



Amman - Jordan

## **An Improved BIRCH Algorithm for Breast Cancer Clustering**

**تصنيف مرض سرطان الثدي باستخدام خوارزمية (BIRCH) المحسنة**

**Prepared By**

**Maysarah Mohammad Barham**

**Supervisor**

**Dr. Ahmad Gazi Alzu'bi**

**A Thesis Submitted in Partial Fulfillment of the Requirements for  
the Master Degree in Computer Science**

**Department of Computer Science**

**Faculty of Information Technology**

**Middle East University**

**June 2020**

## Authorization

I, **Maysarah Mohammad Barham**, authorize Middle East University to provide libraries, organizations, and individuals with copies of my thesis on request.

**Name: Maysarah Mohammad Barham**

**Date: 30/6/2020**




**Signature:**

A handwritten signature in blue ink, consisting of several stylized, overlapping loops and curves, positioned above a horizontal line.

## Thesis Committee Decision

This is to certify that the thesis entitled "An Improved BIRCH Algorithm for Breast Cancer Clustering" was successfully defended and approved on 30/6/2020.

### Examination Committee Members:

<b>Dr. Ahmad Gazi Alzu'bi</b> Middle East University	(Supervisor/Chairman) .....	
<b>Dr. Bassam AL-Shargabi</b> Middle East University	(Internal Member) .....	
<b>Prof. Sadeq AL-Hmouz</b> The World Islamic Sciences and Education University	(External Member) .....	

## Acknowledgment

This Thesis is dedicated to my beloved family who has provided and continues to provide me with an abundance of support. Their memories travelled with me even in my educational journey. I am grateful first and foremost to **My father and mother** (May Allah shower her in his mercy), who was my companion even after she passed away. I am also deeply indebted to my beloved wife Narmin and my kids, Adam and Juwan, for their love and patience. I am incredibly grateful for the important role that **My wonderful siblings** Rakan, Dr. Hadi, and Rula played in my life. Last but not least, a special thanks to **Dr. Ahmad Gazi Alzu'bi** who was not only my teacher, but also the best of mentors.

**Maysarah.M.Barham**

**The Researcher**

## **Dedication**

This thesis is dedicated to my family who has always been a constant source of support, encouragement during the challenges of our whole college life, and have been taught us to work hard for the things that we aspire to achieve.

# Table of Contents

<b>Authorization.....</b>	<b>II</b>
<b>Thesis Committee Decision.....</b>	<b>III</b>
<b>Acknowledgment.....</b>	<b>IV</b>
<b>Dedication .....</b>	<b>V</b>
<b>Table of Contents.....</b>	<b>VI</b>
<b>List of Tables.....</b>	<b>VIII</b>
<b>List of Figures.....</b>	<b>IX</b>
<b>Table of Abbreviations .....</b>	<b>X</b>
<b>Abstract.....</b>	<b>XI</b>
<b>Abstract in Arabic .....</b>	<b>XIII</b>
<b>Chapter One: Study Background and Motivation.....</b>	<b>1</b>
1.1 Introduction .....	1
1.2 Study Motivation.....	5
1.3 Problem Statement .....	6
1.4 Research Questions .....	7
1.5 Research Objectives.....	7
1.6 Research Contributions.....	8
<b>Chapter Two: Related work .....</b>	<b>9</b>
2.1 Introduction .....	9
2.2 Background on Clustering Approaches .....	9
2.3 Literature Review on BIRCH Algorithm .....	11
2.4 BIRCH Algorithm for Medical Records .....	16
<b>Chapter Three: Methodology and Proposed Model .....</b>	<b>19</b>
3.1 Introduction .....	19
3.2 Datasets .....	19
3.2.1 Breast Cancer Wisconsin Dataset .....	19
3.2.2 Breast Cancer Wisconsin (Diagnostic) Dataset .....	21
3.3 The Basic BIRCH Algorithm.....	22
3.4 The Framework of Improved BIRCH Algorithm.....	25
3.5 Data Preprocessing .....	26
3.6 Automatic Threshold Initialization.....	28
3.7 Adopted Approaches of Linkage and Distance Measures .....	31
3.7.1 Linkage Methods.....	32

3.7.2	Distance Similarity Metrics. ....	34
3.8	Pseudocode of Improved BIRCH Algorithm .....	35
3.9	Performance Evaluation Measures.....	36
<b>Chapter Four: Results and Discussion.....</b>		<b>39</b>
4.1	Introduction .....	39
4.2	Experiments Setup .....	39
4.3	Results of Basic BIRCH Algorithm.....	39
4.3.1	Data set Breast Cancer Wisconsin .....	40
4.3.2	Data set Breast Cancer Wisconsin (Diagnostic) .....	43
4.4	Results of Improved BIRCH Algorithm .....	44
4.4.1	Data set Breast Cancer Wisconsin .....	44
4.4.2	Data set Breast Cancer Wisconsin (Diagnostic) .....	48
4.5	Comparisons.....	49
<b>Chapter Five: Conclusions and Future Work.....</b>		<b>50</b>
5.1	Conclusions .....	50
5.2	Future Work .....	50
<b>REFERENCES.....</b>		<b>52</b>

## List of Tables

Table No	Contents	Page
Table 2.1	Comparison between Single Threshold and Multiple Thresholds	15
Table 3.1	Samples records of Breast Cancer Wisconsin dataset	21
Table 3.2	Sample records of Breast Cancer Wisconsin (Diagnostic) dataset	22
Table 3.3	Linkage methods and definitions	33
Table 3.4	Distance metrics and definitions	34
Table 3.5	Samples predictions	36
Table 4.1	Clustering results on Breast Cancer Wisconsin executed by linkage (Ward) and different distances Metrics.	40
Table 4.2	Clustering results on Breast Cancer Wisconsin executed by linkage (Centroid) and different distances Metrics.	40
Table 4.3	Clustering results on Breast Cancer Wisconsin executed by linkage (Average) and different distances Metrics.	41
Table 4.4	Clustering results on Breast Cancer Wisconsin executed by linkage (Single) and different distances Metrics.	41
Table 4.5	Clustering results on Breast Cancer Wisconsin (Diagnostic) executed by linkage (Ward) and different distances Metrics.	44
Table 4.6	Clustering results on Breast Cancer Wisconsin executed by linkage (Ward) and different distances Metrics.	44
Table 4.7	Clustering results on Breast Cancer Wisconsin (Diagnostic) executed by linkage (Ward) and different distances Metrics.	48
Table 4.8	the latest data-related findings (Wisconsin breast cancer)	49



## List of Figures

Figure No	contents	Page
Figure 2.1	An Overview of Clustering Algorithms for Big Data Mining.	9
Figure 3.1	The standard BIRCH Algorithm.	23
Figure 3.2	A graphical depiction of improved BIRCH Algorithm.	26
Figure 3.3	The flowchart of improved BIRCH algorithm.	31
Figure 4.1	Clustering results of baseline BIRCH using threshold (0.2) on Breast Cancer Wisconsin.	42
Figure 4.2	Dendrogram of baseline BIRCH using Ward linkage on Breast Cancer Wisconsin.	43
Figure 4.3	Clustering results of Improved BIRCH using threshold ( $T=0.4354$ ) on Breast Cancer Wisconsin, methods linkage (Average) and distances Metrics (Euclidean).	46
Figure 4.4	Clustering results of Improved BIRCH using threshold ( $T=0.3759$ ) on Breast Cancer Wisconsin, methods linkage (Ward) and distances Metrics (Euclidean).	47
Figure 4.5	Dendrogram of Improved BIRCH using Ward linkage on Breast Cancer Wisconsin.	48

## Table of Abbreviations

Abbreviation	Full Description
BIRCH	Balanced Iterative Reducing and Clustering Using Hierarchies
KDD	Knowledge Discovery in Databases
T	Threshold
SOM	Self-Organizing Feature Map
KNIME	Konstanz Information Miner
MBD-BIRCH	Multiple Branch Descent BIRCH
MRI	Magnetic Resonance Imaging
CF	Clustering Feature
HCM	Hierarchical Clustering Method
CF-tree	Clustering Feature Tree
A-BIRCH	Automatic threshold estimation for the BIRCH clustering algorithm
FF	Farthest First
EM	Expectation Maximization
GKA	Genetic k-means Algorithm
MWMOTE	Majority Weighted Minority Oversampling Technique
HBCA	Hybrid Bees Colonies Algorithm

# **An Improved BIRCH Algorithm for Breast Cancer Clustering**

**Prepared By**

**Maysarah Mohammad Barham**

**Supervised By**

**Dr. Ahmad Gazi Alzu'bi**

## **Abstract**

Breast cancer became a popular disease affects women over the world, but in most cases, treatment is possible when discovered early. Screening tests play an important role in identifying tumors before they become cancerous, where diagnosis of breast cancer is more effective compared to other tests. Over the past few decades, the computer-aided diagnosis of cancer has been the subject of research and achieved significant advances. However, the automatic clustering and analysis of patients records in real-time is still a challenging task associated with the selection criteria of BIRCH parameters, and linkage and similarity metrics.

Clustering is an unsupervised machine learning technique used to group data elements without advance knowledge of group definitions. Using aggregation algorithms for a large amount of data could lead to efficiency and accuracy problems. In order to help specialists in making proper decisions while dealing with patients' records, we propose in this thesis work an improved version of the clustering algorithm called balanced iterative reducing and clustering using hierarchies (BIRCH). This approach aims at transforming and clustering the medical records including the disease features into sub-clusters so that the similar features are grouped and analyzed. The proposed improved BIRCH consists of four main components: features selection, features rescale, an efficient automatic threshold initialization, and empirical selection of linkage methods and distance metrics. Specifically, the basic BIRCH clustering is fed with normalized selected features and automatic threshold value to control the tree-based sub-clustering as well as different linkage and similarity measures are involved.

The Breast Cancer Wisconsin dataset is used to evaluate the proposed algorithm. The experimental results show that the improved BIRCH algorithm achieves a clustering accuracy of 97.7% during only 0.0004 seconds, which confirms its efficiency in helping doctors in analyzing the patients' records and making decisions.

**Keywords: BIRCH Clustering, Clustering Feature Tree, Threshold, Dataset Pre-processing, Features Rescaling, linkage and Distance metrics, Data Fitting, Data Prediction.**

## تصنيف مرض سرطان الثدي باستخدام خوارزمية (BIRCH) المحسنة

إعداد: ميسرة محمد حسين برهم

إشراف: الدكتور أحمد الزعبي

### المُلخص

أصبح سرطان الثدي مرضًا شائعًا يصيب النساء في جميع أنحاء العالم، ولكن في معظم الحالات يكون العلاج ممكنًا عند اكتشافه مبكرًا. تلعب اختبارات الفحص دور في تحديد الأورام قبل أن تصبح سرطانية ، حيث يكون تشخيص سرطان الثدي أكثر فعالية مقارنة بالاختبارات الأخرى. على مدى العقود القليلة الماضية ، كان تشخيص مرض السرطان بمساعدة الكمبيوتر موضوعًا للبحث وحقق تقدمًا كبيرًا. ومع ذلك ، فإن التجميع والتحليل التلقائي لسجلات المرضى في الوقت الحقيقي لا يزال مهمة صعبة بسبب معايير الاختيار لمعلمات BIRCH، ومقاييس الربط والتشابه.

التجميع عبارة عن تقنية للتعلم الآلي غير خاضعة للرقابة تُستخدم لتجميع عناصر البيانات دون معرفة مسبقة بتعريفات هذه المجموعة. قد يؤدي استخدام خوارزميات التجميع لكمية كبيرة من البيانات إلى مشاكل في الكفاءة والدقة. من أجل مساعدة المتخصصين في اتخاذ القرارات المناسبة أثناء التعامل مع سجلات المرضى ، نقترح في هذه الأطروحة نسخة محسنة من خوارزمية (BIRCH) للتجميع الهرمي. يهدف هذا النهج إلى تحويل السجلات الطبية وتجميعها بما في ذلك سمات المرض إلى مجموعات فرعية بحيث يتم تجميع وتحليل السمات المماثلة. يتكون (BIRCH) المحسن المقترح من أربعة مكونات رئيسية: اختيار الميزات وإعادة تحديدها، وتهيئة عتبة تلقائية فعالة، واختيار تجريبي لطرق الربط ومقاييس المسافة. على وجه التحديد ، يتم تغذية التجميع الأساسي (BIRCH) بميزات مختارة وقيمة عتبة تلقائية للتحكم في التجمعات الفرعية القائمة على الأشجار بالإضافة إلى إجراءات الربط والتشابه المختلفة.

يتم استخدام مجموعة بيانات القياس (Breast Cancer Wisconsin) لتقييم الخوارزمية المقترحة. حيث أظهرت النتائج التجريبية أن خوارزمية (BIRCH) المحسنة يمكنها تحقيق دقة تجميعية بنسبة (97.7%) خلال (0.0004) ثانية فقط، مما يؤكد كفاءتها في مساعدة الأطباء في تحليل سجلات المرضى واتخاذ القرارات.

الكلمات الرئيسية: خوارزمية (BIRCH) ، شجرة ميزة التجميع ، العتبة ، معالجة بيانات مجموعة البيانات ، إعادة قياس الميزات ، مقاييس الربط والمسافة ، ربط البيانات ، التنبؤ بالبيانات.

---

# Chapter One: Study Background and Motivation

---

## 1.1 Introduction

There is no doubt that we live in an era of wide technological expansion. The amount of data, systems, and users has increased exponentially. Because of this expansion, we are in direct need of methods that allow us to extract, arrange, manipulate and organize data efficiently. Data mining is among such essential methods. The concept of data mining, sometimes called knowledge discovery, is characterized by extracting or mining important information from a large amount of data (Han; Kamber, 2006). Data mining approaches involve a wide range of algorithms that help in pulling a great deal of information that is stored in large databases or information repositories, which allows users to categorize data using different criteria and parameters.

One of important benefits of data mining would be transforming data into information (Jackson, 2002). It has been used in many important daily-life applications including education, commerce, medical domain, and environment. Most importantly, extracting and managing meaningful information from medical records is a challenging task. However, The medical data are significant and need to be investigated carefully many challenges are usually encountered in the medical domain due to their Large amounts of medical data from data generated by media sensors in health monitoring systems and medical data among these challenges :

- 1) Have characteristics of disease diversity.
- 2) Heterogeneity of treatment and outcome.

3) The complexity of data collection, processing, and interpretation, through medical diagnostics that results from media various (audio, visual, image, and text content).

4) Health service providers are so complex that they cannot be treated and analyzed in traditional ways (Jackson, 2002).

Clustering is one of the simplest and yet most beneficial unsupervised approaches that assign data elements into groups of similar objects, i.e. clusters. Accordingly, data in a cluster are similar to each other and dissimilar with objects of other clusters (Tsai, C.; Wu, H.; Tsai, C., 2002). Specifically, the objects in a particular group are very similar or the groups are different from each other. Among the requirements of clustering is scalability (i.e. highly scalable clustering algorithms are required to deal with large databases), versatility of algorithms to work with different kinds of attributes, clustering any type of data such as interval-based (numerical) data, categorical data, and binary data, discovery of clusters with attribute shape (i.e. it is important to develop algorithms that can detect clusters of arbitrary shapes, the ability to deal with noisy data (i.e. databases contain noise, like erroneous data. some algorithms are sensitive to data and this may lead to deterioration of the quality of the clusters), and interpretability, i.e., the clustering results should be interpretable, comprehensible, and usable (Sajana, T.; Rani, C.M.S.; Narayana, V., 2016).

Clustering is beneficial in many practical applications of aggregation algorithms in biomedical research that exist everywhere, and there are exemplary examples that have been applied including analysis of gene expression data, analysis of genome sequences, extraction of biomedical documents, and analysis of MRI images, and magnetic and cancer diagnosis (Obermeyer Z; Emanuel E.j.,2016). However, given the diversi-

ty of clustering analysis, the different terms, objectives, and assumptions underlying the different clustering algorithms can be daunting. Therefore, determining the correct congruence between aggregation algorithms and biomedical applications has become particularly important.

Lately, different methods of clustering have been used to extract the useful clustering the group and increased focus on an exploratory analysis of very large data sets to discover beneficial and/or relationships between traits. However, a proper selection of data and methods for clustering is an important task in medical diagnosis, which needs a sufficient domain knowledge and experience.

An efficient and scalable data clustering method is based on a memory data structure called clustering feature tree (CF-tree), which serves as a summary of the data distribution and called Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH). BIRCH clustering has the ability to cluster multi-dimensional metric data points, either incrementally or dynamically and it also has the ability to produce high clustering accuracy in a single scan, but it improves the clustering quality with more few scans. Indeed, obtaining high clustering performance depends on two elements:

- 1) High intra-class similarity.
- 2) Low inter-class similarity.

Moreover, the clustering quality also depends on both the similarity measure, implementation, and its ability to discover some or all of the hidden patterns.

BIRCH has been used to solve several real-life problems such as building iterative and interactive classifiers and generating codebooks for image compression (Zhang, T.; Ramakrishnan, R.; & Linvy, M., 1996).



In BIRCH clustering tree, a node is known as a clustering feature (CF). It is a small representation of an underlying cluster of one point or many points. BIRCH builds on the idea that points that are close enough to one another should always be considered as a group. The CFs provides this level of abstraction. In other words, the core of the BIRCH clustering algorithm is the CF. Generally, BIRCH algorithm consists of four phases (Zhang, T; Ramakrishnan, R.; Linvy, M., 1996):

- 1) Scanning a database to formulate an in-memory CF tree.
- 2) Building smaller CF trees.
- 3) Performing a global clustering.
- 4) Refining clusters, which is not mandatory and requires more scans of the dataset.

The downside of BIRCH algorithm is that it can only work with numerical data and that it is sensitive to the order of the data records (Zhang, T; Ramakrishnan, R.; Linvy, M., 1996). Basically, BIRCH uses three parameters: the branching factor  $B_r$ , cluster count  $k$ , and the threshold ( $T$ ). While the data points of given dataset are entered into BIRCH, a height-balanced CF tree of hierarchical clusters is built. Each node represents a cluster in the cluster hierarchy where leaf nodes are the actual clusters and intermediate nodes are super clusters. The branching factor  $B_r$  is the maximum number of children a node can have. Then, when a leaf is reached, the new point is added to this leaf cluster, which will not increase the radius of the cluster beyond the threshold ( $T$ ). Otherwise, the new point is assigned into a new created cluster as its only member.

As a result, the size of the clusters is obviously controlled by the threshold parameter  $T$ . Therefore, choosing an optimal threshold is crucial for getting high accuracy of BIRCH clustering. Moreover, BIRCH algorithm could be affected by various meth-

ods of linkage and distance metrics used while constructing the tree sub-clusters and measuring the distance between clusters data points and their centroids.

To investigate and address such important factors of BIRCH algorithm, this research proposes an improved version of basic BIRCH by applying three additional steps as follows:

- 1) Applying data points rescale as a preprocessing step.
- 2) Developing an automatic threshold initialization.
- 3) Utilizing various linkage methods and distance metrics.

We aim at providing more accurate hierarchical clustering approach to diagnose the medical records of breast cancer patients, as a case study, while maintaining the time and memory constraints, which makes it favorable in medical domain. This work studies the BIRCH performance in terms of clustering accuracy, run time complexity, and stability. Stability is the most important characteristic of clustering algorithm that shows the ability to create same data partitions irrespective of the order in which patterns are presented to the algorithm. Thus, stability is considered as an important parameter for achieving an effective clustering performance. We evaluate the performance of the proposed clustering algorithm using two standard datasets: Breast Cancer Wisconsin and Breast Cancer Wisconsin (Diagnostic) (Dua, D.; Graff, C,2019).

## 1.2 Study Motivation

Exploring and managing medical data is challenging task using traditional data mining techniques. Tree-based BIRCH clustering is among algorithms popularly used but it depends on the quality of submitted data, i.e. breast cancer records in our case, and the clusters size controlled by a threshold parameter that should be selected careful-

ly. Additionally, the clustering accuracy is also influenced by the linkage procedure adopted as well as the distance metrics used for data points assignments into clusters. Such important factors have motivated us to propose more additional improvements to the basic BIRCH algorithm's capabilities that could be applied in the medical sector, which in turn improves the quality of healthcare offered to breast cancer patients as well as the service that patients receive.

### **1.3 Problem Statement**

Finding useful clustering approaches for medical datasets has recently attracted a considerable attention, which treats with large amount of obtained medical records. Such medical information provides valuable analytical baseline for diagnosing diseases such as cancer. Several clustering techniques could be used to extract and analyze useful the medical patients' records in order to explore their underlying features and patterns. However, many challenges are usually encountered while applying hierarchical clustering algorithms such as BIRCH in the medical domain due to their limitation in identifying, disseminating and clustering relevant and accurate patients' records.

By enhancing the basic BIRCH clustering algorithm applied on breast cancer patients, we achieve tremendous benefits and overcome several problems associated with the selection criteria of BIRCH parameters, branching factors of sub clustering, and linkage and similarity metrics. As a consequence of improving BIRCH accuracy, the accuracy of disease diagnosis is increased and reduce the processing effort and time.

## 1.4 Research Questions

This work attempts to answer the following research questions:

1. What is the impact of features/data points preprocessing and rescaling on BIRCH clustering quality and performance?
2. Could the automatic threshold initialization improve the accuracy of BIRH algorithm while dealing with diverse breast cancer records?
3. How could changing the linkage methods, utilized in BIRCH, affect the complexity of tree sub-clustering?
4. How could changing the distance metrics, used in assigning medical records into clusters, affect the ability of determining similar/dissimilar records?

## 1.5 Research Objectives

The objectives of this thesis work can be summarized as follows:

1. To analyze and explore the existing BIRCH clustering technique in the medical domain, which highlights the outlier's patterns and investigates the most important parameters that affect the performance of BIRCH algorithm.
2. To propose an improved implementation of BIRCH algorithm by applying data preprocessing along with an efficient automatic threshold initialization.
3. To conduct thorough experiments to investigate the impact of linkage methods and similarity distance metrics on the BIRCH hierarchical clustering during its execution on breast cancer records.
4. To evaluate the improved BIRCH using several standard performance metrics including accuracy and time constraints.

## 1.6 Research Contributions

The main contributions of this thesis work are three-fold and can be summarized as follows:

1. A preprocessing rescaling procedure proposed and applied on data points and dataset features.
2. An automatic initialization of BIRCH algorithm is proposed and discussed thoroughly.
3. An empirical investigation is applied on various linkage methods and distance metrics to find the optimal setups for BIRCH clustering algorithm.

---

## Chapter Two: Related works

---

### 2.1 Introduction

Chapter two provides a brief background of clustering approaches, and it reviews the recent works studied BIRCH algorithm but focusing on the BIRCH Algorithm has been utilized in medical records clustering. Section 2.2 discusses the clustering approaches briefly. Section 2.3 presents a literature review on BIRCH algorithm in general. And Section 2.4 reviews the recent research works applied BIRCH clustering algorithm on medical records.

### 2.2 Background on Clustering Approaches

Many algorithms have been formulated to assist users in achieving their clustering tasks. These algorithms are categorized into five major groups as shown in Figure 2.1 They are the hierarchical, partitioning, density-based, grid-based, and model-based algorithms (Bhardwaj, S.2017). This thesis work will focus on the hierarchical clustering but some clustering algorithms will be defined beforehand.

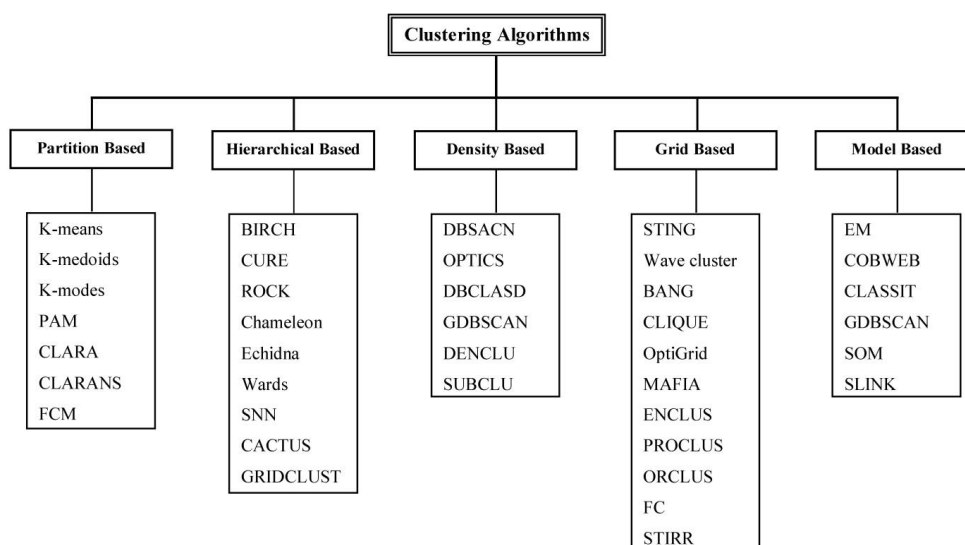


Figure 2.1. An Overview of Clustering Algorithms for Big Data Mining (Sajana, T.; Rani, C.M.S; Narayana, V. ,2016).

**Density-based Clustering Algorithms.** Data objects are classified into core points, border points, and noise points. All the core points are connected together based on their densities to form cluster. It groups points that are closely packed together. The arbitrarily-shaped clusters are formed by the various density-based clustering algorithms like the DBSCAN, OPTICS, DBCLASD, GDBSCAN, DENCLU, and SUBCLU algorithms, listed in Figure 2.1.

**Partitioning Clustering Method.** This method is used to classify data into multiple groups based on their similarity. The various partitioning procedures commonly result in a group of (M) clusters. Ideally, each item belongs to a unique cluster. A centroid or a cluster representative may denote each cluster, which is some sort of summary description of all the entities enclosed within a cluster.

**Hierarchical Clustering Method.** In Hierarchical clustering, data objects are grouped into a tree-like cluster, and every cluster noted contains child clusters within it. This approach is ideal when a user needs exploring data at different levels of complexity and granularity. The hierarchical clustering algorithms build clusters gradually (Sajana, T.; Rani, C.M.S.; Narayana, V., 2016).

Hierarchical clustering can be approached in two ways, either bottom-up clustering or top-down clustering (also known as divisive hierarchical clustering). Agglomerative clustering (hierarchical or bottom-up clustering) starts by merging each basic object in one cluster. It then begins to other clusters into an even larger cluster. The process continues to repeat until all clusters are merged into one, which is the top level of the hierarchical shape.

In divisive hierarchical clustering (top-down clustering), however, we start with one large cluster that contains all objects of the objects, then, subdivide this cluster into

smaller and smaller pieces. This process is repeated until a stopping criterion (the requested number of clusters,  $k$ ) is obtained.

However, there are advantages and disadvantages associated to hierarchical clustering. Among the advantages is the fact that hierarchical clustering is easy to implement and that it occasionally achieves the best results (Berkhin, P., 2006). On the other hand, the biggest disadvantage is the fact that there is no ‘undo’ in this algorithm. It is also sometimes difficult to identify the number of required clusters using this approach (Berkhin, P., 2006). In order to improve the efficiency and quality of the hierarchical clustering method we do have the option to combine it with other clustering methods such as using hierarchies algorithms it as ROCK (Robust Clustering algorithm for categorical attributes) and Chameleon (Tsai, C.; Wu, H.; Tsai, C.2002 & Bhardwaj, S., 2017).

### **2.3 Literature Review on BIRCH Algorithm.**

Many studies have been focused on the utilization of data clustering and mining. In this section, we outline many of these studies to highlight their contributions in the domain of data mining and clustering.

Bhardwaj et al. (2017) compared the accuracy of different data mining techniques. Accuracy is of utmost importance when it involves patients’ health. Therefore, computerizing this vast amount of knowledge improves the standard of the entire system. Data mining is part of the Knowledge Discovery in Databases(KDD) process and the data mining tools perform a comparison of symptoms, treatments, and negative effects so as to investigate the particular action that can be proven to be the simplest and provide best results for a group of patients. Data which are obtained from aid organizations are voluminous and heterogeneous and should be collected and kept in organized



manner. Data mining in health offers unlimited possibilities for analyzing models that are hidden or less visible to common analysis techniques.

Abikoye al. (2018) revolved around the classification algorithms used in data mining, despite the fact that there are different types of algorithms available in data mining for the prediction of a business's future strategy. The decision tree classification technique employed in this work focused mainly on data of student's performance in a high school during a quiz using the KNIME tool.

Chayadevi al. (2012) discussed patterns that were used in the past and their application in medical image processing and searching. There was also discussion that revolved around the need for automated tools that quickly recognize microbes in order to examine medical data prior to expiry. Digital image processing is a key aspect of microscopy. The automated color image segmentation for bacterial image is proposed to classify the bacteria into two broad categories of gram images. Edge detection algorithm with eight neighbor-connectivity contours is used. Bacterial morphological geometric features extracted from microscopy images are used for classification and clustering. The potential and distinguished features are extracted from each bacterial cell. The experimental testing results using the self-organizing map revealed that the obtained bacterial cluster patterns are better than those obtained following the statistical approach.

Tsai et al. (2002) introduced a new data-clustering method for data mining in large databases. The results of simulation concluded that the clustering method they proposed performs better than the Fast SOM (FSOM) combined with the k-means approach (FSOM+K-means) and the Genetic k-means Algorithm (GKA) in all the cases that they studied. Their proposed method also produced much smaller errors than both the FSOM+K-means approach and the GKA.

Pagudpud et al. (2018) illustrated how cluster analysis is applied and analyzed some of the typically used methods of cluster analysis. They underscored that clustering can be executed using a variety of algorithms such as the hierarchical, partitioning, and grid algorithms. There are two parts to hierarchical clustering: partitioning and grid-based hierarchical clustering. Partitioning is the Centroid-based clustering. the value of the k-mean is set. Grid-based clustering is the fastest in terms of the processing time, but that typically depends on the size of the grid rather than the size of the data. The grid-based methods use the single uniform grid mesh to slice up the entire problem domain into cells. The urgency to apply cluster analysis is dramatically increasing. As technology continues to develop, cluster areas will achieve a critical breakthrough in the near future.

Madhumitha et al. (2018) examined various clustering techniques and analyzed the pros and cons of each technique. In addition, it provided information about some commonly used clustering methods, Choice of the clustering algorithm plays a crucial role in cluster analysis. According to the need of the user, these techniques can be applied for better clustering results.

Zhang et al. (1996) examined the BIRCH method and its suitability for dealing with extremely large databases. BIRCH uses its available resources in order to incrementally and dynamically cluster incoming multi-dimensional metric data points in order to produce the best quality clustering. This clustering algorithm can typically find a good clustering solution with a single scan of the data and improve the quality further with few additional scans. BIRCH is also the first clustering algorithm proposed in the database area to handle noise (data points that are not part of the underlying pattern) effectively. Authors evaluated the time/space efficiency, data input order sensitivity, and clustering quality of BIRCH through several experiments. They also presented a com-

parison of the performance of BIRCH against that of CLARANS, which is a clustering method proposed recently for large datasets, and showed that BIRCH is consistently superior to CLARANS.

Ismael et al. (2014) attempted to overcome the previous solutions proposed to overcome the shortcomings of the BIRCH algorithm when a single threshold is used. They suggested, instead, an algorithm that is suitable for very large sets of data. In the algorithm, a CF-tree is built whose all entries in each leaf node must fulfill a uniform threshold ( $T$ ), and the CF tree is rebuilt at each stage using different threshold. This was achieved using multiple thresholds rather than a single threshold.

Table 2.1 demonstrates the difference between using multiple thresholds and single threshold in BIRCH algorithm (Ismael, N.; Alzaalan, M.; Ashour, W., 2014. Owen, R.K.; Cooper, N.J.; Quinn, T.J., Lees, R.; Sutton, A.J., 2018. Mitra, Suh, S.C., & Nandy, J., 2011)

Table 2.1. Comparison between single threshold and multiple threshold in BIRCH Algorithm.

properties	Single Threshold	Multiple Thresholds
<b>Uses</b>	Used in the basic BIRCH algorithm.	Used in the modified (or advanced) BIRCH algorithm.
<b>Performance</b>	Lower performance than multiple thresholds.	Higher performance than single threshold.
<b>Accuracy</b>	Accuracy of single threshold selection depends on whether the histogram is bimodal or not.	Accuracy of multiple threshold selection depends on clear multiple peaks in the histogram.
<b>RAM</b>	Only increases when the random-access memory (RAM) is full.	Does not require full RAM to increase.
<b>Threshold</b>	In most particular situations, sizes of clusters are not equal. So, there is no optimal threshold to use in building the whole CF tree and its CF entries. Hence, using single threshold in building the CF has many shortcomings.	The number of thresholds used in the CF tree will be equal to the number of the CF entries in that tree. These thresholds will not be equal and will be dynamically changed during the clustering operation.
<b>Sensitivity</b>	Have an increased specificity but decreased sensitivity.	Have an increased sensitivity but decreased specificity.
<b>Efficiency</b>	Less than multiple threshold and results in lower clustering efficiency.	Better than single threshold and results in higher clustering efficiency.
<b>Densities and Noise</b>	Less than multiple thresholds to handle data with different densities and noise.	Better than single threshold to handle data with different densities and noise.

Lorbeer al. (2017) introduced-BIRCH, which automates the threshold estimation for BIRCH clustering algorithms. This approach calculates the optimal threshold parameter of this clustering algorithm from the data in order for BIRCH to properly cluster even without the global clustering phase that is usually the final step of this algorithm. This is possible as long as the data meets certain requirements. If those requirements are not met, then A-BIRCH will issue a relevant warning before presenting the results. This method causes the final global clustering step to be unnecessary in many cases, which results in two advantages. First, no need to know the expected number of clusters prior to executing the algorithms. Second, without the computationally demanding final clustering, the fast BIRCH algorithm will perform even faster. For very large data sets, these researchers introduced another variation of BIRCH, called MBD-BIRCH. This version of BIRCH is of particular advantage in conjunction with A-BIRCH but is independent of it and is also of general benefit.

## 2.4 BIRCH Algorithm for Medical Records

Vijayarani al. (2013) evaluated two performance factors such as clustering accuracy and outlier detection accuracy used for analysis. They used the Pima Indian Diabetes and Wisconsin Breast Cancer datasets. Their main objective was to implement the aggregation process in data flows and discover extreme values in data flows. Two clustering algorithms namely (BIRCH with K-Means) and (BIRCH with CLARANS) are used for clustering the data items and finding the outliers in data streams. In order to find the best clustering algorithm for outlier detection, several performance measures are used. They observed that the clustering and outlier detection accuracy is more efficient in BIRCH with CLARANS clustering while compare to BIRCH with K-means with clustering.

Jahanvi al. (2014) studied an early diagnosis of breast cancer patients. They used four different clustering algorithms: k-means, Expectation Maximization (EM), Hierarchical Clustering Method (HCM), and Farthest First (FF) Algorithms for diagnosing the health of the patient using the WEKA environment. The EM algorithm tracks an iterative loop, sub-optimal, which seeks to get the constraints of the probability distribution that can say maximum probability of its characteristics. EM Clustering is model based clusters which is nothing but the abstraction of the k-means clustering data mining algorithm. The FF clustering algorithm performs fast analysis rather than other clustering technique. It is an option of k-means clustering algorithm that seats each cluster center in turn at the peak extreme from the presented cluster centers. This peak must relax contained by the data part because of lesser amount of relocation and modification. They concluded that k-means clustering algorithm and FF algorithm are helpful to the early

diagnosis of breast cancer patients. In the HCM algorithm, their experiment found a high error rate. In EM technique research cannot able to diagnosis 36% of patients.

Lavanya al. (2016). They suggested varied distribution of data samples among different classes, based on theory Majority Weighted Minority Oversampling Technique (MWMOTE) most of the samples get grouped under some classes and rest of the samples belong to the remaining classes. They reported that this approach produces the artificial samples from the biased instructive alternative class samples by means of a clustering approach. Average-linkage agglomerative clustering is used to form clusters. The agglomerative clustering is not appropriate for large databases and has time complexity and highly sensitive to noise. Their proposed system introduces an aggregation algorithm for adoption even in a large database. BIRCH is used in their proposed system to cluster incoming multi-dimensional metric dataset and to produce the unsurpassed clustering with the available resources dynamically. They use the MWMOTE with k-mean clustering.

Gurpreet al. (2018) proposed an algorithm called hybrid been colonies algorithm (HBCA). The HBCA algorithm combines the features of BIRCH clustering algorithm whose feature of insertion and splitting is same as B-Tree algorithm and Partitioning clustering algorithm k-means algorithm. In addition, they implemented using WEKA this algorithm on cancer dataset which is collected. The HBCA algorithm first make call to tree algorithm which is named as k-means algorithm that build a tree containing more than 1500 clusters on cancer dataset.

The procedure of insertion and splitting of this tree algorithm is same as B Tree algorithm but in this algorithm each node of the tree stores the node or tree label, the cluster number and the number of instances in that cluster. These large numbers of clusters are difficult to predict and understand. After that the algorithm make call to k-

means clustering algorithm which clusters the leaf nodes of the clustering algorithm. They compared of proposed algorithm with the existing algorithm k-Means & k-Medoid on Cancer dataset using WEKA data mining tool. They analyzed the results by changing the No. of iterations, Error Rates value specifies that the proposed method gives better performance than k-Means & k-Medoid by reducing the sum of square error, which signifies that HBCA have high intra classification similarity, and is more accurate. In Addition, the proposed algorithm can handle large datasets more effectively.

---

## Chapter Three: Methodology and Proposed Model

---

### 3.1 Introduction

The real-world medical datasets are huge and sometimes have missing and inconsistent data, and such datasets are usually of low quality and could lead to low quality of mining and clustering results. Pre-processing techniques including handling missing data, adjustment, and aggregation could overcome dataset problems and improve the content quality and clustering accuracy. Several improvements are proposed throughout the cycle of BIRCH clustering algorithm, which focus on the datasets pre-processing, features selection, threshold initialization, and linkage/distance alternatives.

The remaining part of this chapter is organized as follows: Section 3.2 presents a detailed description of the datasets utilized in this work. Section 3.3 describes the basic BIRCH algorithm. Section 3.4 illustrates the proposed framework of the improved BIRCH algorithm. Section 3.5 introduces the preprocessing technique adopted in this work. Section 3.6 describes the proposed automatic threshold initialization. Section 3.7 presents the adopted approaches of linkage and distance measures. Section 3.8 summarizes the procedure of improved BIRCH in a Pseudocode algorithm. And Section 3.9 describes the standard measures we used to evaluate the performance of proposed improved BIRCH.

### 3.2 Datasets

**3.2.1 Breast Cancer Wisconsin Dataset** (O. L. Mangasarian; W. H. Wolberg; W. Nick Street, 1992).



This dataset contains of 11 attributes and 699 instances to perform data mining tasks, and its data is divided into different partitions. The dataset contains the following attributes/features:

- 1) Sample code number: id number.
- 2) Clump Thickness.
- 3) Uniformity of Cell Size.
- 4) Uniformity of Cell Shape.
- 5) Marginal Adhesion.
- 6) Single Epithelial Cell Size.
- 7) Bare Nuclei.
- 8) Bland Chromatin.
- 9) Normal Nucleoli.
- 10) Mitoses.
- 11) Class: (2 for benign, 4 for malignant).

The last attribute is used as a cluster label, i.e. 2 for benign and 4 for malignant. The list of features used in our experiments are the attributes within the range 2-10, and ID number is excluded. Sample records of this dataset are shown in Table 3.1.

Table 3.1. Samples records of Breast Cancer Wisconsin dataset.

id	clump	thickuniformity	uniformity	marginal	$\epsilon$ single	epitbare	nuclebland	chrcnormal	numitoses	class
1000025	5	1	1	1	2	1	3	1	1	
1002945	5	4	4	5	7	10	3	2	1	
1015425	3	1	1	1	2	2	3	1	1	
1016277	6	8	8	1	3	4	3	7	1	
1017023	4	1	1	3	2	1	3	1	1	
1017122	8	10	10	8	7	10	9	7	1	
1018099	1	1	1	1	2	10	3	1	1	
1018561	2	1	2	1	2	1	3	1	1	
1033078	2	1	1	1	2	1	1	1	5	
1033078	4	2	1	1	2	1	2	1	1	
1035283	1	1	1	1	1	1	3	1	1	
1036172	2	1	1	1	2	1	2	1	1	
1041801	5	3	3	3	2	3	4	4	1	
1043999	1	1	1	1	2	3	3	1	1	
1044572	8	7	5	10	7	9	5	5	4	
1047630	7	4	6	4	6	1	4	3	1	
1048672	4	1	1	1	2	1	2	1	1	
1049815	4	1	1	1	2	1	3	1	1	
1050670	10	7	7	6	4	10	4	1	2	
1050718	6	1	1	1	2	1	3	1	1	
1054590	7	3	2	10	5	10	5	4	4	
1054593	10	5	5	3	6	7	7	10	1	
1056784	3	1	1	1	2	1	2	1	1	
1057013	8	4	5	1	2	1	7	3	1	
1059552	1	1	1	1	2	1	3	1	1	
1065726	5	2	3	4	2	7	3	6	1	
1066373	3	2	1	1	1	1	2	1	1	
1066979	5	1	1	1	2	1	2	1	1	
1067444	2	1	1	1	2	1	2	1	1	
1070935	1	1	3	1	2	1	1	1	1	
1070935	3	1	1	1	1	1	2	1	1	
1071760	2	1	1	1	2	1	3	1	1	
1072179	10	7	7	3	8	5	7	4	3	
1074610	2	1	1	2	2	1	3	1	1	

### 3.2.2 Breast Cancer Wisconsin (Diagnostic) Dataset (O. L. Mangasarian and W. H. Wolberg, 1995)

This dataset contains of 31 attributes and 569 instances to perform data mining tasks, and it is divided into different partitions. We mention the most important of these attribute/features:

- 1) Diagnosis (M = malignant, B = benign) used as cluster label, i.e.0 and 1.
- 2) Ten real-valued features are computed for each cell nucleus:
  - Radius (mean of distances from center to points on the perimeter).
  - Texture (standard deviation of gray-scale values).
  - Perimeter.
  - Area.
  - Smoothness (local variation in radius lengths).
  - Compactness ( $\text{perimeter}^2 / \text{area} - 1.0$ ).

- Concavity (severity of concave portions of the contour).
- Concave points (number of concave portions of the contour).
- Symmetry.
- Fractal dimension ("coastline approximation" - 1).

The list of features used in our experiments are the attributes within the range 2-30, and ID number is excluded. Sample records of this dataset are shown in Table 3.2.

Table 3.2. Sample records of Breast Cancer Wisconsin (Diagnostic) dataset

id	diagnosis	radius_me	texture_me	perimeter_me	area_me	smoothness	compactness	concavity	concave points	symmetry	fractal_dim	radius_se	texture_se	perimeter_se	area_se	smoothness
842302 M		17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419	0.07871	1.095	0.9053	8.589	153.4	0.006399
842517 M		20.57	17.77	132.9	1326	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	0.5435	0.7339	3.398	74.08	0.005225
84300903 M		19.69	21.25	130	1203	0.1096	0.1599	0.1974	0.1279	0.2069	0.05999	0.7456	0.7869	4.585	94.03	0.00615
84348301 M		11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	0.2597	0.09744	0.4956	1.156	3.445	27.23	0.00911
84358402 M		20.29	14.34	135.1	1297	0.1003	0.1328	0.198	0.1043	0.1809	0.05883	0.7572	0.7813	5.438	94.44	0.01149
843796 M		12.45	15.7	82.57	477.1	0.1278	0.17	0.1578	0.08089	0.2087	0.07613	0.3345	0.8902	2.217	27.19	0.00751
844359 M		18.25	19.98	119.6	1040	0.09463	0.109	0.1127	0.074	0.1794	0.05742	0.4467	0.7732	3.18	53.91	0.004314
84458202 M		13.71	20.83	90.2	577.9	0.1189	0.1645	0.09366	0.05985	0.2196	0.07451	0.5835	1.377	3.856	50.96	0.008805
844981 M		13	21.82	87.5	519.8	0.1273	0.1932	0.1859	0.09353	0.235	0.07389	0.3063	1.002	2.406	24.32	0.005731
84501001 M		12.46	24.04	83.97	475.9	0.1186	0.2396	0.2273	0.08543	0.203	0.08243	0.2976	1.599	2.039	23.94	0.007149
845636 M		16.02	23.24	102.7	797.8	0.08206	0.06669	0.03299	0.03323	0.1528	0.05697	0.3795	1.187	2.466	40.51	0.004029
84610002 M		15.78	17.89	103.6	781	0.0971	0.1292	0.09954	0.06606	0.1842	0.06082	0.5058	0.9849	3.564	54.16	0.005771
846226 M		19.17	24.8	132.4	1123	0.0974	0.2458	0.2065	0.1118	0.2397	0.078	0.9555	3.568	11.07	116.2	0.003139
846381 M		15.85	23.95	103.7	782.7	0.08401	0.1002	0.09938	0.05364	0.1847	0.05338	0.4033	1.078	2.903	36.58	0.009769
84667401 M		13.73	22.61	93.6	578.3	0.1131	0.2293	0.2128	0.08025	0.2069	0.07682	0.2121	1.169	2.061	19.21	0.006429
84799002 M		14.54	27.54	96.73	658.8	0.1139	0.1595	0.1639	0.07364	0.2303	0.07077	0.37	1.033	2.879	32.55	0.005607
848406 M		14.68	20.13	94.74	684.5	0.09867	0.072	0.07395	0.05259	0.1586	0.05922	0.4727	1.24	3.195	45.4	0.005718
84882001 M		16.13	20.68	108.1	798.8	0.117	0.2022	0.1722	0.1028	0.2164	0.07356	0.5692	1.073	3.854	54.18	0.007026
849014 M		19.81	22.15	130	1260	0.09831	0.1027	0.1479	0.09498	0.1582	0.05395	0.7582	1.017	5.865	112.4	0.006494
8510426 B		13.54	14.36	87.46	566.3	0.09779	0.08129	0.08664	0.04781	0.1885	0.05766	0.2699	0.7886	2.058	23.56	0.008462
8510653 B		13.08	15.71	85.63	520	0.1075	0.127	0.04568	0.0311	0.1967	0.06811	0.1852	0.7477	1.383	14.67	0.004097
8510824 B		9.504	12.44	60.34	273.9	0.1024	0.06492	0.02956	0.02076	0.1815	0.06905	0.2773	0.9768	1.909	15.7	0.009606
8511133 M		15.34	14.26	102.5	704.4	0.1073	0.2135	0.2077	0.09756	0.2521	0.07032	0.4388	0.7096	3.384	44.91	0.006789
851509 M		21.16	23.04	137.2	1404	0.09428	0.1022	0.1097	0.08632	0.1769	0.05278	0.6917	1.127	4.303	93.99	0.004728
852552 M		16.85	21.38	110	904.6	0.1121	0.1457	0.1525	0.0917	0.1995	0.0633	0.8088	0.9017	5.455	102.6	0.006048
852631 M		17.14	16.4	116	912.7	0.1186	0.2276	0.2229	0.1401	0.304	0.07413	1.046	0.976	7.276	111.4	0.008029
852763 M		14.58	21.53	97.41	644.8	0.1054	0.1868	0.1425	0.08783	0.2252	0.06924	0.2545	0.9832	2.11	21.05	0.004452
852781 M		18.61	20.25	122.1	1094	0.0944	0.1066	0.149	0.07731	0.1697	0.05699	0.8529	1.849	5.632	93.54	0.01075
852973 M		15.3	25.27	102.4	732.4	0.1082	0.1697	0.1683	0.08751	0.1926	0.0654	0.439	1.012	3.498	43.5	0.005233
853201 M		17.57	15.05	115	955.1	0.09847	0.1157	0.09875	0.07953	0.1739	0.06149	0.6003	0.8225	4.655	61.1	0.005627
853401 M		18.63	25.11	124.8	1088	0.1064	0.1887	0.2319	0.1244	0.2183	0.06197	0.8307	1.466	5.574	105	0.006248
853612 M		11.84	18.7	77.93	440.6	0.1109	0.1516	0.1218	0.05182	0.2301	0.07799	0.4825	1.03	3.475	41	0.005551
85382601 M		17.02	23.98	112.8	899.3	0.1197	0.1496	0.2417	0.1203	0.2248	0.06382	0.6009	1.398	3.999	67.78	0.008268
854002 M		19.27	26.47	127.9	1162	0.09401	0.1719	0.1657	0.07593	0.1853	0.06261	0.5558	0.6062	3.528	68.17	0.005015
854039 M		16.13	17.88	107	807.2	0.104	0.1559	0.1354	0.07752	0.1998	0.06515	0.334	0.6857	2.183	35.03	0.004185
854253 M		16.74	21.59	110.1	869.5	0.0961	0.1336	0.1348	0.06018	0.1896	0.05656	0.4615	0.9197	3.008	45.19	0.005776
854268 M		14.25	21.72	93.63	633	0.09823	0.1098	0.1319	0.05598	0.1885	0.06125	0.286	1.019	2.657	24.91	0.005878
854941 B		13.03	18.42	82.61	523.8	0.08983	0.03766	0.02562	0.02923	0.1467	0.05863	0.1839	2.342	1.17	14.16	0.004352
855133 M		14.99	25.2	95.54	698.8	0.09387	0.05131	0.02398	0.02899	0.1565	0.05504	1.214	2.188	8.077	106	0.006883

### 3.3 The Basic BIRCH Algorithm

The basic BIRCH algorithm is introduced in this section as baseline to be improved, and its main steps are shown in Figure 3.1. We refer the readers to the work introduced by Dong et al. (2013) for further details. The basic BIRCH algorithm consists of four phases as follows:

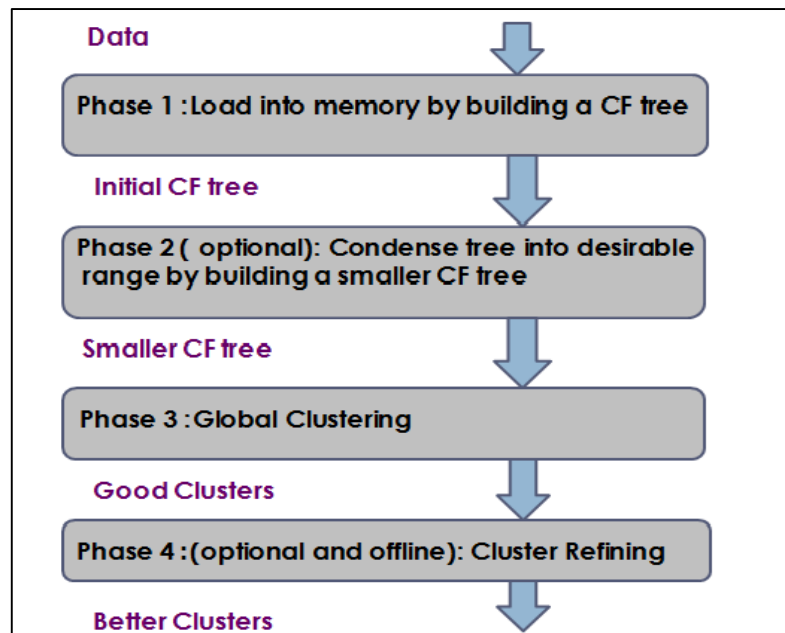


Figure 3.1. The standard BIRCH Algorithm.(Dong, J.; Wang, F.; Yuan, B.; Dong, J.; Wang, F; Yuan, B. (2013)).

- **Phase 1: Load data into memory CF Tree.** This phase applies an initial scan of the database and loads data into memory by building a CF tree. If all memory is consumed, then the tree must be rebuilt from the leaf node.
- **Phase 2: Condense data (Condensing is optional).** This resizes the dataset by building a smaller CF tree. We can also regroup crowded sub-clusters into larger clusters thus creating a smaller CF tree.
- **Phase 3: Global clustering.** It uses an existing clustering algorithm (e.g.,  $k$ -means and HC) on the CF entries.
- **Phase 4: Cluster refining (Refining is optional).** This rescans the original raw data to ensure inaccuracies are corrected. Cluster refining fixes the CF trees problem occurred when the original data get scanned only once.

The steps of basic BIRCH algorithm are summarized as follows (Zhang T.; Ramakrishnan, R ; & Livny, 1996):

1. The dataset records are transformed into the clustering feature (CF). The clustering feature contains three parameters that affect clusters of given data points. It is denoted as follows:

$$CF = (N, LS, SS) \quad (3.1)$$

- $N$  is a number of data points.
- $LS$  is the linear sum of the  $N$  data points,

$$LS: \sum_{i=1}^n x_i \quad (3.2)$$

- $SS$  is the square sum of the  $N$  data points.

$$SS: \sum_{i=1}^n x_i^2 \quad (3.3)$$

A clustering feature is one type of summary of a given cluster. Using it, we can derive many parameters like:

- Centroid:

$$\mathbf{X}_0 = \sum_{i=1}^n \mathbf{X}_i = \frac{LS}{n} \quad (3.4)$$

- Radius:

Average distance from any point of the cluster to its Centroid:

$$R = \sqrt{\frac{\sum_{i=1}^n (x_i - x_0)^2}{n}} = \sqrt{\frac{nSS - 2LS^2 + nLS}{n^2}} \quad (3.5)$$

- Diameter: square root of average mean squared distance between all pairs of points in the cluster:

$$D = \sqrt{\frac{\sum_{i=1}^n \sum_{j=1}^n (x_i - x_j)^2}{n}} = \sqrt{\frac{2nSS - 2LS^2}{n(n-1)}} \quad (3.6)$$

Given two cluster  $C_1$  and  $C_2$ , if they are merged then resultant clustering feature is  $CF$  result=  $CF_1 + CF_2 = (n_1+n_2, LS_1+LS_2, SS_1+ SS_2)$

CF tree is a height-balanced tree that contains clustering features. Non-leaf node has descendants and all non-leaf node store sums of CFs of their children, there are two parameters in CF tree: (threshold  $T$  and branching factor  $B$ ).

- The branching factor ( $B$ ): When all the data set has been changed in the form of CF, then CF-Tree starts working to bring together several formed CFS.
- Threshold ( $T$ ): Before scanning any data points from the database, we must initialize the initial CF tree threshold, this threshold will be used as the initial threshold value for each new CF entry that will not be changed during the grouping process.

2. In a standard BIRCH, you have to initialize  $L$  (number of leaves). Two other parameters  $m$  and  $b$  are added as follows:

- Parameter  $b$  is used to count the number of branches on CF-non leaf.
- Parameter  $m$  is used to count the number of leaf branches on CF-leaf.

3. For a given data record, BIRCH compares the location of the record with the location of each CF at the root node, using a linear number or average CF. Then, BIRCH passes the entry to the CF root node closest to the entry record.

4. The node then descends to the non-leaf child node of the CF nodes selected in step 5. BIRCH compares the location of records with the location of each non-leaf CF. BIRCH passes the node that goes to the non-leaf CF node closest to the entry.

5. The node then descends to the leaf child node of the non-leaf CF node as will be selected in step 6. BIRCH compares the record location with the location of each leaf. BIRCH temporarily passes the entry to the closest leaf with the entry node.

6. Do one (a) or (b):

- (a) If the selected leaf radius ( $R$ ) including a new node does not exceed a threshold  $T$ , then the entry entered is assigned to that leaf. Leaves and all parents CFs are updated to take into account new data points.
- (b) If the selected leaf radius including the new record exceeds a threshold  $T$ , then a new leaf is formed, consisting of incoming notes only. CF parent updated to account for new data points.

7. If the leave ( $m$ ) branch has exceeded the specified Leave ( $L$ ) limit, there will be an additional branch ( $B$ ).

8. If  $B$  has exceeded, there will be a split parent process on CF and then it will be combined again with the new high CF formed.

### 3.4 The Proposed Framework of Improved BIRCH Algorithm

Figure 3.2 shows the main steps involved in our proposed BIRCH algorithm for medical records clustering. In this framework. We summarize the main steps here, and each step is thoroughly illustrated in the next sections:

1. **Dataset preprocessing.** We will use the benchmarking medical datasets (Breast Cancer Wisconsin and Breast Cancer Wisconsin (Diagnostic)), preprocess the data records and features by selecting the most relevant features and fitting them to the corresponding clusters labels (Benign and Malignant), and detecting outliers eliminated using Features rescale.

2. **Automatically threshold initialization.** The threshold value is initialized automatically by a proposed function of three stages based on a subset selected from the given features randomly.
3. **Features Rescaling.** An efficient way of normalizing the dataset features into a range of numbers based on the calculation of minimum and maximum values of all medical records.
4. **Altering linkage and distance metrics in the baseline BIRCH.** We use a range of linkage methods and similarity distances during the execution of BIRCH.
5. **Data Fitting.** Fit the dataset records with their corresponding labels.
6. **Data Prediction.** Predict and assign all records into their proper cluster, and the performance is evaluated using several standard measures.

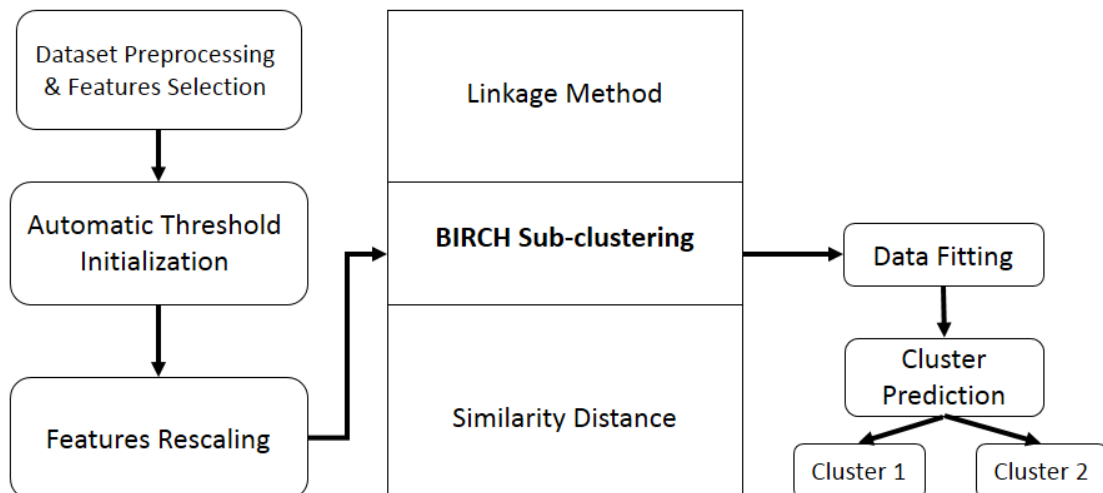


Figure 3.2 A graphical depiction of improved BIRCH Algorithm.

### 3.5 Data Preprocessing

In this research, the breast cancer dataset and Breast Cancer Wisconsin (Diagnostic) dataset were preprocessed, by preprocess the data records and features by selecting the most relevant features and fitting them to the corresponding clusters labels (Benign and Malignant), and detecting outliers eliminated, using Features rescale. The data preprocessing consists of two main processes :

- **Features Selection.** Different features affect clusters differently, some are important for clusters while others may challenge the clustering task. It helps in finding clusters efficiently, understanding the data better, and reducing data size for efficient storage, collection, and processing. In this work, record features were represented as a vector  $x = [x_1, \dots, x_i]$ , which produce a matrix of dataset records size.
- **Features rescale.** In this phase, repetitive records were eliminated using the min-max normalization method. Afterward, the data were divided into two groups of arrays. Were data divided into two arrays of  $x$  and  $y$ , and they could be rewritten as follows:

$x = [x_1, \dots, x_i]$  is the features of each medical record.

$y = [y_1, \dots, y_i]$  is the labels vector.

Random sampling is one of the simplest forms of collecting data from the total dataset. Under random sampling, each member of the subset carries an equal opportunity of being chosen as a part of the sampling process. We determined the size of this sample random to be 50% from all dataset (e.g. 50% of Winconson dataset records). Then, we inserted these random samples in the automatic threshold function, and we created the task scale range for the array element.

The procedure of features rescale to get the new normalized features  $F_{\text{new}}$  is applied using the following formula[MATLAB]:

$$F_{\text{new}} = \text{rescale}(F_{\text{old}}, 'inputMin', Mn, 'inputMax', Mx) \quad (3.7)$$

Where  $F_{\text{old}}$  is a matrix of input features,  $Mn$  is a vector of minimum value of each dataset feature column,  $Mx$  is a vector of maximum value of each dataset feature column,  $inputMin$  is the lower bounding limit of normalization interval, and  $inputMax$  is the upper bounding limit of normalization interval.



This function scales the elements of features matrix into new values within the bounds of  $Mn$  and  $Mx$ . It takes the dataset features into a range of numbers based on the calculation of minimum and maximum values of all medical records, i.e.  $[Mn, Mx]$ . Then, it rescales along the dimension of the input array that corresponds with the shape of the 'InputMin' and 'InputMax' bounding limits.

### 3.6 Automatic Threshold Initialization

The BIRCH algorithm is a matching grouping algorithm for any given datasets, where a CF-tree is built in which all entries in each leaf node must meet the same T threshold using a static (fixed) threshold that usually produces poor cluster quality. In this search, the threshold value will be initialized automatically in the CF-entry, which aims at improving the clustering accuracy. Therefore, the T parameter to CF-Leaf is used to store the latest changes from the threshold used. The T addition parameter is only used for information about CF-Leaf while the CF-Node still uses the formula  $CF = (N, LS, \text{ and } SS)$ .

In the standard BIRCH when the data point has found the CF-node through the calculation of the closest distance. Then the data point will enter the CF-leaf if the radius on the leaf does not exceed the threshold (T). But if it exceeds the threshold value, a new leaf will be built and if it exceeds the leaf limit, there will be a split parent. Whereas in the modified BIRCH, the new data point that goes beyond the threshold will be initialized automatically. That is by enlarging the scale on the leaf radius so that it can reduce split parent in BIRCH.

The main steps of the proposed automatic threshold initialization can be summarized as follows:

**Step 1.** The features matrices initially segmented into two parts using a randomly-select starting threshold value, denoted as  $T(1)$ . Then, the data are clustered into two classes, denoted as  $c1$  and  $c2$ .

---

**Algorithm 1.** Features Splitting

---

**Input:** sample points at random from dataset (I)

**Output:** initial threshold

**Begin**

N: input random sample.

I: features matrix.

Counts: sum of array element.

Sum: sum of array element.

T:threshold, var mu1, sum, sumb, counts, n

**1.1** Compute mean intensity of random sample from dataset, and set  $T(1)=\text{mean}(I)$

[counts, N] = features matrix(I).

//(random sample 50% of records)

i = 1.

//counter for the generations of T

mu1 = cumsum(counts).

//where cumsum is Cumulative sum

**1.2** Round to nearest decimal or integer

$T(i) = (\text{sum}(N.*\text{counts})) / \text{mu1}(\text{end})$ .

**end**

---

**Step 2.** A new threshold value is computed as the average of the above two sample means as described in Algorithm 2.

---

**Algorithm 2.** Mean Computations.

---

**Input:** features mean

**Output:** a new threshold value

**begin**

MBT = calculate mean below current threshold

MAT = calculate mean above current threshold

Counts: sum of array element.

N: input random sample.

T:threshold

Sum: sum of array element.

**2.1** calculate mean below current threshold

$\text{MBT} = \text{sum}(N(N \leq T(i)).*\text{counts}(N \leq T(i))) / \text{mu2}(\text{end})$ .

**2.2** calculate mean above current threshold

$\text{MAT} = \text{sum}(N(N > T(i)).*\text{counts}(N > T(i))) / \text{mu3}(\text{end})$ .

the new threshold is the mean of MAT and MBT

**2.3**  $T(i) = (\text{MAT} + \text{MBT}) / 2$ .

**end**

---

**Step 3.** This step repeats Step 2 until the threshold value does not change anymore, as described in Algorithm 3.

---

**Algorithm 3.** Threshold Recalculations

---

**Input:** setup new threshold

**Output:** best threshold

**begin**

**3.1** repeat step 2 (Algorithm 2)

while  $T(i) \neq T(i-1)$

The features matrix is  $T(i)$

**3.2** while  $\text{abs}(\text{newT}(i) - \text{old } T(i-1)) \neq 1$  do:

$\mu_2 = \text{cumsum}(\text{counts}(N \leq T(i)))$ .

$\text{MBT} = \text{sum}(N(N \leq T(i)) \cdot \text{counts}(N \leq T(i))) / \mu_2(\text{end})$ .

$\mu_3 = \text{cumsum}(\text{counts}(N > T(i)))$ .

$\text{MAT} = \text{sum}(N(N > T(i)) \cdot \text{counts}(N > T(i))) / \mu_3(\text{end})$ .

$i = i + 1$ .

**3.3** repeat step 2 if  $T(i) \neq T(i-1)$

$T(i) = (\text{MAT} + \text{MBT}) / 2$ .

The features matrixes =  $T(i)$ .

end while

**end**

---

During the BIRCH algorithm, the data items are iteratively joined to form clusters, merging first the clusters that are at the minimum distance. However, given two clusters, each one formed by several data observations, there exist many ways of defining the distance between the clusters from the dissimilarities between their constituent individuals. Among these linkage methods (e.g. single, complete, Ward, Centroid). Figure 3.3 shows the flowchart of the improved BIRCH algorithm.

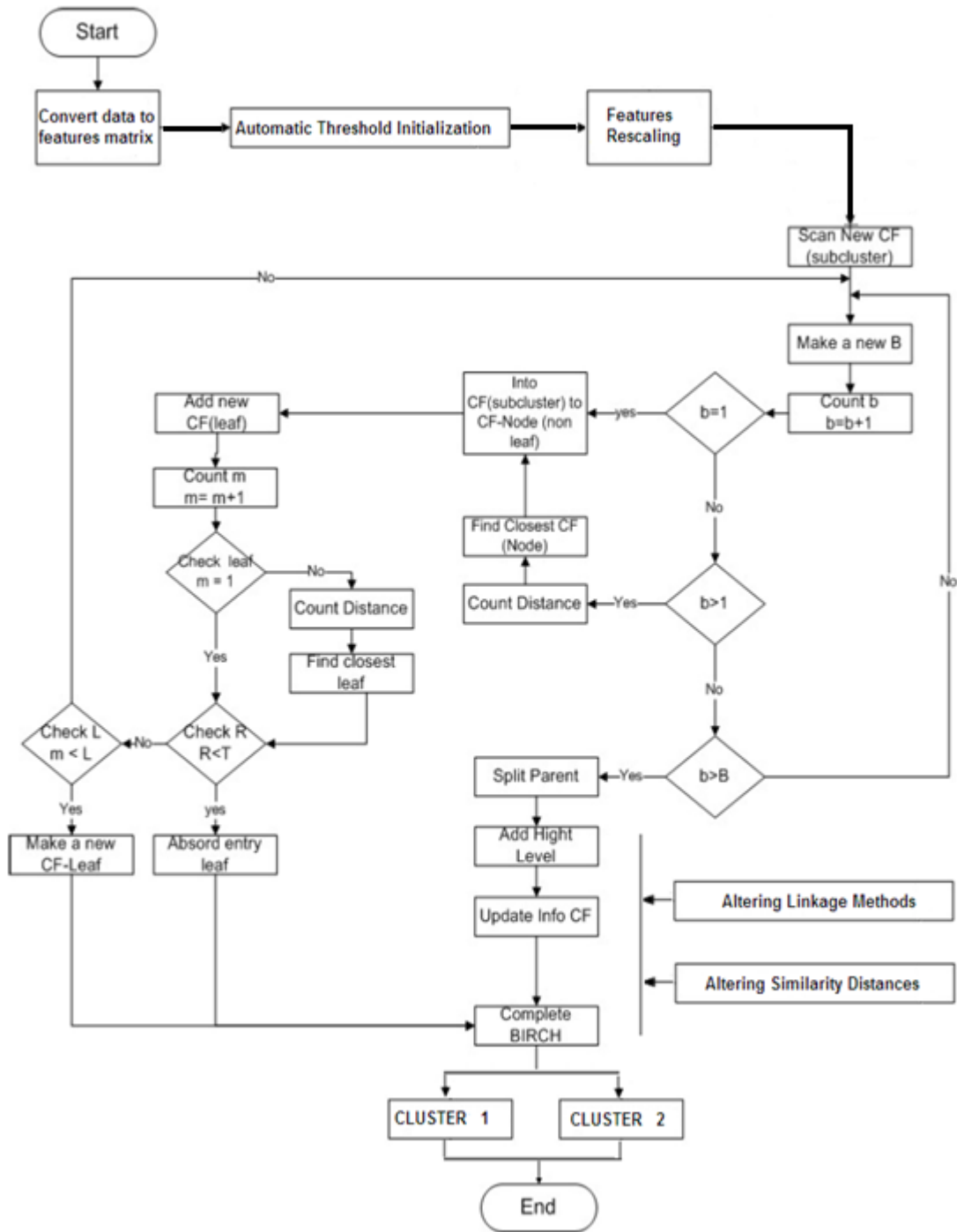


Figure 3.3. The flow chart of improved BIRCH algorithm.

### 3.7 Adopted Approaches of Linkage and Distance Measures

We introduce in this section the linkage methods and distance measures adopted in our experiments while running the improved BIRCH algorithm, which aims at seek-

ing for the best alternative for getting better clustering accuracy, and memory and time complexity.

The difference between the available hierarchical clustering methods rests in the way the distance between clusters is defined. For instance, during the agglomeration process, the data items are iteratively joined to form clusters, merging first the clusters that are at the minimum distance. However, given two clusters, each one formed by several data observations, there exist many ways of defining the distance between the clusters from the dissimilarities between their constituent individuals.

### 3.7.1 Linkage Methods

The difference between the available hierarchical clustering methods rests in the way the distance between clusters is defined. Data items are iteratively joined to form clusters, merging first the clusters that are at the minimum distance. However, given two clusters, each one formed by several data observations, there exist many ways of defining the distance between the clusters from the dissimilarities between their constituent individuals. The following notation describes the linkages used by the various methods:

- Cluster  $r$  is formed from clusters  $p$  and  $q$ .
- $n_r$  Is the number of objects in cluster  $r$ .
- $x_{ri}$  Is the  $i$ th object in cluster  $r$ .

Table 3.3 defines the linkage methods examined in this research.

Table 3.3. Linkage methods and definitions.

Method	Description
<b>Single</b>	The distance between clusters equals the minimum distance between individuals, also called nearest neighbor, uses the smallest distance between objects in the two clusters. $d(r, s) = \min \left( \text{dist}(x_{ri}, x_{sj}) \right), i \in (1, \dots, n_r), j \in (1, \dots, n_s) \quad (3.8)$
<b>Complete</b>	The distance between clusters equals the maximum distance between individuals, also called <i>farthest neighbor</i> , and uses the largest distance between objects in the two clusters. $d(r, s) = \max(\text{dist}(x_{ri}, x_{sj})), i \in (1, \dots, n_r), j \in (1, \dots, n_s) \quad (3.9)$
<b>Ward</b>	The distance between clusters is a weighted squared Euclidean distance between the Centroids of each cluster. $d(r, s) = \sqrt{\frac{2n_r n_s}{(n_r - n_s)}} \ \bar{x}_r - \bar{x}_s\ _2 \quad (3.10)$ <p>Where:</p> <ul style="list-style-type: none"> <li>• <math>\ \cdot\ _2</math> is the Euclidean distance.</li> <li>• <math>\bar{x}_r</math> and <math>\bar{x}_s</math> are the centroids of cluster <math>r</math> and <math>s</math>.</li> <li>• <math>n_r</math> and <math>n_s</math> are the number of elements in clusters <math>r</math> and <math>s</math></li> </ul>
<b>Centroid</b>	The distance between clusters equals the square of the Euclidean distance between the Centroids of each cluster. Also known as WPGMC (weighted version) or UPGMC (unweighted version). $d(r, s) = \ \bar{x}_r - \bar{x}_s\ _2 \quad (3.11)$ <p>Where:</p> <ul style="list-style-type: none"> <li>• <math>\bar{x}_r = \frac{1}{n} \sum_{i=1}^{n_r} x_{ri}</math></li> </ul>
<b>Average</b>	Unweighted average distance (UPGMA), average linkage uses the average distance between all pairs of objects in any two clusters. $d(r, s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} \text{dist}(x_{ri}, x_{sj}) \quad (3.12)$
<b>Median</b>	Weighted center of mass distance (WPGMC), appropriate for Euclidean distances only, <i>Median linkage</i> uses the Euclidean distance between weighted Centroids of the two clusters. $d(r, s) = \ \tilde{x}_r - \tilde{x}_s\ _2 \quad (3.13)$ <p>Where:</p> <p><math>\tilde{x}_r</math> and <math>\tilde{x}_s</math> are weighted centroids for the clusters <math>r</math> and <math>s</math>. If cluster <math>r</math> was created by combining clusters <math>p</math> and <math>q</math>, <math>\tilde{x}_s</math> is defined recursively as : <math>\tilde{x}_r = \frac{1}{2}(\tilde{x}_p - \tilde{x}_q)</math></p>

### 3.7.2 Distance Similarity Metrics

The distance metrics are usually used to know the input data pattern in order to make any data-based decision. A good distance metric helps in improving the performance of classification, clustering, and any information processing significantly. In this research, we will investigate different distance metrics and how do they help in exploring the similarities between records instances and predicting the best cluster for each medical record. Table 3.4 lists all the similarity metrics we used with their definitions.

Table 3.4. Distance metrics and definitions.

Metric	Description
<b>Minkowski</b>	<p>For the special case of <math>p = 1</math>, the Minkowski distance gives the city block distance. For the special case of <math>p = 2</math>, the Minkowski distance gives the Euclidean distance. For the special case of <math>p = \infty</math>, the Minkowski distance gives the Chebychev distance.</p> $d_{st} = \sqrt[p]{\sum_{j=1}^n  x_{sj} - x_{tj} ^p} \quad (3.14)$
<b>Euclidean</b>	<p>The Euclidean distance is a special case of the Minkowski distance, where <math>p = 2</math>.</p> $d_{st}^2 = (x_s - x_t)(x_s - x_t) \quad (3.15)$
<b>Squaredeuclidean</b>	<p>Squared Euclidean distance. (This option is provided for efficiency only. and it's a standard approach to regression analysis)</p> $d_{st}^2 = (x_{s1} - x_{t1})^2 + (x_{s2} - x_{t2})^2 + \dots + (x_{si} - x_{ti})^2 + \dots + (x_{sn} - x_{tn})^2 \quad (3.16)$
<b>Seuclidean</b>	<p>Standardized Euclidean distance. Each coordinate difference between observations is scaled by dividing by the corresponding element of the standard deviation, where <math>V</math> is the <math>n</math>-by-<math>n</math> diagonal matrix whose <math>j</math>th diagonal element is <math>(S(j))^2</math>, where <math>S</math> is a vector of scaling factors for each dimension.</p> $d_{st}^2 = (x_s - x_t) V^{-1} (x_s - x_t) \quad (3.17)$ <p>Where:  <math>V</math> is the <math>n</math>-by-<math>n</math> diagonal matrix whose <math>j</math>th diagonal element is <math>(S(j))^2</math>, <math>S</math> is a vector of scaling factors for each dimension.</p>

Metric	Description
<b>Mahalanobis</b>	<p>Mahalanobis distance is an effective multivariate distance metric that measures the distance between a point and a distribution. It is an extremely useful metric having, excellent applications in multivariate anomaly detection, classification on highly imbalanced datasets and one-class classification. Using the sample covariance of X, where C is the covariance matrix.</p> $d_{st}^2 = (x_s - x_t)C^{-1}(x_s - x_t) \quad (3.18)$ <p>Where: C is the covariance matrix.</p>
<b>Cityblock</b>	<p>The city block distance is a special case of the Minkowski distance, where p = 1.</p> $d_{st} = \sum_{j=1}^n  x_{sj} - x_{tj}  \quad (3.19)$

### 3.8 Pseudocode of Improved BIRCH Algorithm

Algorithm 4 shows the Pseudocode of the proposed improved BIRCH Algorithm including all steps adopted and described in the previous sections.

---

#### Algorithm 4. Improved BIRCH

---

**Input:** The dataset matrix X of size  $s$  and  $m$  features, maximum diameter (or radius) of a cluster R, and the branching factor B

**Output:** Two dissimilar clusters

**Constraints:** Arbitrary linkage methods and similarity distances

**Begin**

**4.1** Select  $m$  features, input X of size  $s$  matrix.

**4.2** Threshold initialization (**Algorithm 1. Algorithm 2. Algorithm 3**).

**4.3** Rescale data features as data = rescale(data,'inputMin',Mn,'inputMax',Mx).

**4.4** The features matrix is initially segmented.

**4.5** Load data into memory and an initial in-memory CF-tree is constructed with one scan of the data.

**4.6** (Condense data) Rebuild the CF-tree.

**4.7** (Global clustering) Use the existing k-means clustering algorithm on CF-leaves.

**4.8** (Cluster refining) Do additional passes over the dataset and reassign data points to the closest centroid from step **4.7**.

**4.9** Connect two vertices if the distance between them is the defined threshold value.

**4.10** Compute the final CF points assigned to their corresponding clusters.

**end**

---



The BIRCH algorithm is a good robust solution in the case of diverse datasets. The worst-case time complexity of the algorithm is  $O(n)$ . The time needed for the execution of the algorithm varies linearly to the dataset size. Computation complexity of the algorithm is  $O(n)$ , where  $n$  is the number of objects.

### 3.9 Performance Evaluation Measures

The performance results of improved BIRCH must be evaluated by a set of standard metrics. The following are the metrics definition used in our study:

1. **Confusion Matrix.** The confusion matrix is exploited to evaluate the position and efficiency of disease classification and diagnosis systems. Analysis of confusion matrix in classification and diagnosis of diseases lead to four modes of positive truth, negative truth, and positive false and negative false. Table 3.5 shows the position of expressed parameters in the confusion matrix (Balayla, Jacques, 2020).

Table 3.5. Samples predictions.

Actual values	Predicted values	
	Unhealthy	Healthy
Unhealthy	TP (true positives)	FN (false negatives)
Healthy	FP (false positives)	TN (true negatives)

- TP: true positives: number of examples predicted positive that are actually positive.
- FP: false positives: number of examples predicted positive that are actually negative.
- FN: false negatives: number of examples predicted negative that are actually positive.
- TN: true negatives: number of examples predicted negative that are actually negative.

2. **Accuracy.** This percentage shows how our model is performing when predicts the testing data, and defined as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (3.20)$$

3. **Recall.** Out of all the positive classes, how much we predicted correctly. It should be high as possible(the recall is calculated to know the percentage of how many true positive are out of actual positives in the data), and is defined as follows:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3.21)$$

4. **Precision.** Out of all the positive class we have predicted correctly, how many are actually positive (deals with how accurate our model can predict out of those positive prediction).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{Fp}} \quad (3.22)$$

5. **F\_measure (The Fowlkes-Mallows score).** It is difficult to compare two models with low precision and high recall or vice versa. Thus, to make them comparable, we use F-Score. F-score helps to measure Recall and Precision at the same time. It uses harmonic mean in place of arithmetic mean by punishing the extreme values more, and it defined as follows:

$$\text{F\_measure} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Precision} + \text{Recall}} \quad (3.23)$$

6. **FM-index.** The Fowlkes–Mallows index is an external evaluation method that is used to determine the dissimilarity between the resulting clusters, and defined as follows:

$$F_m = \sqrt{\frac{TP}{TP + FP} * \frac{TP}{TP + FP}} \quad (3.24)$$

---

## Chapter Four: Experimental Results and Discussion

---

### 4.1 Introduction

In this chapter, the experimental setups and requirements are presented, and the algorithm implementation with all results are discussed. Section 4.2 describes the tools used to implement the algorithm. Section 4.3 presents the results obtained by the basic BIRCH algorithm with manual threshold initialization values and with different distance metrics. Section 4.4 illustrates the implementation of the proposed improvements to BIRCH and the experiments carried out. Section 4.5 analyses the experimental results with relevant comparisons with the most related works.

### 4.2 Experiments Setup

We have implemented our algorithms in MATLAB 9.6 (R2019a). Test environment and all of our experiments are performed on a computer with an Intel (R) core (TM) i7 processor and 8 GB of memory, running on Windows 7 enterprise edition. To test the accuracy and efficiency of the BIRCH algorithm, we compared BIRCH with the improved BIRCH algorithm.

### 4.3 Results of Basic BIRCH Algorithm

In this section, we analyze the obtained results using the baseline BIRCH algorithm with different threshold values (0.2, 0.5, and 0.9) assigned manually to give multiple

thresholds within the range [0,1]. This procedure involves examining various linkage methods and distance metrics.

### 4.3.1 Data set Breast Cancer Wisconsin

Table 4.1. Clustering results on Breast Cancer Wisconsin executed by linkage (Ward) and different distances metrics.

Linkage (Ward) and Distance (Seuclidean)								
TIME	Recall	TP	TN	FP	FN	Fm	Accuracy	Threshold
0.20000 S	0.9282	0.5085	0.4251	0.0274	0.0390	0.9336	0.9283	T=0.2
0.6002 S	0.9329	0.5108	0.4175	0.0350	0.0367	0.9344	0.9283	T=0.5
0.3231 S	0.9256	0.5068	0.4241	0.0283	0.0407	0.9363	<b>0.9309</b>	T=0.9
Linkage (Ward) and Distance (Euclidean)								
TIME	Recall	TP	TN	FP	FN	Fm	Accuracy	Threshold
0.1277 S	0.9264	0.5072	0.4290	0.0234	0.0403	0.9410	<b>0.9363</b>	T=0.2
0.1163 S	0.9451	0.5051	0.4232	0.0293	0.0425	0.9338	0.9283	T=0.5
0.8826 S	0.9279	0.5081	0.3467	0.0558	0.0395	0.9144	0.9047	T=0.9
Linkage (Ward) and Distance (Squaredeuclidean)								
TIME	Recall	TP	TN	FP	FN	Fm	Accuracy	Threshold
0.0954 S	0.9191	0.5032	0.3811	0.0713	0.0443	0.8972	<b>0.8844</b>	T=0.2
0.0810 S	0.9191	0.5032	0.3811	0.0713	0.0443	0.8972	0.8844	T=0.5
0.9997 S	0.9191	0.5032	0.3811	0.0713	0.0443	0.8972	0.8844	T=0.9

Table 4.2. Clustering results on Breast Cancer Wisconsin executed by linkage (Centroid) and different distances metrics.

Linkage (Centroid) and Distance (Euclidean)								
TIME	Recall	TP	TN	FP	FN	Fm	Accuracy	Threshold
0.3798 S	0.9982	0.5466	0.0019	0.4506	0.0009	0.7397	<b>0.5484</b>	T=0.2
0.1603 S	0.9982	0.5466	0.0019	0.4506	0.0009	0.7397	0.5484	T=0.5
0.0092 S	0.9982	0.5466	0.0019	0.4506	0.0009	0.7397	0.5484	T=0.9
Linkage (Centroid) and Distance (Squaredeuclidean)								
TIME	Recall	TP	TN	FP	FN	Fm	Accuracy	Threshold
0.9767 S	0.9319	0.4049	0.4049	0.0476	0.0373	0.9232	<b>0.9151</b>	T=0.2
0.9065 S	0.4232	0.3941	0.3941	0.8965	0.0420	0.9098	0.8996	T=0.5
0.8464 S	0.9279	0.3967	0.3967	0.0558	0.0395	0.9144	0.9047	T=0.9
Linkage (Centroid) and Distance (Seuclidean)								
TIME	Recall	TP	TN	FP	FN	Fm	Accuracy	Threshold
0.1077 S	0.4298	0.5091	0.4192	0.0333	0.0383	0.9342	<b>0.9283</b>	T=0.2
0.0238 S	0.9982	0.5466	0.0019	0.4506	0.0009	0.7397	0.5484	T=0.5
0.0116 S	0.9982	0.5466	0.0019	0.4506	0.0009	0.7397	0.5484	T=0.9

Table 4.3. Clustering results on Breast Cancer Wisconsin executed by linkage (Average) and different distances metrics.

<b>Linkage (Average) and Distance (Euclidean)</b>								
TIME	Recall	TP	TN	FP	FN	Fm	Accuracy	Threshold
0.5770 S	0.9197	0.5163	0.0019	0.0666	0.0440	0.9012	<b>0.8894</b>	T=0.2
0.0625 S	0.9982	0.5466	0.0019	0.4506	0.0009	0.7397	0.5484	T=0.5
0.0015 S	0.9964	0.5456	0.0019	0.4487	0.0020	0.7391	0.5493	T=0.9
<b>Linkage (Average) and Distance (Squareeuclidean)</b>								
TIME	Recall	TP	TN	FP	FN	Fm	Accuracy	Threshold
0.9440 S	0.9164	0.5018	0.3851	0.0673	0.0458	0.8989	<b>0.8869</b>	T=0.2
0.6474 S	0.9982	0.5466	0.0019	0.4506	0.0009	0.7397	0.5484	T=0.5
0.9076 S	0.9982	0.5466	0.0019	0.4506	0.0009	0.7397	0.5484	T=0.9
<b>Linkage (Average) and Distance (Seuclidean)</b>								
TIME	Recall	TP	TN	FP	FN	Fm	Accuracy	Threshold
0.2839 S	0.9460	0.5180	0.0657	0.3867	0.0296	0.7360	0.5837	T=0.2
0.0078 S	0.9447	0.5173	0.0676	0.3849	0.0303	0.7360	<b>0.5849</b>	T=0.5
0.9053 S	0.9526	0.5216	0.0563	0.3461	0.0259	0.7358	0.5779	T=0.9

Table 4.4. Clustering results on Breast Cancer Wisconsin executed by linkage (Single) and different distances metrics.

<b>Linkage (Single) and Distance (Euclidean)</b>								
TIME	Recall	TP	TN	FP	FN	Fm	Accuracy	Threshold
0.0142 S	0.9982	0.5466	0.0019	0.4506	0.0259	0.7397	<b>0.5484</b>	T=0.2
0.1854 S	0.9982	0.5466	0.0019	0.4506	0.0259	0.7397	0.5484	T=0.5
0.8840 S	0.9982	0.5466	0.0019	0.4506	0.0259	0.7397	0.5484	T=0.9
<b>Linkage (Single) and Distance (Squareeuclidean)</b>								
TIME	Recall	TP	TN	FP	FN	Fm	Accuracy	Threshold
0.0013 S	0.9982	0.5466	0.0019	0.4506	0.0259	0.7397	<b>0.5484</b>	T=0.2
0.0006 S	0.9982	0.5466	0.0019	0.4506	0.0259	0.7397	0.5484	T=0.5
0.8689 S	0.9982	0.5466	0.0019	0.4506	0.0259	0.7397	0.5484	T=0.9
<b>Linkage (Single) and Distance (Seuclidean)</b>								
TIME	Recall	TP	TN	FP	FN	Fm	Accuracy	Threshold
0.9199 S	0.9982	0.5466	0.0019	0.4506	0.0259	0.7397	<b>0.5484</b>	T=0.2
0.9388 S	0.9982	0.5466	0.0019	0.4506	0.0259	0.7397	0.5484	T=0.5
0.9216 S	0.9982	0.5466	0.0019	0.4506	0.0259	0.7397	0.5484	T=0.9

Figure 4.1 represents a preliminary picture of the best results obtained using threshold (0.2) on Breast Cancer Wisconsin and the methods linkage (Ward) and distances Metrics (Euclidean).

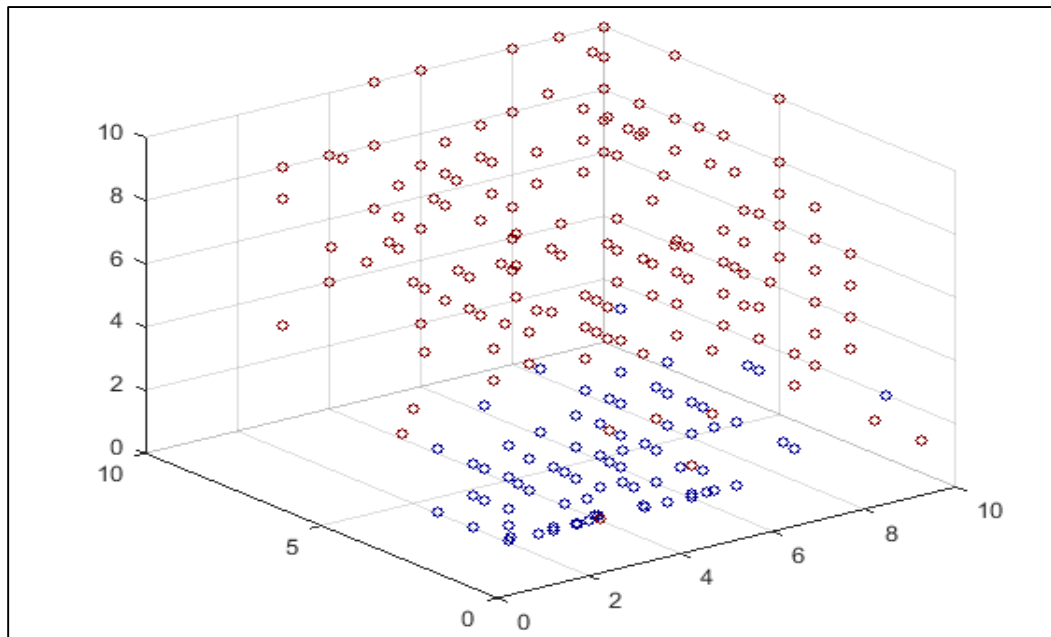


Figure 4.1. Clustering results of baseline BIRCH using threshold (0.2) on Breast Cancer Wisconsin.

It shows how the set of features are transformed and assigned into subsets so that features in the same subset are similar in some sense as shown in Figure 4.1, we have two groups, which the red dots represent the Unhealthy people, and the blue dots represent the Healthy people. The threshold value defines the class of the dataset and Euclidian distance from the central point is calculated according to that similar and dissimilar values are clustered.

We also evaluated the clusters results using the Dendrogram plot that shows the hierarchical relationship between objects. It is most commonly created as output from hierarchical clustering, as shown in Figure 4.2.

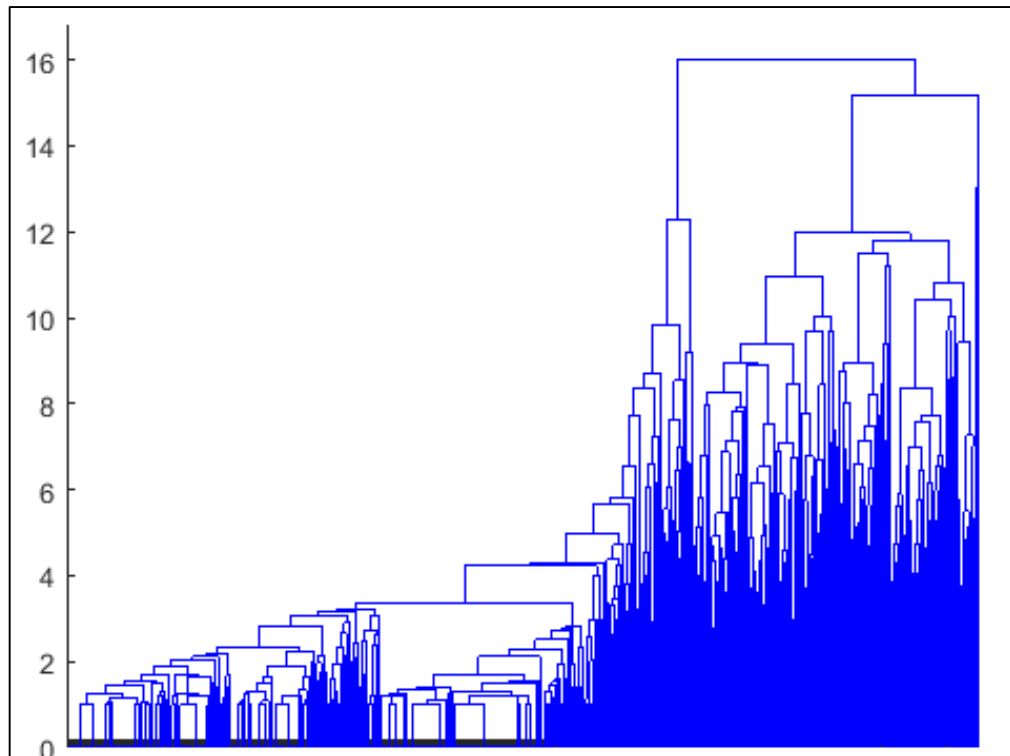


Figure 4.2. Dendrogram of baseline BIRCH using Ward linkage on Breast Cancer Wisconsin.

The Dendrogram can be interpreted as follows: At the bottom, we start with 699 records data points, each assigned to separate clusters. Two closest clusters are then merged till we have just one cluster at the top. The height in the dendrogram at which two clusters are merged represents the distance between two clusters in the data space. The highest mean and median FM scores were obtained for the BIRCH algorithm with a threshold of  $T = 0.2$ . This setting also leads to the best minimal (worst-case) FM-index. After applying with the methods linkage (Ward) and distances metric (Euclidean), we found the best results since inter-cluster distances can be defined by means of centroids.

### 4.3.2 Breast Cancer Wisconsin (Diagnostic) Data Set

Table 4.5. Clustering results on Breast Cancer Wisconsin (Diagnostic) executed by linkage (Ward) and different distances Metrics.



<b>Linkage (Ward) and Distance (Seuclidean)</b>								
TIME	Recall	TP	TN	FP	FN	Fm	Accuracy	Threshold
0.6424	0.8738	0.4645	0.1899	0.2783	0.0670	0.7392	0.6545	T=0.2
<b>Linkage (Ward) and Distance (Euclidean)</b>								
TIME	Recall	TP	TN	FP	FN	Fm	Accuracy	Threshold
0.5274	0.8738	0.4645	0.1899	0.2783	0.0670	0.7392	0.6545	T=0.2

## 4.4 Results of Improved BIRCH Algorithm

In this section, we show and analyze the obtained results using the improved BIRCH algorithm. In our proposed scheme, the threshold is initialized automatically after the features of medical records get processed.

### 4.4.1 Breast Cancer Wisconsin

Table 4.6 shows the results when different methods linkage and different distances metrics are executed.

Table 4.6. Clustering results on Breast Cancer Wisconsin executed by linkage (Ward) and different distances Metrics.

<b>methods linkage (Ward) and distances Metrics (Seuclidean)</b>										
Threshold	Accuracy	Fm	Fm-score	FN	FP	TN	TP	Precision	Recall	TIME
T=0.4771	0.9490	0.9460	0.9701	0.0313	0.0362	0.4302	0.5114	0.9917	0.9954	0.0002
<b>linkage (Ward) and distance (Euclidean)</b>										
Threshold	Accuracy	Fm	Fm-score	FN	FP	TN	TP	Precision	Recall	TIME
<b>T=0.3759</b>	<b>0.9771</b>	<b>0.9587</b>	<b>0.9772</b>	<b>0.0411</b>	<b>0.0174</b>	<b>0.4351</b>	<b>0.5201</b>	<b>0.9917</b>	<b>0.9955</b>	<b>0.0004</b>
<b>linkage (Ward) and distance (Squareeuclidean)</b>										
Threshold	Accuracy	Fm	Fm-score	FN	FP	TN	TP	Precision	Recall	TIME
T=0.4719	0.9370	0.8893	0.9380	0.0747	0.0435	0.4090	0.4729	1	1	0.0002
<b>linkage (Average) and distance (Seuclidean)</b>										
Threshold	Accuracy	Fm	Fm-score	FN	FP	TN	TP	Precision	Recall	TIME
T=0.3803	0.9699	0.9463	0.9700	0.0330	0.0253	0.4271	0.5145	0.9751	0.9866	0.0009
<b>linkage (Average) and distance (Euclidean)</b>										
Threshold	Accuracy	Fm	Fm-score	FN	FP	TN	TP	Precision	Recall	TIME
T=0.4354	0.9628	0.9342	0.9628	0.376	0.341	0.4183	0.5099	0.9672	0.9757	0.0007
<b>linkage (Average) and distance (Squareeuclidean)</b>										
Threshold	Accuracy	Fm	Fm-score	FN	FP	TN	TP	Precision	Recall	TIME
T=0.4452	0.9685	0.9436	0.9687	0.0371	0.0239	0.4285	0.5104	0.9875	0.9932	0.0005

<b>linkage (Single) and distance (Seuclidean)</b>										
Threshold	Accuracy	Fm	Fm-score	FN	FP	TN	TP	Precision	Recall	TIME
T=0.4383	0.6566	0.7396	0.6954	0.0009	0.4506	0.0019	0.5466	1	1	0.0007
<b>linkage (Single) and distance (Squaredeuclidean)</b>										
Threshold	Accuracy	Fm	Fm-score	FN	FP	TN	TP	Precision	Recall	TIME
T=0.3774	0.6566	0.7396	0.6954	0.0009	0.4506	0.0019	0.5466	1	1	0.0008
<b>linkage (Single) and distance (Euclidean)</b>										
Threshold	Accuracy	Fm	FN	FN	FP	TN	TP	Precision	Recall	TIME
T=0.4387	0.6566	0.7396	0.6954	0.0009	0.4506	0.0019	0.5466	1	1	0.0006
<b>linkage (Centroid) and distance (Seuclidean)</b>										
Threshold	Accuracy	Fm	FN	FN	FP	TN	TP	Precision	Recall	TIME
T=0.4762	0.6552	0.7355	0.6940	0.0056	0.4495	0.0030	0.5419	1	1	0.0009
<b>linkage (Centroid) and distance (Squaredeuclidean)</b>										
Threshold	Accuracy	Fm	Fm-score	FN	FP	TN	TP	Precision	Recall	TIME
T=0.4679	0.9670	0.9416	0.9671	0.0332	0.0306	0.4219	0.5144	0.9716	0.9780	0.0010
<b>linkage (Centroid) and distance (Euclidean)</b>										
Threshold	Accuracy	Fm	Fm-score	FN	FP	TN	TP	Precision	Recall	TIME
T=0.4760	0.655222	0.7396	0.6954	0.0009	0.4506	0.0019	0.5466	1	1	0.0007

From Tables 1, 2, 3, 4 and 5, it is observed that the improved BIRCH clustering algorithm performs better than BIRCH standard algorithms for detecting outliers in Breast Cancer Wisconsin and Breast Cancer Wisconsin (Diagnostic) dataset. Additionally, we found that dataset pre-processing and rescaling along the automatic threshold initialization affects the BIRCH performance in term of accuracy and time when dealing with various dataset features. Moreover, after applying various methods of linkage and similarity distances, we found that the best performing setups for BIRCH are with (Ward) linkage and (Euclidean) distance.

We compare here between the basic and improved BIRCH in term of time and accuracy measures:

- In terms of execution time, the improved BIRCH algorithm provides good results when examined with 699 records, and it takes (0.0004) seconds but the basic BIRCH algorithm takes (0.1277) seconds.
- In term of accuracy, the Basic BIRCH achieved clustering accuracy of (% 93.6) and the improved BIRCH achieved clustering accuracy of (97.7 %).

Figure 4.3 and Figure 4.4 show the clustering results of improved BIRCH on Breast Cancer Wisconsin using methods linkage (Average) and distances Metrics (Euclidean)

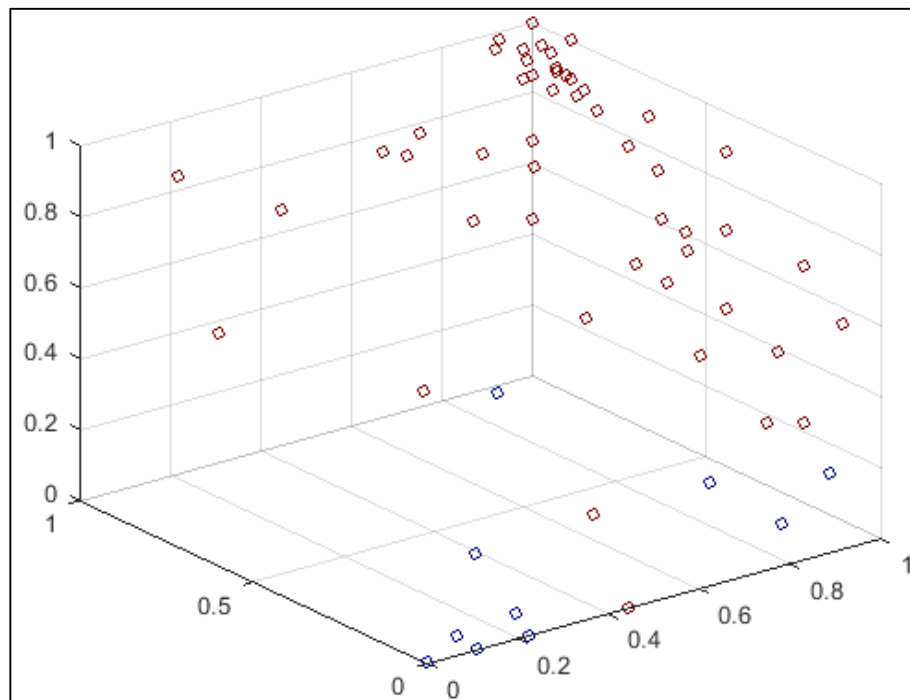


Figure 4.3. Clustering results of Improved BIRCH using threshold ( $T=0.4354$ ) on Breast Cancer Wisconsin, methods linkage (Average) and distances Metrics (Euclidean).

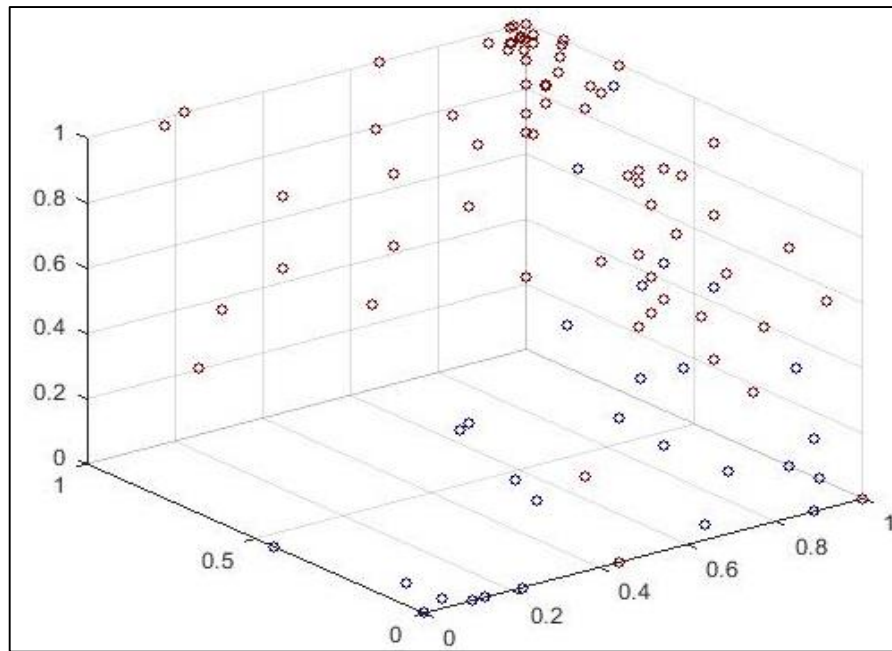


Figure 4.4. Clustering results of Improved BIRCH using threshold ( $T=0.3759$ ) on Breast Cancer Wisconsin, methods linkage (Ward) and distances Metrics (Euclidean).

As shown in Figure 4.4, applied improved BIRCH algorithm using preprocessing and automatic Threshold Initialization with linkage [Ward] and distance Metrics [Euclidean]. The final clusters are generated according to which will calculate the automatic Threshold Initialization to cluster similar and dissimilar values.

The results of hierarchical clustering can be shown using Dendrogram of improved BIRCH using ward linkage on Breast Cancer Wisconsin Figure 4.5. As observed from the figure, the method of collection is better than what shown in Figure 4.2.

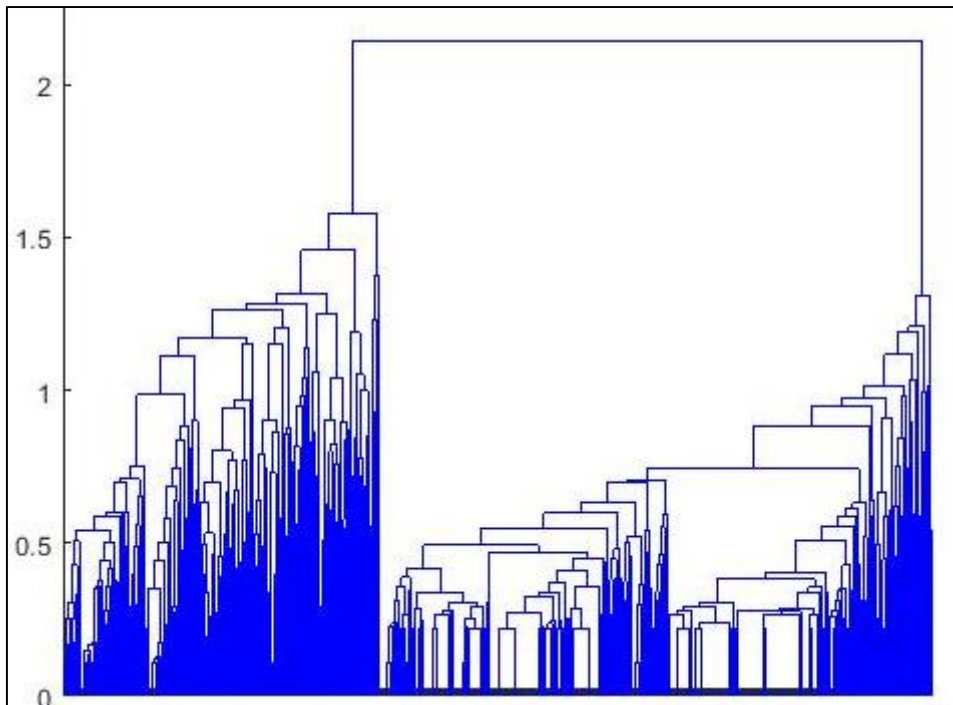


Figure 4.5 Dendrogram of Improved BIRCH using Ward linkage on Breast Cancer Wisconsin.

The highest mean and median FM scores were obtained for the BIRCH algorithm with an Automatic Threshold Initialization of  $T = 0.3759$  this setting also leads to the best minimal (worst-case) FM-index. After applying with the methods linkage (Ward) and distances Metrics (Euclidean), we found the best results since inter-cluster distances can be defined by means of Centroids.

#### 4.4.2 Breast Cancer Wisconsin (Diagnostic) Data Set.

Table 4.7 shows the performance results after executing the linkage (Ward) and distances Metrics (Euclidean, Seuclidean).

Table 4.7: Clustering results on Breast Cancer Wisconsin (Diagnostic) executed by linkage (Ward) and different distances Metrics.

<b>linkage (Ward) and distance (Euclidean)</b>										
Threshold	Accuracy	Fm	Fm-score	FN	FP	TN	TP	Precision	Recall	TIME
T=0.4751	0.9156	0.8612	0.9134	0.0544	0.1002	0.3680	0.4771	0.9767	0.9887	0.0002
<b>linkage (Ward) and distance (Seuclidean)</b>										
Threshold	Accuracy	Fm	Fm-score	FN	FP	TN	TP	Precision	Recall	TIME
<b>T=0.5607</b>	<b>0.9332</b>	<b>0.8843</b>	<b>0.9326</b>	<b>0.0544</b>	<b>0.0704</b>	<b>0.3979</b>	<b>0.4771</b>	<b>0.9438</b>	<b>0.9691</b>	<b>0.0008</b>

After we tried the steps that we took on the improved Birch algorithm, we did the experiment on data (Breast Cancer Wisconsin (Diagnostic) ) as shown in a Table 4.7, and we can say that it can be applied to a different data and it could give better results than the basic BIRCH.

## 4.5 Comparisons

In this section, we will discuss the latest data-related findings (Wisconsin breast cancer), especially the clustering topic, as described in the following table 4.8:

Title paper	Algorithm	Description	Accuracy	Precision	Recall
<b>Vijayarani al. (2013)</b>	<b>Birch with K-Means</b>	They analyzed the clustering and outlier performance of Birch with Clarans and Birch with K-Means clustering algorithm for detecting outliers.	70.38%	74.84%	76.88%
<b>Vijayarani al. (2013)</b>	<b>Birch with Clarans</b>	-	76.39%	76%	76%
<b>Lavanya al. (2016)</b>	<b>Mwmote with Birch</b>	They suggested varied distribution of data samples among different classes, Based on Majority Weighted Minority Oversampling Technique (MWMOTE) most of the samples get grouped under some classes and rest of the samples belong to the remaining classes	96.87%	94%	97%
<b>Improved BIRCH (ours)</b>	<b>Birch with features rescaling and automatic threshold initialization</b>	Linkage: Ward Distance: Euclidean Rescaled features. Automatic threshold.	<b>97.7%</b>	<b>99.5%</b>	<b>99.1%</b>

---

## Chapter Five: Conclusions and Future Work

---

### 5.1 Conclusions

Data clustering is one type of grouping method for specific objects in such a way that the similarity between groups is the minimum and the similarity within the block is the maximum. In this research we have improved the hierarchical aggregation algorithm (BIRCH) in extracting data for medical data sets by conducting many experiments until Reaching the best ranking results, there is a very clear difference between the standard BIRCH algorithm and the BIRCH algorithm in many aspects that were introduced to the algorithm where we added: features selection, data rescaling, automatic threshold initialization, and different linkage methods and distances metrics.

We found that the effects of pre-processing dataset and redefining its data points affect BIRCH performance, and the automatic threshold configuration improves the accuracy of basic BIRH algorithm in dealing with various dataset features. In addition, changing the binding methods used in BIRCH can affect the complexity of Subgroups of the tree. We achieved an improvement on the clustering accuracy by (97%) with a much better cluster quality compared to the standard BIRCH algorithm.

### 5.2 Future Work

The proposed algorithms provided on the baseline BIRCH could be further investigated and improved in the future as follows:

- 1- The resulting clusters of improved BIRCH algorithm can be passed to another clustering algorithm such as k-means to improve the accuracy by feeding it with

the previous centroids. This procedure could combine the clustering algorithm in a sequential or parallel order.

- 2- More medical datasets could be used especially Covid-19 dataset once the features of this new emerging disease are explored.



## REFERENCES

- Han, J., & Kamber, M. (2006). *Data mining: Concepts and techniques* (2ed Ed.). Beijing: China Machine Press.
- Jackson, J. (2002). Data mining: A conceptual overview. *Communications of the Association for Information Systems*, 8, 267-296.
- Tsai, C., Wu, H., & Tsai, C (2002). A new data clustering approach for data mining in large databases. *Proceedings of the International Symposium on Parallel Architectures, Algorithms and Networks –IEEE, Makati City, Philippines, 22-24 May 2002*, pp. 278-283.
- Sajana, T., Rani, C.M.S., & Narayana, V. (2016). A survey on clustering techniques for big data mining. *Indian Journal of Science and Technology*, 9(3), 1-12.
- Obermeyer Z., Emanuel E.J. (2016). Predicting the future-Big data, machine learning, and clinical medicine. *N. Engl. J. Med.* 2016.375:1216–1219. doi: 10.1056/NEJMp1606181.
- Zhang, T., Ramakrishnan, R., & Linvy, M. (1996). BIRCH: An efficient data clustering method for very large databases. In: *Proceedings of ACM SIGMOD International Conference on Management of Data*. Montreal, Quebec, Canada, 4-6 June 1996, pp. 103-114.
- Dua, D. and Graff, C. (2019). *UCI Machine Learning Repository* [<http://archive.ics.uci.edu/ml>]. Irvine, CA: University of California, School of Information and Computer Science.
- Bhardwaj, S. (2017). Data mining clustering techniques – A review. *International Journal of Computer Science and Mobile Computing*, 6(5), 183-186.
- Berkhin, P. (2006). A survey of clustering data mining techniques. In: Kogan J., Nicholas C., Teboulle M. (eds) *Grouping Multidimensional Data*. Berlin: Springer.
- Abikoye, O., Oladeji, O., & Aro, O., Taye, O.A. (2018). Text Classification using data mining techniques: A review. *Information Systems Education Journal*, 22, 1-8.
- Chayadevi, M.L., & Raju, G.T. (2012). Data mining, classification and clustering with morphological features of microbes. *International Journal of Computer Applications*, 52(4), 1-5.
- Pagudpud, M.V., Palaoag, T.T., & Padirayon, L.M. (2018). Mining the national career assessment examination result using clustering algorithm. *IOP Conference Series: Materials Science and Engineering*, 3, 1-6.
- Madhumitha, G., & Kathiresan, K. (2018). A survey on clustering techniques in data mining. *International Journal of Computer Science and Mobile Computing*, 7(8), 192-195.

Ismael, N., Alzaalan, M., & Ashour, W. (2014). Improved multi threshold BIRCH clustering algorithm. *International Journal of Artificial Intelligence and Applications for Smart Devices*, 2(1), 1-10.

Owen, R.K., Cooper, N.J., Quinn, T.J., Lees, R., & Sutton, A.J. (2018). Network meta-analysis of diagnostic test accuracy studies identifies and ranks the optimal diagnostic tests and thresholds for health care policy and decision-making. *Journal of Clinical Epidemiology*, 99, 64-74.

Mitra, S., & Nandy, J. (2011). KDD Clus: A simple method for multi-density clustering. In: *Proceedings of International Workshop on Soft Computing Applications and Knowledge Discovery (SCAKD'2011)*, Moscow (pp. 72–76).

Suh, S.C. (2011). *Particular applications of data mining*. Massachusetts, USA: Jones & Bartlett Learning.

Lorbeer, B., & Kosareva, A., Deva, B., Softić, D., Ruppel, P., & Küpper, A. (2017). A-BIRCH: Automatic threshold estimation for the BIRCH clustering algorithm. *Advances in Big Data: Proceedings of the 2nd INNS Conference on Big Data*, October 23-25, 2016, Thessaloniki, Greece (pp.169-178).

Vijayarani.S. Dr., & MsJothi.P .Dr. (2013). An Efficient Clustering Algorithm for Outlier. Detection in Data Streams. *International Journal of Advanced Research in Computer and Communication Engineering* Vol. 2, Issue 9, September 2013.

Jahanvi Joshi,Rinal Doshi,Jigar Patel, (2014). Diagnosis of Breast Cancer using Clustering Data Mining Approach: *International Journal of Computer Applications* (0975 – 8887) Volume 101– No.10.

S. Lavanya and S. Palaniswami, (2016). Hierarchical Sampling Techniques for Imbalanced Datasets. *Asian Journal of Information Technology*, 15: 2887-2896. *Asian Journal of Information Technology*, 18: 250-260.

Gurpreet Singh. Prof, Karan Jamla. M.T, (2018). Implementation & Analysis of Clustering Techniques in Bioinformatics: Cancer Research. *International Journal of Engineering Science and Computing*, Volume 8 Issue No.7, July 2018.

O. L. Mangasarian and W. H. Wolberg, W.Nick Street (1992). Uci repository of machine learning databases. Available: <https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+%28original%29>. [Accessed 1/3/2020]

O. L. Mangasarian and W. H. Wolberg,(1995). Uci repository of machine learning databases. Available: [https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+\(Diagnostic\)](https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)). [Accessed 1/3/2020]

Dong, J., Wang, F., Yuan, B., Dong, J., Wang, F., & Yuan, B. (2013). Accelerating BIRCH for clustering large scale streaming data using CUDA dynamic parallelism. *Intelligent Data Engineering and Automated Learning*, 409-416.

MATLAB. *User's Guide*. The Math-Works, Inc., Natick, MA 01760, 1994-2020. Available: <http://www.mathworks.com/access/helpdesk/help/techdoc/matlab.shtml>. [Accessed 3/3/2020]

Balayla, Jacques (2020). "Prevalence Threshold and the Geometry of Screening Curves." arXiv preprint arXiv:2006.00398 (2020) doi: <https://arxiv.org/abs/2006.00398>.

MATLAB. *User's Guide*. The Math-Works, Inc., Natick, MA 01760, 1994-2020. Available: <https://www.mathworks.com/products/matlab/whatsnew.html>. [Accessed 1/1/2020]