4 saniye
30 örnek için ortalama kapılama süresi	~ 72,18 saniye

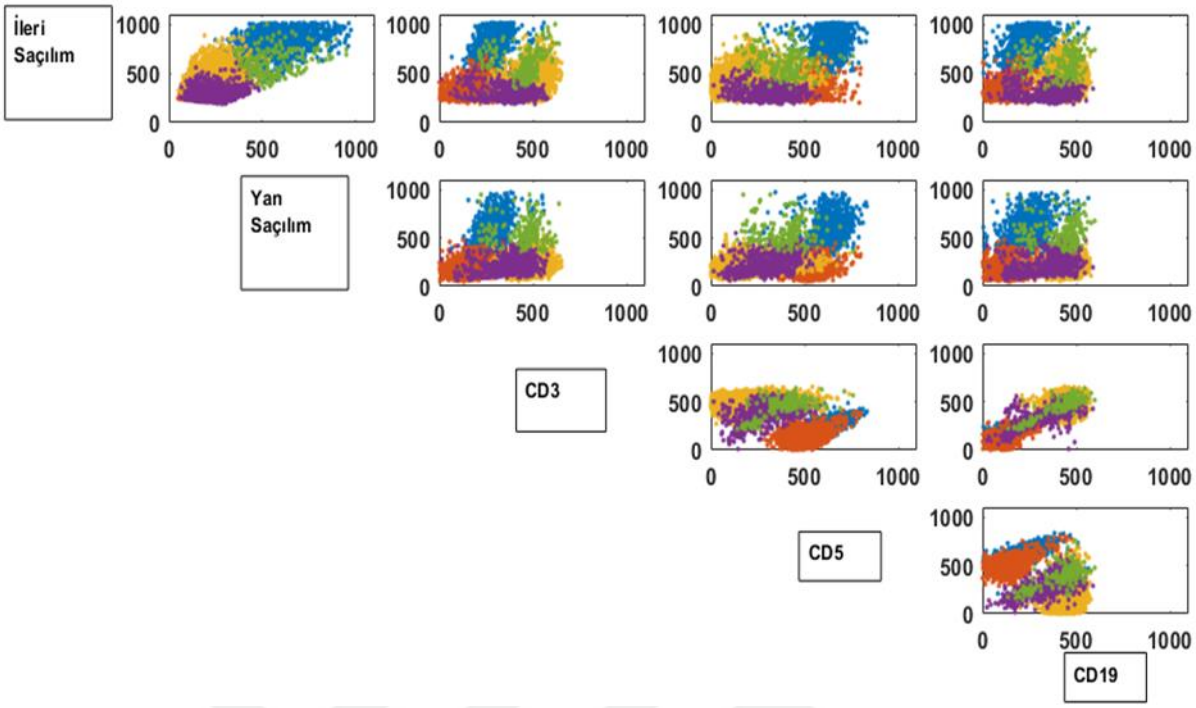
Tablo 4.8'de kapılama süreleri verilmiştir. Elde edilen sonuçlara göre her örneğin ortalama kapılanma süresi 2,4 saniye, veri kümesindeki tüm örneklerin kapılanma süresi ise 72,18 saniye sürmektedir. Sonuçlar Tablo 4.4'le karşılaştırıldığında başlangıç ve bitiş değerlerinin yaklaşık

olarak 7 saniyede belirlendiği görülür. Bunun sebebi K kümesindeki her değer için verilerin dağılıma uydurulup kriterlerin hesaplanmasının yapılmasıdır.

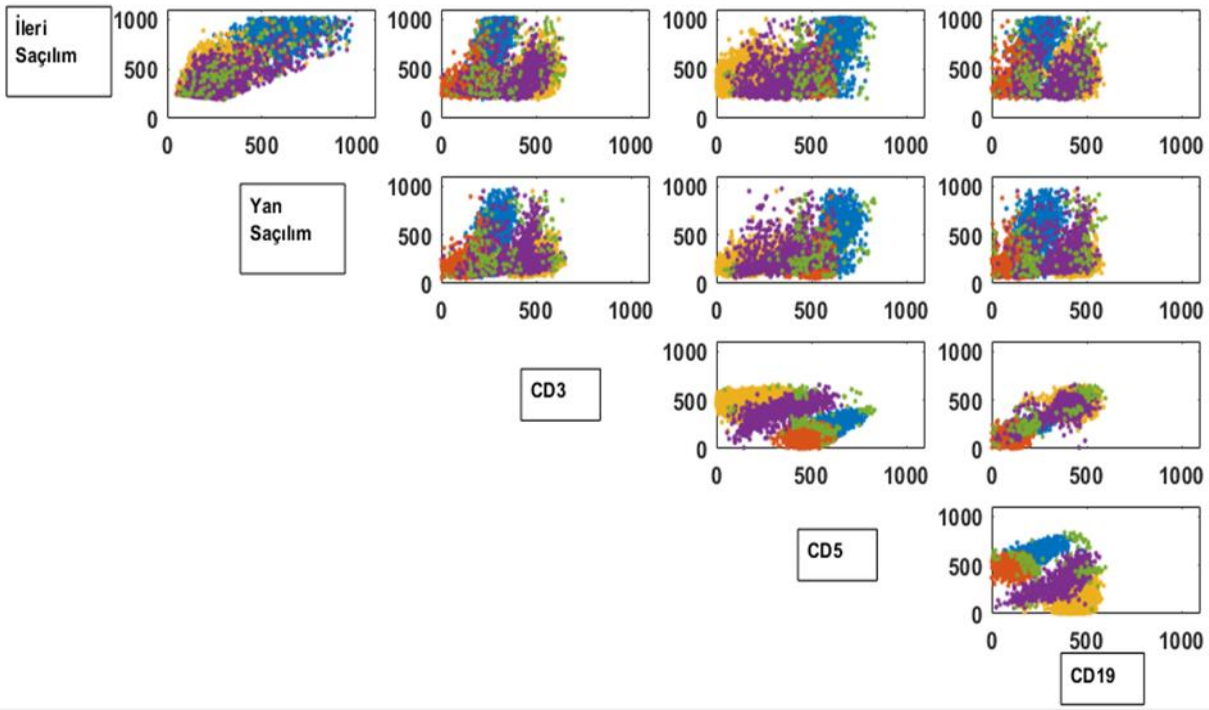
Tablo 4.9'da her örneğe ait en yüksek, en düşük ve ortalama kapılama başarı oranları verilmiştir. Sonuçlara göre 18 örnek ortalamanın üzerinde başarıyla kapılanmıştır. 13. ve 17. örnek %70'in altında başarı oranıyla en düşük doğruluğun elde edildiği örnekler olmuşlardır. Buna karşın 6, 16 ve 22. örnekler %95'in üzerinde başarıyla kapılanmıştır.

Tablo 4.9: DLBCL veri kümesi örnekleri için parametre girişli 300 deneyin kapılama sonuçları

Örnek / ***Başarı Oranı***	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
En Yüksek(%)	96.31	98.43	98.86	78.70	92.88	99.03	85.14	92.59	95.11	93.16
En Düşük(%)	88.61	70.83	81.95	71.35	70.13	75.24	66.37	68.39	90.46	88.85
Ortalama(%)	90.86	79.58	93.02	74.22	84.06	96.86	77.55	85.94	93.28	91.87
Örnek / ***Başarı Oranı***	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
En Yüksek(%)	86.35	91.18	79.35	95.18	92.69	97.06	76.43	92.02	96.86	94.41
En Düşük(%)	63.56	80.16	44.14	79.22	68.23	94.36	52.26	63.53	58.95	81.71
Ortalama(%)	80.84	87.43	64.12	88.69	82.29	96.62	66.77	86.5	78.51	92.82
Örnek / ***Başarı Oranı***	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
En Yüksek(%)	94.56	97.53	94.6	85.02	94.35	96.68	89.67	99.53	95.91	96.15
En Düşük(%)	72.43	76.85	75.79	64.47	85.45	88.32	86.22	68.05	77.14	72.21
Ortalama(%)	87.66	95.35	89.5	75.32	92.32	90.77	88.85	94.05	91.44	85.45



Şekil 4.4: Örnek 12 için önerilen kümeleme yöntemiyle bulunan ortalama kapılama (küme) sonuçları



Şekil 4.5: Örnek 12 için manuel kapılama (küme) sonuçları

Şekil 4.4 ve Şekil 4.5 örnek 12'ye ait uygulanan yöntemin bulduğu kapılama sonuçlarıyla manuel yapılan kapılama sonuçlarını göstermektedirler. Örnek toplamda 5 küme içermektedir. Turuncu mavi ve sarı renkle gösterilen hücre grupları başarıyla bulunmuştur. Buna karşın uygulanan kapılama yöntemi yeşil renkle gösterilen hücre kümesini daha çok mor renkle gösterilen hücre kümesinin büyük değerli grubu olarak bulmuştur. Gerçek kapılama sonuçlarına göre bu durum kısmen doğrudur. Fakat yeşil renkli hücre grubu belirli bir topluluk oluşturmamaktadır. Böyle bir sonucun ortaya çıkması uygulanan kümeleme yönteminin belirli bir dağılımla ifade edilmekte zorlanılan hücre kümelerini tam olarak bulamadığını göstermektedir.

5. TARTIŞMA VE SONUÇ

DLBCL Akış sitometrisi veri kümesi üzerinde bölüm 3'te açıklanan yöntemle yapılan kapılama sonuçları bölüm 4'te paylaşılmıştır. Sonuçlara göre örneklerin içerdiği küme sayılarının BSO kriteriyle tahmin edildiği deneyde ortalama kapılama/kümeleme başarısı %87,44 olarak bulunmuştur. Örneklerin içerdiği küme sayılarının giriş parametresi olarak verilen ikinci deneyde ortalama kapılama/kümeleme başarısı %86,06 çıkmıştır.

Bu tez çalışmasında, AS verilerinin kapılanması olarak adlandırılan, verilerin doğru şekilde kümelenmesi üzerinde çalışılmıştır. AS verileri çok boyutlu verilerdir ve bunların doğru şekilde kapılanması uzmanlar tarafından çoğunlukla bir ve iki boyutlu nadiren üç boyutlu saçılım grafikleri üzerinden manuel yapılmaktadır. Bu şekilde kapılama yapmak, daha maliyetlidir ve daha zaman alıcı olmaktadır. Ayrıca, bu kapılama işleminde, uzman kişinin bilgi seviyesine subjektif olarak bağımlılık söz konusudur. Bununla beraber, manuel kapılamada çok boyutlu verilerin aynı anda tüm boyutlarıyla gözlemlenmesi de mümkün değildir.

Bu gibi sebepler göz önüne alındığında, kümeleme çalışmaları bilim insanlarını en çok yoran ve zaman alan analiz adımı olarak karşımıza çıkmaktadır. Dolayısıyla, AS verilerini otomatik olarak kapılayan yöntemlerin geliştirilmesine ihtiyaç duyulmuştur. Bu amaçla yapılan çalışmalar, daha çok kümeleme yöntemleri üzerine yoğunlaşmıştır. Kümeleme yöntemlerinde, verileri bir olasılıksal dağılım modeline uydurarak ifade etme, çalışmalarda sıklıkla tercih edilen fikir ve yöntemler olmuştur. Literatürde, T-dağılımı ve normal dağılım gibi farklı dağılım yöntemleri, verileri modellemek için kullanılan metotlardan bir kısmıdır. Fakat verileri tek bir dağılımla ifade etmek çoğunlukla mümkün değildir. Bu sebeple, birden çok dağılımın katsayılarla toplamı şeklinde tanımlanan, karışım modelleri kullanılmıştır. Karışım modelleri, AS verilerini ifade etmede büyük esneklik ve başarı göstererek, verileri yüksek başarıyla kümelemektedirler.

Bununla beraber, AS verilerinin kapılanmasında hala geliştirilmesi gereken kısımlar mevcuttur. Kan içerisinde kaç çeşit hücre olduğu bilinse de verilerin kaç adet kümeye sahip olduğunun tahmin edilmesi, literatürde tam olarak çözülememiş önemli bir problemdir. Küme sayılarının az veya fazla tahmin edilmesi ileriki analizleri etkilemektedir. Diğer bir problem ise kapılama yöntemlerinin küçük sayıda hücre içeren veri kümelerini tespit etmesinde sorunlar yaşamasıdır. Bu gibi durumlarda, küçük kümeler büyük kümelerin bir parçası olarak yorumlanmaktadır. Bu

kapsamda, yanlış sonuçlar elde edilebilir. Kapılama sonrası bulunan hücre kümelerinin, diğer örneklerde yer alan aynı tür hücre kümeleriyle eşleştirilmesi de üzerinde çalışılması gereken bir başka araştırma alanıdır.

Sonuç olarak, AS cihazının ürettiği verilerin daha doğru ve daha hızlı analiz edilebilmesi için yöntemler geliştirilmesine ihtiyaç duyulmaktadır. AS veri analizi düzeltme, dönüşüm, kapılama, küme eşleştirme gibi farklı aşamalara sahiptir ve her aşama için yöntemler geliştirilmeye çalışılmaktadır. Bu tez çalışmasında da temelde Gauss Karışım Modeli (GKM) kullanılarak, AS verilerini kapılayan bir yöntem önerilmiştir. Bu yöntem, DLBCL veri kümesi üzerinde test edilmiş ve elde edilen sonuçlar bu tez çalışması içerisinde paylaşılmıştır. Sonuçlara göre, önerilen yöntem verileri yüksek başarıyla kapılama yapabilmektedir. Önerilen bu yöntem, verileri fazla sayıda küçük kümeye ayırarak, kapılama işlemini yapmaya başlamaktadır. Sonrasında, birbirlerine en yakın kümeler birleştirilir. Birleştirilen bu kümeler, içerdikleri toplam bileşen sayısına göre Gauss Karışım Modeliyle ifade edilirler. İki küme arası uzaklık ölçümünde, her bileşen kendi ağırlığı oranında sonuca etki eder. Sonuç itibariyle, elde edilen kümeler farklı sayıda bileşene ve şekil formatına sahip olmaktadır. Önerilen yöntem, tüm boyutlarda topluluk halinde bulunan veriler için kümeleme performansı yönünden, var olan bazı diğer yaklaşımlara göre daha başarılı şekilde çalışmaktadır. Ayrıca bu yöntemin, İstanbul Üniversitesi, İstanbul Tıp Fakültesi, Temel Tıp Bilimleri Bölümü araştırmalarında kullanılmaya başlanması planlanmaktadır.

Veri kümelerinde, belirli bir şekil formatına sokulamayan verilerin bulunmasında tezde kullanılan yöntem itibariyle, yeni geliştirmeler yapılması gerekmektedir. Bu durum, ileride yapılması muhtemel tez, yayın ve deneysel çalışmalarda üzerinde çalışılabilecek bir alan olarak karşımıza çıkmaktadır. Bu tez çalışmasında kullanılan DLBCL veri kümesi 5 boyutlu olup, ileriki çalışmalarda daha fazla boyutlu AS veri kümeleri üzerinde kapılama yapılması planlanmaktadır.

KAYNAKLAR

[1] U. Güner, "Flow Sitometrinin Hidrobiyolojide Kullanımı," *J. Fish. Sci. com*, vol. 6, no. 1, pp. 9–17, 2012.

[2] T. S. Hawley ve R. G. Hawley, "Flow Cytometry Protocols," *Springer Sci. Bus. Media*, vol. 263, 2004.

[3] F. Taneli, "Flow Sitometri Tekniği ve Klinik Laboratuvarlarda Kullanımı," *Türk Klin. Biyokim. Derg*, vol. 5, no. 2, pp. 75–82, 2007.

[4] Star CellBio, "Flow Cytometry," 2015. [Online]. Available: http://star.mit.edu/CellBio/animations/index.html. [Erişim Tarihi: 17.10.2018].

[5] R. Misha, *Introduction to Flow Cytometry*. AbD serotec, BIO-RAD, 2014.

[6] İ. Karaboz, E. Kayar, ve S. Akar, "Flow Sitometri ve Kullanım Alanları," *Elektron. Mikrobiyoloji Derg.*, vol. 6, no. 2, pp. 1–18, 2008.

[7] K. O'Neill, N. Aghaeepour, J. Špidlen, ve R. Brinkman, "Flow Cytometry Bioinformatics," *PLoS Comput. Biol.*, vol. 9, no. 12, 2013.

[8] R. R. Jahan-Tigh, C. Ryan, G. Obermoser, ve K. Schwarzenberger, "Flow cytometry," *J. Invest. Dermatol.*, vol. 132, no. 10, pp. 1–6, 2012.

[9] A. Adan, G. Alizada, Y. Kiraz, Y. Baran, ve A. Nalbant, "Flow cytometry: basic principles and applications," *Crit. Rev. Biotechnol.*, vol. 37, no. 2, pp. 163–176, 2017.

[10] M. Roederer, "Compensation in Flow Cytometry," *Curr Protoc Cytom Chapter 1*, p. Unit 1.14, 2002.

[11] A. Azad, "An Algorithmic Pipeline For Analyzing Multi-parametric Flow Cytometry Data," Purdue University, 2014.

[12] BIO-RAD, *Flow Cytometry Basics Guide*. 2016.

[13] K. Dalva, "Hematoloji ' de Akım Sitometri Kullanımı," *Temel Moleküler Hematol. Kursu*, pp. 73–86, 2005.

[14] D. N. Software, "FCS Express." [Online]. Available: https://www.denovosoftware.com/. [Erişim Tarihi: 17.10.2018].

[15] FlowingSoftware, "Flowing Software." [Online]. Available: http://flowingsoftware.btk.fi/index.php?page=1. [Erişim Tarihi: 17.10.2018].

[16] A. Bashashati ve R. R. Brinkman, "A Survey of Flow Cytometry Data Analysis Methods," *Adv. Bioinformatics*, vol. 2009, pp. 1–19, 2009.

[17] M. J. Boedigheimer ve J. Ferbas, "Mixture modeling approach to flow cytometry data," *Cytom. Part A*, vol. 73, no. 5, pp. 421–429, 2008.

[18] N. Aghaeepour *ve diğ.*, "Critical assessment of automated flow cytometry data analysis techniques," *Nat. Methods*, vol. 10, no. 3, pp. 228–238, 2013.

[19] Y. Saeys, S. Van Gassen, ve B. N. Lambrecht, "Computational flow cytometry: Helping to make sense of high-dimensional immunology data," *Nat. Rev. Immunol.*, vol. 16, no. 7, pp. 449–462, 2016.

[20] R. F. Murphy, "Automated identification of subpopulations in flow cytometric list mode data using cluster analysis," *Cytometry*, vol. 6, no. 4, pp. 302–309, 1985.

[21] N. Aghaeepour, R. Nikolic, H. H. Hoos, ve R. R. Brinkman, "Rapid cell population identification in flow cytometry data," *Cytom. Part A*, vol. 79 A, no. 1, pp. 6–13, 2011.

[22] C. Chan, F. Feng, J. Ottinger, D. Foster, M. West, ve T. B. Kepler, "Statistical mixture modeling for cell subtype identification in flow cytometry," *Cytom. Part A*, vol. 73A, no. 8, pp. 693–701, 2008.

[23] K. Lo, R. R. Brinkman, ve R. Gottardo, "Automated gating of flow cytometry data via robust model-based clustering," *Cytom. Part A*, vol. 73, no. 4, pp. 321–332, 2008.

[24] G. Finak, A. Bashashati, R. Brinkman, ve R. Gottardo, "Merging Mixture Components for Cell Population Identification in Flow Cytometry," *Adv. Bioinformatics*, vol. 2009, pp. 1–12, 2009.

[25] Y. Ge ve S. C. Sealfon, "Flowpeaks: A fast unsupervised clustering for flow cytometry data via K-means and density peak finding," *Bioinformatics*, vol. 28, no. 15, pp. 2052–2058, 2012.

[26] G. Walther *ve diğ.*, "Automatic Clustering of Flow Cytometry Data with Density-Based Merging," *Adv. Bioinformatics*, vol. 2009, pp. 1–7, 2009.

[27] Y. Qian *ve diğ.*, "Elucidation of seventeen human peripheral blood B-cell subsets and quantification of the tetanus response using a density-based method for the automated identification of cell populations in multidimensional flow cytometry data," *Cytom. Part B - Clin. Cytom.*, vol. 78, no. SUPPL. 1, pp. 69–82, 2010.

[28] A. Cron *ve diğ.*, "Hierarchical Modeling for Rare Event Detection and Cell Subset Alignment across Flow Cytometry Samples," *PLoS Comput. Biol.*, vol. 9, no. 7, 2013.

[29] B. E. Köktürk ve B. Karaçali, "Model-free expectation maximization for divisive hierarchical clustering of multicolor flow cytometry data," *Proc. - 2014 IEEE Int. Conf. Bioinforma. Biomed. IEEE BIBM 2014*, pp. 267–272, 2014.

[30] M. Dundar, F. Akova, H. Z. Yerebakan, ve B. Rajwa, "A non-parametric Bayesian model for joint cell clustering and cluster matching: Identification of anomalous sample phenotypes with random effects," *BMC Bioinformatics*, vol. 15, no. 1, pp. 1–15, 2014.

[31] K. Johnsson, J. Wallin, ve M. Fontes, "BayesFlow: Latent modeling of flow cytometry cell populations," *BMC Bioinformatics*, vol. 17, no. 1, pp. 1–16, 2016.

[32] M. Lux *ve diğ.*, "flowLearn: Fast and precise identification and quality checking of cell populations in flow cytometry," *Bioinformatics*, vol. 34, no. February, pp. 2245–2253, 2018.

[33] L. M. Weber ve M. D. Robinson, "Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data," *Cytom. Part A*, vol. 89, no. 12, pp. 1084–1096, 2016.

[34] A. K. Jain, M. N. Murty, ve P. J. Flynn, "Data clustering: a review," *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264–323, 1999.

[35] D. Everitt, B. S., Landau, S., Leese, M., Stahl, *Cluster Analysis*. 2001.

[36] Wikipedia, "k-means clustering." [Online]. Available: https://en.wikipedia.org/wiki/K-means_clustering#cite_note-lloyd1957-4. [Erişim Tarihi: 17.10.2018].

[37] C. M. Bishop, *Pattern recognition and machine learning*. 2006.

[38] S. M. Ross, *Introduction To Probability and Statistics for Engineers and Scientists*. Elsevier Inc., 2004.

[39] S. K. Zhou ve R. Chellappa, "From sample similarity to ensemble similarity probabilistic distance measures in reproducing kernel Hilbert space," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 6, pp. 1–34, 2006.

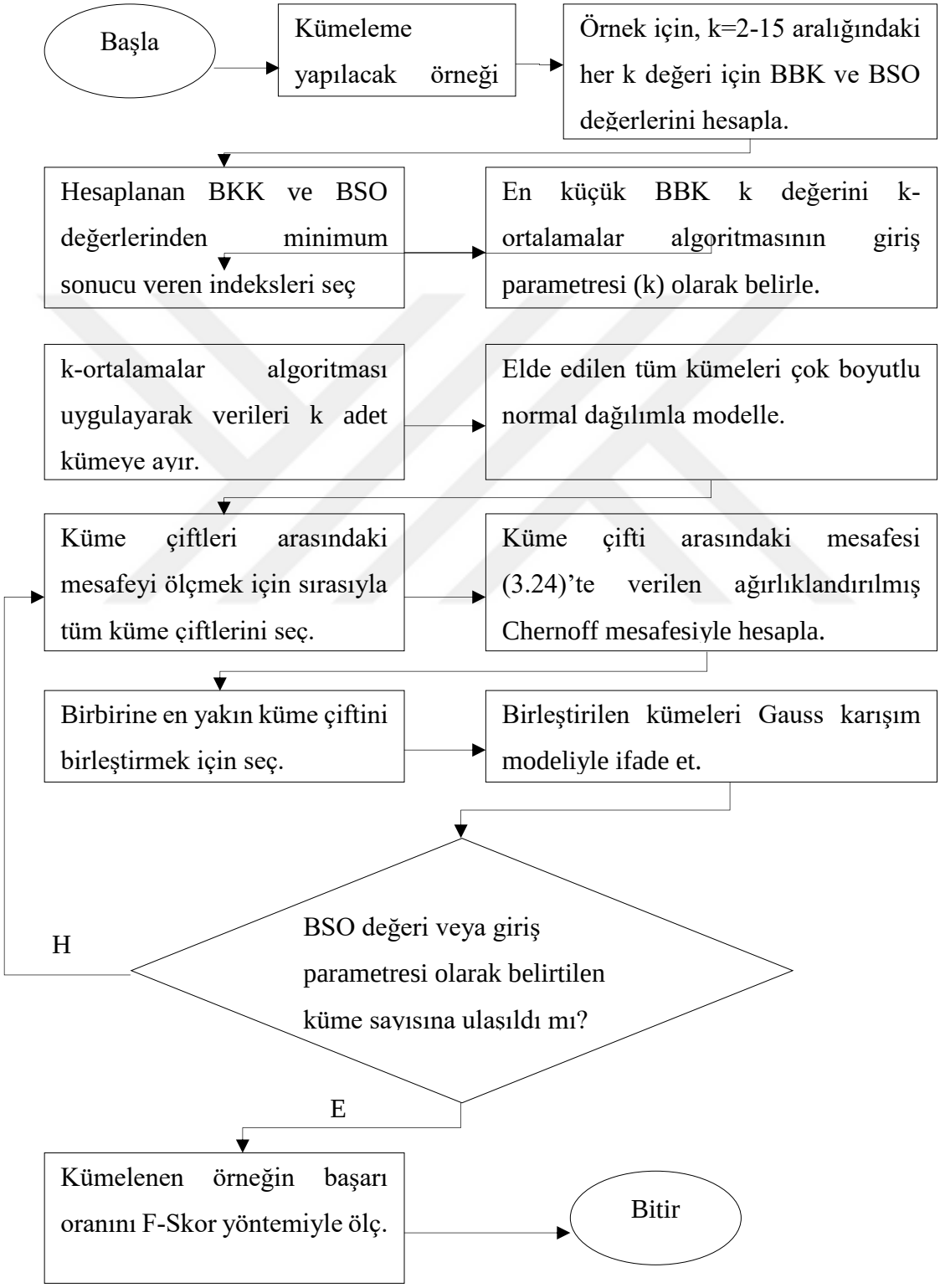
[40] J. P. Baudry, A. E. Raftery, G. Celeux, K. Lo, ve R. Gottardo, "Combining mixture components for clustering," *J. Comput. Graph. Stat.*, vol. 19, no. 2, pp. 332–353, 2010.

[41] "The number of parameters in Gaussian mixture model." [Online]. Available: https://stats.stackexchange.com/questions/229293/the-number-of-parameters-in-gaussian-mixture-model. [Erişim Tarihi: 10.10.2018].

[42] G. J. Mclachlan ve S. Rathnayake, "On the number of components in a Gaussian mixture model," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 4, no. 5, pp. 341–355, 2014.

EKLER

Bölüm 3.7.'de önerilen kümeleme algoritmasının akış diyagramı:



ÖZGEÇMİŞ

Kişisel Bilgiler	
Adı Soyadı	Eyyüp
Doğum Yeri	Yıldız
Doğum Tarihi	14.09.1992
Uyruğu	☑ T.C. ❒ Diğer:
Telefon	+90 446 2240088 | Dahili: 43033
E-Posta Adresi	eyypyldz@gmail.com
Web Adresi	http://aves.erzincan.edu.tr/eyyup.yildiz/

Eğitim Bilgileri	
Lisans	
Üniversite	İstanbul Üniversitesi
Fakülte	Mühendislik Fakültesi
Bölümü	Bilgisayar Mühendisliği Bölümü
Mezuniyet Yılı	31.05.2016

Yüksek Lisans	
Üniversite	İstanbul Üniversitesi-Cerrahpaşa
Enstitü Adı	Lisansüstü Eğitim Enstitüsü
Anabilim Dalı	Bilgisayar Mühendisliği Anabilim Dalı
Programı	Bilgisayar Mühendisliği Programı

Makale ve Bildiriler

[1] M. Günay, E. Yıldız, Y. Nalçakan, B. Aşıroğlu, A. Zencirli, B. R. Mete, T. Ensari, "Digital Data Forgetting: A Machine Learning Approach", IEEE International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Ankara, Turkey, October 19-21, 2018.

[2] E. Yıldız, T. Ensari, "Gender Classification from Face Images", International Journal of Naval Science and Engineering, Vol. 13, No. 1, pp. 31-42, April, 2017.

[3] Book Chapter: T. Ensari, M. Günay, Y. Nalçakan, and E. Yıldız, Machine Learning for Wireless Communications, IGI Global Pub., 2019 [Basım aşamasında].