



**Department of Information Technology**

**An Enhancement over BIRCH Hierarchical Clustering Algorithms  
for Better Partitioning of Medical Data**

**Prepared by**

**Ra'ed Mohammad Jamil Alnusour**

**Supervised by**

**Dr. Faiz Al Shrouf**

**This Thesis is Submitted to Faculty of Information Technology  
as a Partial Fulfillment of the Requirement for Master Degree in  
Software Engineering**

**July 2020**



**Department of Information Technology**

**An Enhancement over BIRCH Hierarchical Clustering Algorithms  
for Better Partitioning of Medical Data**

**Prepared by**

**Ra'ed Mohammad Jamil Alnusour**

**Supervised by**

**Dr. Faiz Al Shrouf**

**This Thesis is Submitted to Faculty of Information Technology  
as a Partial Fulfillment of the Requirement for Master Degree in  
Software Engineering**

**July 2020**

### Authorization Statement

Ra'ed Muhammad Jameel AL-Nsoor authorizes Isra university to provide hard and soft copies of his thesis to libraries for the institutions or individuals upon their request.

Ra'ed Muhammad Jameel AL-Nsoor

Signature



Date: 2020/7/15

## أقرار تفويض

انا رعد محمد جميل النصور افوض جامعة الاسراء للدراسات العليا بتزويد نسخ من رسالتي ورقيا و الكترونيا للمكتبات والمنظمات او الهيئات والمؤسسات المعنية بالابحاث والدراسات العليا عند طلبها.

رعد محمد جميل النصور

التوقيع 

التاريخ: 2020/7/15

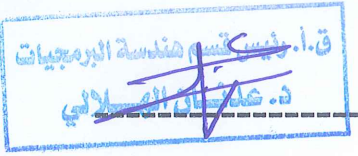
The undersigned have examined the thesis entitled **An Enhancement over BIRCH Hierarchical Clustering Algorithms for Better Partitioning of Medical Data**, presented by **RA'ED MUHAMMAD JAMEEL AL-NSOOR**, a candidate for the degree of master in Software Engineering and hereby certify acceptance.



د. فايز الشروف  
ق.أ. رئيس قسم هندسة البرمجيات

Dr. Faiz Al-Shrouf


Date 15.7.2020




ق.أ. رئيس قسم هندسة البرمجيات  
د. عدنان الحلالي

Dr. Adnan Al-Helali

Date 15-7-2020



د. محمد عليا



DEPARTMENT OF SCIENTIFIC RESEARCH - AL ZAYTOONAH UNIVERSITY OF JORDAN  
جامعة الزيتونة الأردنية - عمادة البحث العلمي

Prof. Dr. Mohammad Alia

Date

15-7-2020

## اهداء

أقدم هذا العمل المتواضع الى :

أبي العطوف....مصدر الفخر ورمز التضحية والصبر، انت مدرستي الاولى والاخيرة في الحياة، تفانيت في تعلمي، وربيتني على الاحترام والتواضع والعيش بكرامة وعزة نفس وشهامة

امي الحنون...الجنة التي ترعرت تحت ظلالها، لقد ضحيتي من اجل راحتنا وتفانيتي من اجل سعادتنا، انت مصنع الرجولة، انت من تعبتني في تربيتنا التربية الصالحة وانت المصدر الذي نستمد منه العطف والود والتسامح

اخي واختي الاعزاء انتم سندي الاول والاخير وعزوتي في كل سراء وضراء، انتم بسمتي ومصدر تفائلي، وجودكم بالقرب مني هو سر سعادتي، اتمنى لكم المستقبل المشرق والمزيد المزيد من التقدم والعطاء، فنجاحكم هو نجاحي

عماتي الكريمات ، انتم حجر الاساس، انتم ياقوت التضحية ولؤلؤ الصبر وزمرد التفاني، انتم مصدر الهامي الاول والاخير، انتم من انار دربي، لولا دعمكم ودعواتكم لما تمكنت من اتمام عملي بنجاح، اطال الله بعمركم وادامكم فوق رؤوسنا

عائلتي واقاربي واصدقائي....انتم مصدر المحبة والاخلاص والسعادة، احبكم جميعا وافخر بكم

اهديكم بحثي هذا واتمنى ان ينال اعجابكم،،،

## **Acknowledgments**

I would like to express my appreciation to my supervisor Dr. Faiz Al Shrouf, I am so grateful for his patient guidance; he guided me to the right path to in order to complete this work.

I am very thankful to the whole staff of faculty of Information Technology and the dean of the faculty Dr. Jamal Zraqo and the section head of department of software engineering Dr. Adnan Al-Helali, they do an exceptional work at administrating researches and supporting master students for brighter future of IT research and development.

I am always thankful to my great and lovely family...father, mother, brother and sister you are the most precious thing in my life, I would like to take advantage of this happy occasion and also introduce many thanks to my example in my professional life Prof. Dr. Ayman Al-Nsoor, his continuous encouragement is the main reason why I continue in pursuing a higher education in software engineering.

## Abstract

Over the years, technology has revolutionized our world and daily lives, information is getting to be more accessible and shared to the public users, big data across the web are being collected and saved in all forms from texts to different media files, machine learning algorithms are utilizing these data to learn more about it which in response, could improve these algorithms to be more useful and applicable in the real world, Clustering algorithms are unsupervised machine learning algorithms that can be used in many fields including pattern recognition and image analysis, There are many clustering algorithms such as K-means and Agglomerative Hierarchical Clustering (AHC), however they work fine in specific data sets.

Clustering algorithms can be used to cluster medical data to find an undiscovered pattern which in result improves the medical field's knowledge about patients and different diseases, This thesis will focus on one of the most dangerous diseases cancer, SEER databases provides a big amount of data from the year of 1973 until now about cancer patients from various locations and sources throughout the United States, to find useful patterns through these data a good clustering algorithm is needed to cluster such big data, BIRCH is one of the most effective clustering algorithms on big data.

This thesis investigates the development of new technologies to propose the MD-BIRCH algorithm which is an enhanced version of BIRCH algorithm by implementing Manhattan distance over multiple phases of BIRCH algorithm from early stages of compacting data points into an initial Clustering Feature (CF) tree to the middle stages while descending the tree into more depth to the late stages of removing the outliers and performing global clustering on the whole tree by another modified clustering algorithm based on Manhattan distance.

The experiments have been conducted on SEER medical dataset over multiple clustering iterations, where each BIRCH and MD-BIRCH has been executed 8 times over cancer patients big data sample, the results showed that the MD-BIRCH algorithm has outperformed BIRCH algorithm in terms of quality and has a slightly an enhanced performance. This work has been implemented by Python 3.7 programming language.

**Keywords:** Machine Learning, Big Data, Knowledge Discovery, Data Mining, Data Clustering, BIRCH Algorithm.



# Table of Contents

<b>1. Chapter One: Introduction .....</b>	<b>1</b>
1.1 Introduction .....	1
1.2 Problem Statement .....	2
1.3 Objectives of the Study .....	4
1.4 Limitations of the Study .....	4
1.5 Assumptions .....	5
1.6 Significance of Study .....	6
1.7 Thesis Organization .....	6
<b>2. Chapter Two: Literature Review .....</b>	<b>7</b>
2.1 Introduction .....	7
2.2 Data Mining of Medical Data: Clustering .....	8
2.3 Density-based Clustering Algorithms .....	9
2.4 The Partitioning Clustering Method .....	9
2.5 The Hierarchical Clustering Method .....	10
2.6 The BIRCH Hierarchical Clustering Algorithm .....	10
2.7 Related Works .....	11
<b>3. Chapter Three: Methodology .....</b>	<b>20</b>
3.1 Introduction .....	20
3.2 Manhattan Distance over K-Means Clustering Algorithm .....	21

3.3 Proposed Methodology .....	22
<b>4. Chapter Four: Design, Analysis &amp; Implementation .....</b>	<b>26</b>
4.1 SEER Dataset .....	26
4.2 Experiment .....	34
<b>5. Chapter Five: Results.....</b>	<b>37</b>
<b>6. Chapter Six: Conclusion &amp; Future Work.....</b>	<b>43</b>
6.1 Conclusion .....	43
6.2 Future Work .....	44
<b>7. References .....</b>	<b>46</b>
<b>8. Appendix .....</b>	<b>48</b>

## List of Figures

1.1 Initial CF tree built by BIRCH algorithm [3] .....	3
1.2 Basic Steps in the Knowledge Discovery in Databases Process [4] .....	5
2.1 Presenting Clustering Algorithms for Data Mining [8] .....	9
3.1 MD-BIRCH Clustering Algorithm .....	24
4.1 SEER*Stat 8.3.6 is used as the main tool to provide cancer statistics .....	27
4.2 SEER Research Data .....	28
4.3 Case Selections of Variables .....	29
4.4 Columns Selection .....	30
5.1 Cluster results of Standard BIRCH vs. MD-BIRCH quality measured by validation through SBdw index over multiple clustering iterations from 2 to 9 clusters.....	37

## List of Tables

2.1 Comparison between Single Threshold and Multiple Thresholds in the BIRCH Algorithm	16
2.2 Summery of Related Work .....	17
5.1 SDbw cluster validation score for Standard BIRCH and MD-BIRCH clustering algorithms under certain number of clusters (lower value > better quality) .....	40
5.2 Summery quality table of execution of Standard BIRCH and MD-BIRCH clustering algorithms on number of clustering iterations from 2 to 9 clusters .....	41
5.3 Time of execution of Standard BIRCH and MD-BIRCH clustering algorithms under certain number of clusters .....	41
5.4 Summery performance table of execution of Standard BIRCH and MD-BIRCH clustering algorithms on number of clustering iterations from 2 to 9 clusters .....	42

## List of Symbols

B	Max no.CF in a leaf node
L	Max no.CF in a non-leaf node
N	No. Cluster's data points
LS	Linear sum of N- data points
SS	Squared sum of N- data points
$D(i,j)$	Distance between objects i and j
T	The Threshold value
C	Number of desired Clusters

## **List of Abbreviations**

KDD: Knowledge Discovery in Databases

BIRCH: Balanced Iterative Reducing and Clustering using Hierarchies

NIH: National Cancer Institute

SEER: Surveillance, Epidemiology, and End Results

# Chapter One

## Introduction

### 1.1 Introduction

Nowadays, with large and great expansion of the Internet, world technology, and users of system, we need systems that help us to organize, analyze, and arrange the data in a way that helps us to use them better. The data mining system is one of these systems. The concept of data mining (sometimes called knowledge discovery) refers to extracting (mining) important information from large amount of data<sup>i</sup>. In other words, we can say that the data mining systems involve a lot of technology and many algorithms that help us to extract great variety of information that are either stored in large database or other information repositories. It allows users to categorize data from many dimensions.

Data mining is part of a large process called ‘knowledge discovery of data’, which means transforming raw data stored in data warehouse into meaningful pattern<sup>ii</sup>. This process consists of some steps that start with data cleaning (for removal of noise); data integration, which is a process whereby multiple data source are combined; data selection (keeping only the data that are relevant for the analysis task); data transformation; data mining (the process of extraction of meaning from data); pattern evaluation (identification of the patterns of interest); and knowledge presentation (knowledge representation technology is used to present knowledge to users). The process of Knowledge Discovery in Databases (KDD) is introduced in Section 2.

Data mining is used in many important areas in our life, including educational and commercial areas. It is also used for solving business and scientific problems. As well, data mining is used for

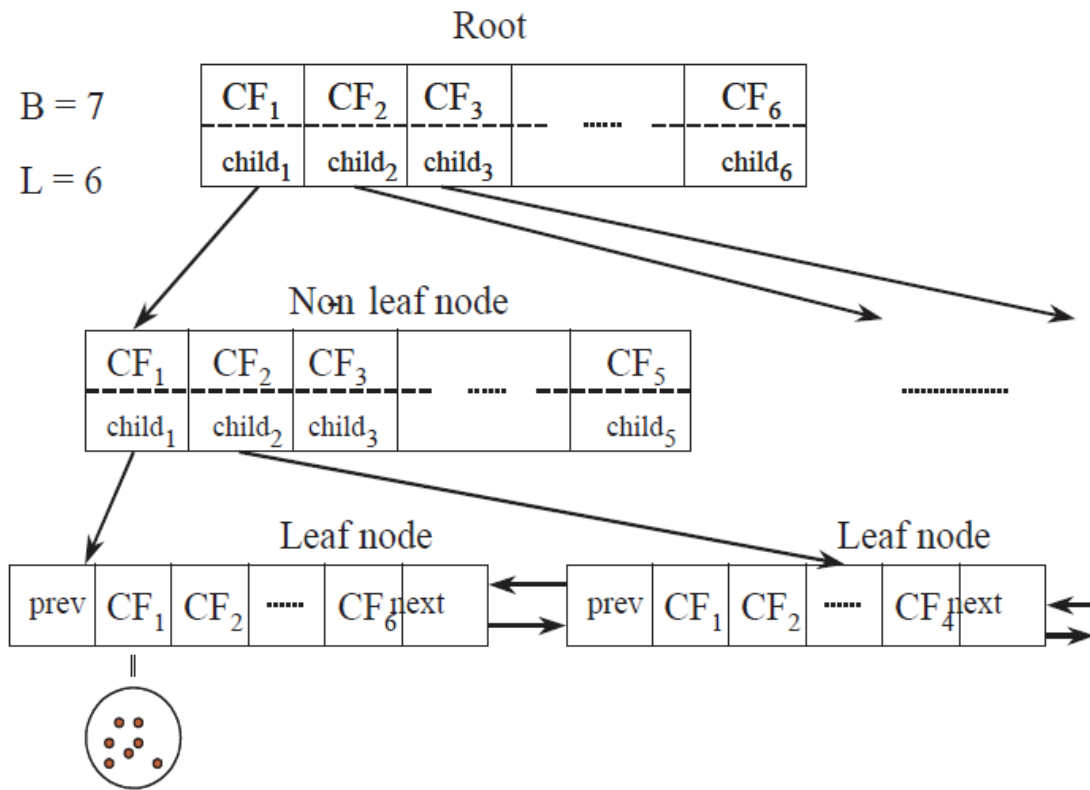
processing big data, which means large amount of complex set of data that is difficult to process by traditional data-processing applications. Data mining consists of three major processes: classification, predication, and clustering.

## **1.2 Problem Statement**

Enhancing or improving the Balanced Iterative Reducing and clustering using Hierarchies algorithm in medical applications may increase the accuracy as well as the performance of the disease diagnosis operation. In consequence, the medical services presented to patients are improved. The hierarchical clustering methods have many advantages, including that they can specify and determine the reasons for some health defects of humans and, thus, enhance the medical diagnosis operations and reduce the processing effort and time.

Balanced Iterative Reducing and Clustering using Hierarchies briefly know as BIRCH is one of the most effective algorithm when the data are too heavy for the memory to handle, medical data such as cancer data are usually known to be too massive, and when we apply usual clustering algorithms such as K-Means algorithm on these type of data, it tend to be on the end too hard to handle because of the lack of the efficiency since K-Means algorithm need to scan all the data points in advance in order to be able to progress its first steps of clustering, while BIRCH algorithm starts by building an initial CF tree as shown in figure 1.1 and if the computer memory reached it maximum size then the threshold will be increased and will rebuild a smaller CF tree (Zhang&Linvy, ,1996, 103-114)<sup>iii</sup>.





**Figure1.1 Initial CF tree built by BIRCH algorithm [3]**

Clustering Feature (CF)

CF= (N, LS, SS)

Additivity Theorem

If Cluster Feature (1) and Cluster feature (2) are two disjoint sub-cluster then by applying additivity theorem – merging CF1 and CF2 like the following:

$$\text{Cluster Feature (1) + Cluster Feature (2) = (N1+N2, LS1+LS2, SS1+SS2)}$$

Clustering feature (CF) can be considered as a compressed storage of data points in a cluster, CF is calculated while the algorithm is running and its update itself dynamically while scanning the

data as well as the birch algorithm can create a good clustering with only its first initial scan unlike the K-Means which need multiple iterations to perform a rational cluster after updating its first initial random centroids multiple times until reaching the convergence.

### **1.3 Objectives of the Study**

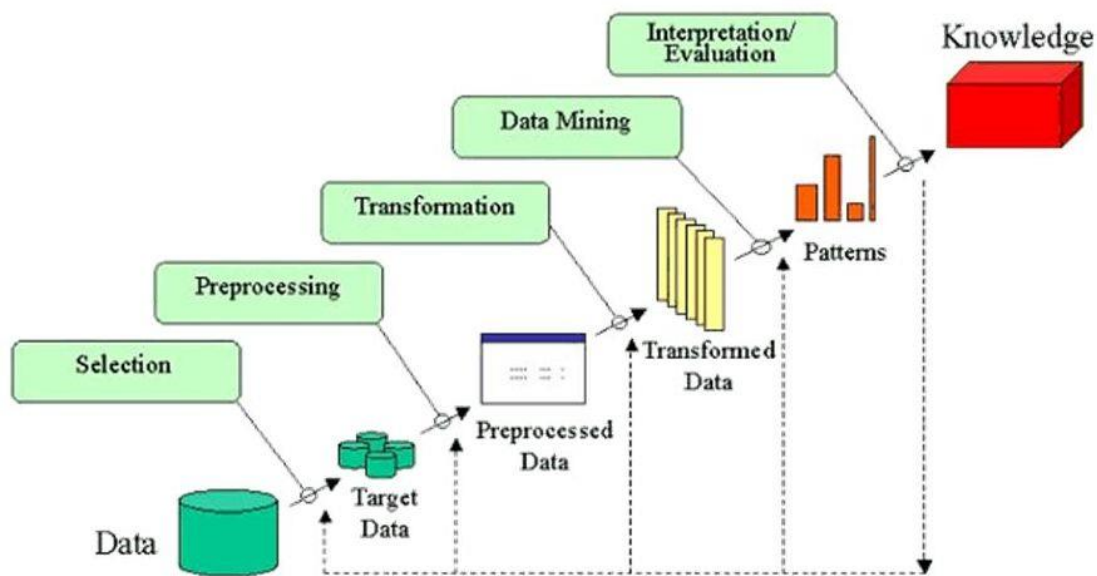
This research aims at achieving enhancement in the BIRCH hierarchical clustering algorithm in data mining for the medical datasets by running many experiments until reaching to the best classification results.

### **1.4 Limitations of the Study**

The acronym SEER refers to Surveillance, Epidemiology, and End Results. It is related to a program that provides information about, and insights on cancer diagnosis and characteristics. The SEER-Medicare database links SEER data with Medicare files of the patients, which results in adding new information collected by Medicare like other diagnosis. It is source for cancer statistics in USA. The benefit of linking two data sources is production of a population focused source of information which will be used for setoff health care services and epidemiological research. In fact, SEER data files are too complex and large. We can create clusters of them using the BIRCH algorithm because this algorithm works for huge databases.

## 1.5 Assumptions

In this study, the research method is based on KDD and the BIRCH algorithm applied to the SEER medical dataset [4]<sup>iv</sup>. The KDD process comprises five essential steps shown in figure 1.2: data selection, pre-processing, transformation, data mining, and generation of knowledge. Further details on each of these steps are provided in the following paragraphs.



**Figure 1.2 Basic Steps in the Knowledge Discovery in Databases Process** (National Cancer Institute, Surveillance, epidemiology and end results (seer) program ([www.seer.cancer.gov](http://www.seer.cancer.gov)) SEER Research data file(1975-2017).

The following assumptions in KDD process are given below:

- 1. Data selection.**
- 2. Data cleaning and pre-processing.**
- 3. Data transformation.**
- 4. Data mining.**

## **5. Interpretation and Evaluation.**

## **6. Knowledge generation and use.**

### **1.6 Significance of Study**

Searching inside medical data, whether they are records or images, is a challenge to the traditional information search techniques. The present research will enhance the BIRCH algorithm for data mining for the medical sector, which will help in distribution of patients to groups so as to provide the best services for them and improve the work quality.

### **1.7 Thesis Organization:**

This thesis contains six chapters: first chapter describes data mining as a base of this study; in the second chapter a literature survey is introduced; the third chapter presents research methodology, the fourth chapter show the design, analysis and implementation phase, the results is shown in the fifth chapter and finally the conclusion and future works are presented in the sixth chapter.

## Chapter Two

### Literature Review

#### 2.1 Introduction:

Data mining classification refers to search for function that describes the data classes. In other words, it is a process that is used to group some items based on some key characteristics such as similarity(Andritsos, ,2002)<sup>v</sup>.

For example, classification is used when an insurance officer needs to analyze information about customers to know which insurance applicants are safe and who one not safe. There are several types of algorithms and techniques that can be employed in a classification process such as classification by decision tree inductions, Bayesian classifications, and back propagation.

Data predication is similar to data classification. It implies identification of data points purely on the description of another related data value<sup>vi</sup> (National Cancer Institute, Surveillance, epidemiology and end results (seer) program ([www.seer.cancer.gov](http://www.seer.cancer.gov)) SEER Research data file (1975-2017)).

That is, prediction models predict continuous value functions. It is used to find a numerical output. For instance, prediction models are used by marketing managers to predict how much a specific customer or group of customers will spend during sales. There are some popular methods for prediction like liner regression analysis and non-liner regression analysis. In general, both classification and predication are types of data analysis that are used in data mining.

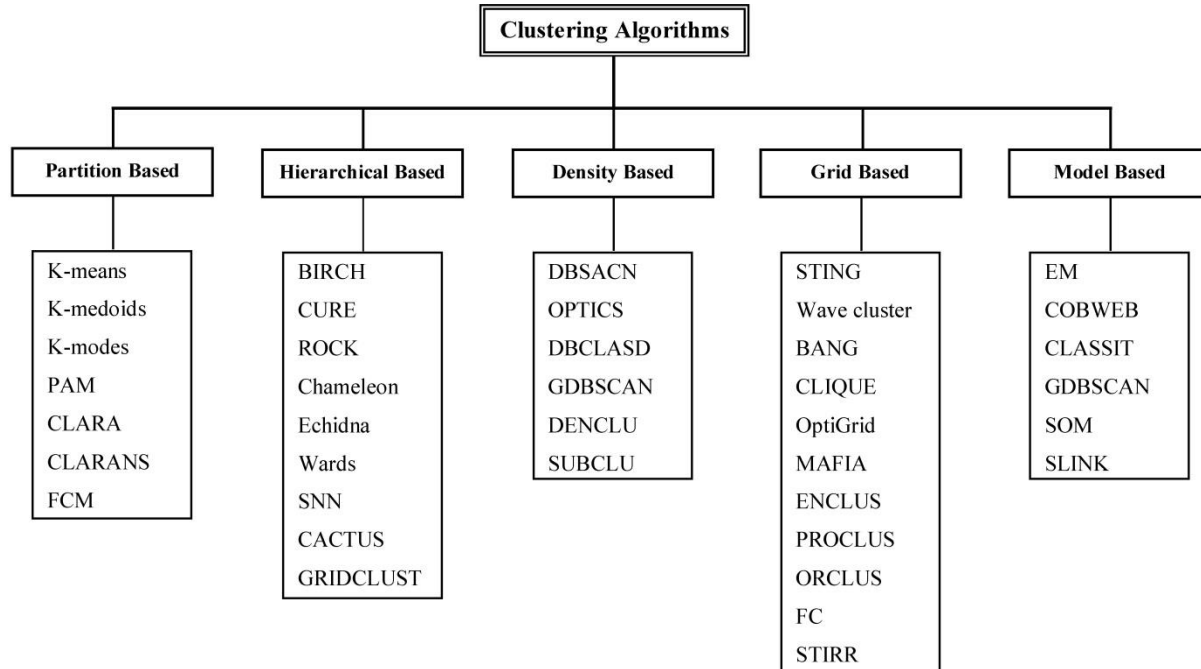
The third main data-mining process is clustering, which is the process of grouping a set of similar objects together. Accordingly, the cluster is a collection of data objects that are similar to each other and dissimilar with objects of other clusters<sup>vii</sup> (Chayadevi & Raju, 2012, 1-5).

This implies that the objects in the same group (i.e., cluster) are more similar to each other than to the objects in other groups (clusters). This means that the function of clustering is to group the similar groups of entities (objects) together. These objects meet one of two conditions; either the objects in a group are very similar or the groups are different from each other.

## **2.2 Data Mining of Medical Data: Clustering**

The key challenge in data mining is to extract meaningful information, that is, patterns, from big datasets, especially in the field of medical data. Extraction of knowledge from medical data is sometimes a great challenge in data mining. Though, the medical data are considered as interesting data and need to be followed up.

Clustering is the main task in data mining. So far, there are many clustering algorithms that have been collectively categorized into five major groups which are the hierarchical, partitioning, density-based, grid-based, and model-based algorithms<sup>viii</sup> (Sajana & Narayana, , 2016, 1-12) . This study will focus on hierarchical clustering but first there is a need for explaining some clustering algorithms. A graphical representation of the foregoing major clustering algorithm groups and the algorithms categorized within each major group of them is given in Figure 2.1<sup>ix</sup>( Bhardwaj,2017, 183-186).



**Figure 2.1: Presenting Clustering Algorithms for Data Mining [8].**

### 2.3 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. The arbitrarily-shaped clusters are formed by the various density-based clustering algorithms like the DBSCAN, OPTICS, DBCLASD, GDBSCAN, DENCLU, and SUBCLU algorithms.

### 2.4 The Partitioning Clustering Method

The various partitioning procedures commonly result in a group of (M) clusters. Ideally, each item belongs to a unique cluster. Each cluster may be denoted by a centroid or a cluster representative, which is some sort of summary description of all the entities enclosed within a cluster.

## **2.5 The Hierarchical Clustering Method**

Hierarchical clustering works by grouping data objects into a tree cluster and every cluster node contains child clusters. This approach allows for exploring data at different levels of granularity. The hierarchical clustering algorithms build clusters gradually. There are two approaches to hierarchical clustering: hierarchical clustering (bottom-up) and divisive hierarchical clustering (top-down). Agglomerative clustering (hierarchical or bottom-up clustering) starts by merging each object in one cluster. After that, these objects (clusters) are merged into large clusters. This process is then repeated till all the clusters are merged into one cluster, which is the top level of the hierarchical shape. In divisive hierarchical clustering (top-down clustering), however, we start with all objects in one cluster and, 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.

## **2.6 The BIRCH Hierarchical Clustering Algorithm**

The balanced iterative reducing and clustering using hierarchies (BIRCH) algorithm is a hierarchical clustering algorithm used with very large data sets. It also has the ability to cluster multi-dimensional metric data points, either incrementally or dynamically. That is to say that BIRCH can produce good clustering in a single scan. It also improves the clustering quality with few scans. It is the first clustering method that could handle noise. 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 the other 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.BIRCH algorithm has some



disadvantages such as that it can work with numerical data only and that it is sensitive to the order of the data records.

BIRCH has been used to solve two real-life problems:

(i) Building an iterative and interactive pixel classification tool and (ii) generating an initial codebook for image compression<sup>x</sup>. BIRCH progresses in four phases:

Phase 1: Scanning all data then building an initial CF tree in memory by using the given amount of memory and recycling space on the disk.

Phase 2: Building a smaller CF tree.

Phase 3: Performing global clustering.

Phase 4: Refining the clusters. This step is optional and it requires additional passes over the data to refine the results.

## **2.7 Related Works:**

There are many articles, studies about use and improvements of BIRCH algorithm. This section presents a briefing on such studies and articles.

Zhang et al. <sup>xi</sup>( Zhang & Linvy, 1997) presented a paper having the title: “BIRCH: An efficient data clustering method for very large databases.” The study presents the BIRCH clustering method and demonstrates how it’s suitable for large datasets. “BIRCH incrementally and dynamically clusters incoming multi-dimensional metric data points to try to produce the best quality clustering

with the available resources (i.e., available memory and time constraints). 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”. They also presented a comparison 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 <sup>xiii</sup>( Garg & Bhatnagar,2006, 34).

Zhang et al. <sup>xiii</sup>( Zhang & Linvy, 1997) also represented a paper having the title: “BIRCH: A New Data Clustering Algorithm and Its Applications”. which proposed BIRCH algorithm and its real world applications, The performance of BIRCH algorithm, K-Means algorithm and CLARANS algorithm have been compared on the workload base, when they applied both of K-Means and CLARANS algorithms on a very large dataset it has been found that the memory can’t hold the whole dataset, which results in needing more memory than BIRCH algorithm needs. Two real world problems have been showed: 1-“Interactive and Iterative Pixel Classification” 2-“Codebook Generalization in Image Compression”, then these problems have been solved by applying BIRCH clustering algorithm, additional work such as handling non-metric data and heuristic techniques for increasing threshold in dynamic ways<sup>xiv</sup>.

Garget et al. <sup>xv</sup>Presented conference paper having the title: “PBIRCH: A Scalable Parallel Clustering algorithm for Incremental Data”. Proposed a new way in dealing with data by dividing multidimensional data into parallel processes where each processor get  $N/P$  amount of data where  $N$  is the number of data items and  $P$  is number of processors, data distributed among processors in

a cyclic way then after Multiple CF-Trees has been built parallel K-Means algorithm is used to global cluster all the high balanced CF-Trees by broadcasting to all processors using exchanged messages, an experiment has been performed with data size ranging from 5000 to 10000 with multidimensional data and found that the PBIRCH speed up the original BIRCH linearly as data increases in size<sup>xvi</sup> (Ismael & Ashour,2014,1-10)

Chayadevi and Raju published <sup>xvii</sup>presented a work titled: “Data mining, classification, and clustering with morphological features of microbes. "In this research, they discussed the old patterns used in data mining and applied in medical image processing and searching. They also discussed the need for an automated tool for fast recognition of microbes in order to examine the medical data before they expire. Digital image processing is an integral part 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 contour 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<sup>xviii</sup> (Han &Kamber, 2001).

Dong et al. <sup>xix</sup>(Dong, et. Al 2013) presented a work with the following title: “Accelerating BIRCH for Clustering Large Scale Streaming Data Using CUDA Dynamic Parallelism”. The Study introduced G-BIRCH algorithm which is an improved version of the BIRCH algorithm by using GPUs dynamic Parallelism feature in CUDA programming platform which has been

developed by NVIDIA, the proposed work featured the methodology of launching a master kernel in the beginning, then multiple slave kernels dealing with sub-group of data in GPU memory that assigned to them, each slave kernel fetches some data records from the master, when building CF-Tree each node uses a storage array and insert inside it the children CF values as the following : “parent ID, child ID and the number of children”, according to the mentioned results: “GBIRCH achieved encouraging speedups from 7 to 154 times over the original BIRCH on six benchmark datasets.”<sup>xx</sup>(Dong, et. Al 2013, 25)

Lorbeer et al.<sup>xxi</sup>(Lorbeer, et. al.2017, 169-178) presented a study that is titled: “A-BIRCH: Automatic threshold estimation for the BIRCH clustering algorithm.” In this paper, these researchers presented A-BIRCH, which is “an approach to automatic threshold estimation for the BIRCH clustering algorithm, this approach computes the optimal threshold parameter of this clustering algorithm from the data” such that BIRCH does proper clustering even without the global clustering phase that is usually the final step of this algorithm, this is possible if the data satisfies certain requirements, if those requirements are not satisfied, then A-BIRCH will issue a pertinent warning before presenting the results, this approach renders the final global clustering step of BIRCH unnecessary in many situations, which results in two advantages: first, no need to know the expected number of clusters beforehand, second, without the computationally-expensive final clustering, the fast BIRCH algorithm will become even faster for very large data sets<sup>xxii</sup>((Lorbeer, et. al.2017, 169-178).<sup>xxiii</sup>

Ramadhani et al.<sup>xxiv</sup>(Ramadhani & Suwilo,2019). Improve BIRCH algorithm for big data clustering.)proposed a work with the following title: “Improve BIRCH algorithm for big data

clustering”. The study represented a new method when improving birch algorithm, instead of the previous methods to try to improve BIRCH algorithm by modify its static threshold value into dynamic for instance, the study uses modifications of CF-Leaf value by modifying CF-Leaf formula from (N, LS, and SS) into CF-Leaf (modif) = (N, LS, SS, T), where is the addition of T parameter in order to track the changes of T value, as the methodology stated “If the leaf radius selected including the CF sub cluster exceeds the Threshold T, the system will enlarge the cluster scale. Then check again. If the radius does not exceed the new threshold value, the change in the threshold value of T will be updated and the sub cluster will enter the leaf (leaf-CF (modif)).” The modified BIRCH algorithm produces 65% less CFs than the original Birch, using silhouette coefficient to measure accuracy the Modified BIRCH averaged around 152.34% better accuracy than the standard one<sup>xxv</sup>( Ramadhani & Suwilo,2019, 19)

Ismael et al. <sup>xxvi</sup>( Ismael & Ashour, 1-10. ) presented a work carrying the title: “Improved multi threshold BIRCH clustering algorithm.” These researchers proposed a solution to the shortcomings of the BIRCH algorithm when a single threshold is used. The clustering algorithm they suggested is a clustering algorithm that is suitable for very large data sets. In the algorithm, a CF-tree is built whose all entries in each leaf node must satisfy a uniform threshold T, and the CF-tree is rebuilt at each stage using different threshold. This was achieved using multiple thresholds instead of a single threshold<sup>xxvii</sup>(Ismael & Ashour, 2014, 1-10).

Within the context of thresholds, Table 2.1 presents a comparison between multiple thresholds and single threshold in the BIRCH algorithm<sup>xxviii</sup>( Ismael & Ashour,2014,1-10):

**Table 2.1 Comparison between Single Threshold and Multiple Thresholds in the BIRCH Algorithm**

No.	Single Threshold	Multiple Thresholds
1	Used in the basic BIRCH algorithm.	Used in the modified (or advanced) BIRCH algorithm.
2	Lower performance than multiple thresholds.	Higher performance than single threshold.
3	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.
4	Only increases when the random-access memory (RAM) is full.	Does not require full RAM to increase.
5	Have an increased specificity but decreased sensitivity.	Have an increased sensitivity but decreased specificity.
6	Less accurate than multiple threshold and results in lower clustering efficiency.	More accurate than single threshold and results in higher clustering efficiency.
7	Less able than the multiple thresholds to handle data with different densities and noise.	Better able than the single threshold to handle data with different densities and noise.

**Table 2.2: Summery of Related Work**

<b>Paper</b>	<b>Year</b>	<b>Description</b>
BIRCH: An Efficient Data Clustering Method for Very Large Databases	1996	Presents the BIRCH clustering method and demonstrates how it's suitable for large datasets, a comparison of the performance of BIRCH against that of CLARANS has been performed, which is a clustering method proposed recently for large datasets, and showed that BIRCH is consistently superior to CLARANS
BIRCH: A New Data Clustering Algorithm and Its Applications	1997	Proposed BIRCH algorithm and its real world applications, The performance of BIRCH, K-Means and CLARANS have been compared, Two real world problems have been solved by BIRCH algorithm: 1- "Interactive and Iterative Pixel Classification" 2- "Codebook Generalization in Image Compression"
PBIRCH: A Scalable Parallel Clustering algorithm for Incremental Data	2006	Proposed a new way in dealing with data by dividing multidimensional data into parallel processes, an experiment with data size ranging from 5000 to 10000 with multidimensional data has been performed and found that the PBIRCH speed up the original BIRCH linearly as data increases in size.

Data mining, Classification and Clustering with Morphological features of Microbes	2012	Discussed the old patterns used in data mining and applied in medical image processing and searching, proposed the automated color image segmentation for bacterial image, 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
Accelerating BIRCH for Clustering Large Scale Streaming Data Using CUDA Dynamic Parallelism	2013	The Study introduced G-BIRCH algorithm which is an improved version of the BIRCH algorithm by using GPUs dynamic Parallelism feature, on a six benchmark datasets, GBIRCH achieved speedups from 7 to 154 times over the original BIRCH,
Improved Multi Threshold Birch Clustering Algorithm	2014	proposed a solution to the shortcomings of the BIRCH algorithm when a single threshold is used and achieved by using multiple thresholds instead of a single threshold, After tested on two real data sets – Statlog Data Set and Abalone Data Set- A concluded results have been formed as shown in table 2.1
A-BIRCH: Automatic Threshold Estimation for the BIRCH Clustering Algorithm	2017	researchers presented A-BIRCH, which is “an approach to automatic threshold estimation for the BIRCH clustering algorithm, this approach computes the optimal threshold parameter of this clustering algorithm from the data, The evaluated results show that the parallelized



		implementation of Gap Statistic with Sparkis scalable as the computation times decrease linearly with an increasing number of worker nodes
Improve BIRCH algorithm for big data clustering	2020	the study uses modifications of CF-Leaf value by modifying CF-Leaf formula from (N, LS, and SS) into CF-Leaf (modif) = (N, LS, SS, T), The modified BIRCH algorithm produces 65% less CFs than the original Birch, using silhouette coefficient on Modified BIRCH averaged 152.34% better accuracy than the standard one

## Chapter Three

### Methodology

#### 3.1 Introduction

##### BIRCH Clustering Algorithm

A Balanced Iterative Reducing and Clustering using Hierarchies is very effective algorithm when the data are too heavy for the memory to handle BIRCH algorithm is showing big potential in handling big data sets. In addition BIRCH is very time efficient on big data since it can produce good immediate clustering results from the first scan as well as it does not ignore the fact that not all the data points are equally important, in previous approaches such as K-means can be quite good on small to mid-range data, but when it comes to big data the whole story is different since they need all the data to be represented before as a pre-request in order to initiate ignoring the cost on memory as well as time of CPU pre-processing huge amount of data can exceed the size of the memory, they also needed multiple iterations of scans until it can produce a good reasonable results which can be time consuming and ineffective, not only this but they also tend to ignore the fact that not all data points should be treated the same, while BIRCH easily remove the outliers while processing on big data. By taking into consideration all the mentioned reasons it can be quite easy to utilize the BIRCH algorithm as our main algorithm when approaching big data such as medical data.

##### Enhancement over BIRCH Clustering Algorithm

In this chapter, the proposed methodology and the specific steps used to solve the enhancement problem on BIRCH clustering algorithm over big data will be illustrated in depth. An alternative of the Euclidean Distance which is widely used in data clustering usually performed in

calculating the distance of the centroid between data points over a specific data set, will be introduced and compared by the Manhattan Distance which is an alternative similarity measure to the Euclidean Distance, in this chapter Manhattan Distance will be used in BIRCH clustering algorithm in order to replace the standard Euclidean Distance and will be employed in the early stages of BIRCH algorithm in the initial scanning to upload and compact data points in the memory by building a CF-Tree, Manhattan Distance is used to calculate the centroid of these points as well as will be employed during the middle stages of BIRCH algorithm while building the CF-tree, it will be used to calculating the centroid between CF sub clusters, in the late stages of BIRCH algorithm Manhattan Distance is also employed to modify the global clustering algorithm when performing universal clustering over the while CF-Tree, by the end of this chapter MD-BIRCH algorithm is proposed as a full packaged improved solution for clustering big data sets such as medical data.

### **3.2 Manhattan Distance over K-Means Clustering Algorithm**

Sinwar&Kaushik ()represented Study of Euclidean and Manhattan Distance Metrics using Simple K-Means Clustering, based on the study of two popular distance metrics viz. Euclidean and Manhattan. A series of experiments has been performed to validate the study. They use two real and one synthetic datasets on simple K-Means clustering. The theoretical analysis and experimental results show that the Euclidean method outperforms Manhattan method in terms of number of iterations performed during centroid calculation. This work may be extended by taking more clustering algorithms with high dimensional real datasets<sup>xxix</sup>( Ismael & Ashour, 2014, 1-10.)

The Euclidean Distance is defined as:( Han &Kamber, 2001, 17)

$$d(i, j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \dots + (x_{ip} - x_{jp})^2}.$$

The Manhattan Distance [16] is defined as:

$$d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \dots + |x_{ip} - x_{jp}|.$$

### 3.3 Proposed Methodology

The Proposed Methodology is about merging Manhattan Distance into multiple phases of BIRCH clustering algorithm in order to produce an enhancement of performance and quality when applying the extracted MD-BIRCH algorithm on clustering big data sets.

The following step explains in-details applying the proposed methodology:

- 1- Assigning values to the threshold T, the maximum number of clustering features (CF) in non-leaf nodes B, the maximum number of clustering features (CF) in leaf nodes L, number of desired clusters C, optional parameter D which represent the density amount allowed below average density and another optional parameter R which represent a range of sub-clusters of near distances to each other.
- 2- MD-BIRCH algorithm starts by scanning the data set and representing the data set in a number of coordination equals to the number of dimensions in the data sets.
- 3- While scanning the data set MD-BIRCH algorithm keeps compacting the data points into a more compact version which is called clustering features (CF) in order to build an initial CF-Tree, clustering features (CF) are built by applying the calculation of the

Manhattan Distance between data points in order to find the centroid to compact data points to their relative CF.

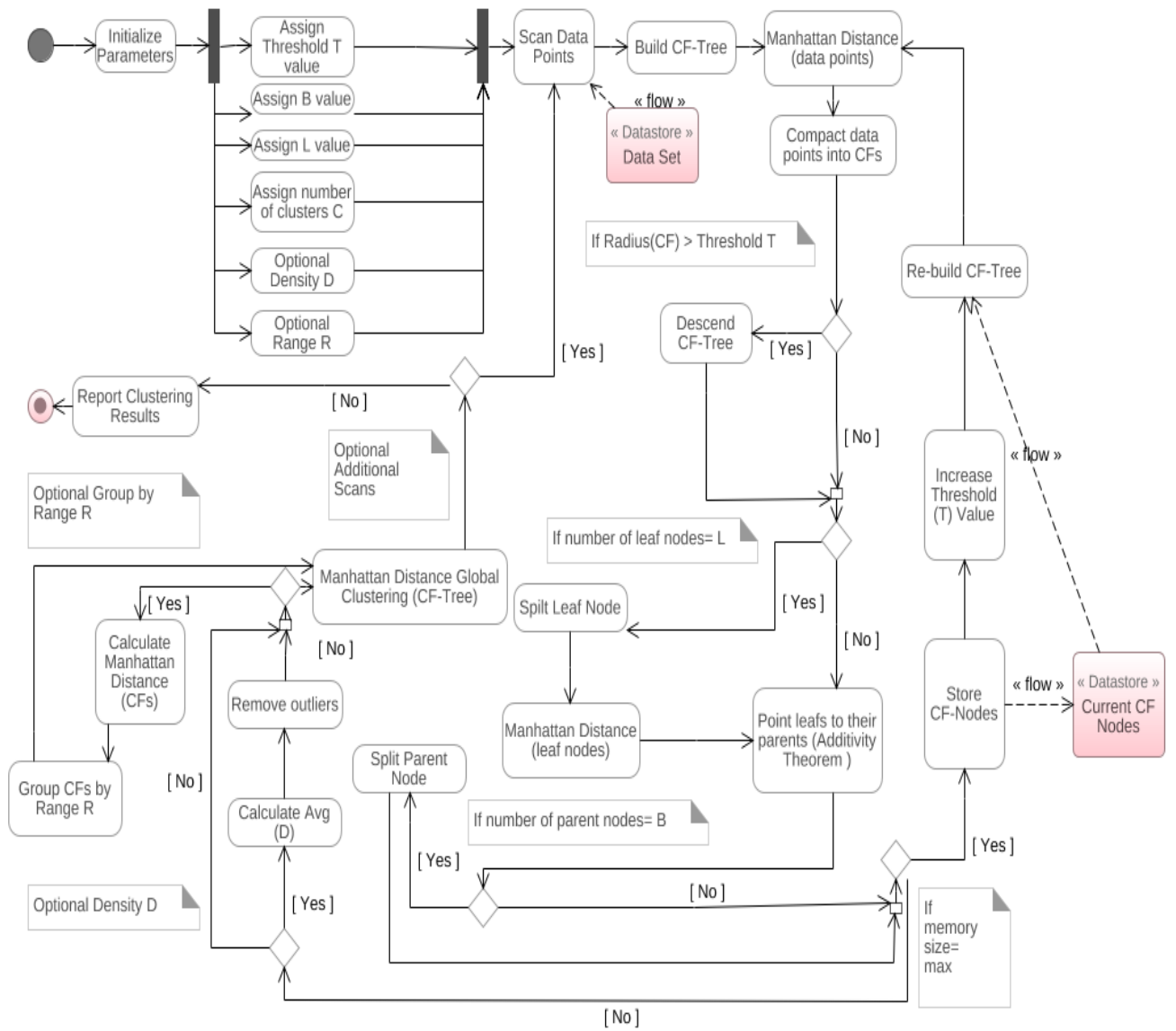
- 4- If the radius of Node exceeds the pre-specified threshold T value, then descend the tree by inserting CF-Node into the nearest leaf child node, Manhattan Distance will be applied here in order to calculate the distance of the nearest leaf child node.
- 5- If the maximum number of CFs in the leaf node L has been reached, then the leaf node is split and the farthest distance is calculated between the split nodes in order to distribute the future CF entries according to the nearest distance based on Manhattan Distance calculation.
- 6- New CF leaf nodes that has been descended while building the CF-tree must have a pointer to their parent nodes, and Additivity Theorem will be applied to calculate the value of the parent node as follows:

$$\text{Cluster Feature (1) + Cluster Feature (2) = (N1+N2, LS1+LS2, SS1+SS2)}$$

- 7- If the maximum number of CFs in parent node B has been reached, then the parent node will be spilt.
- 8- If the memory reached to its maximum limit the threshold value will be increased and a smaller tree is rebuilt by re-input the already mapped CF-Nodes of the old CF-Tree into the current smaller tree, then the algorithm will continue inserting CF leaf nodes into the Tree by repeating the rules starting from step 3.

- 9- Optional step: Calculate the average density of data points in each CF-leaf node and then find the CF-leaf nodes that are less than average in density  $D$  amount allowed and put a pointer on those nodes as they are considered outliers then those outliers can be removed to reduce the noise.
- 10- Manhattan Distance is used in the whole tree in order to find if a group of sub-clusters are less or equal the amount of range  $R$ , if founded then they can be grouped into a larger sub-cluster, which would result into a smaller tree which make future process on it much faster.
- 11- A Manhattan Distance version of any other clustering algorithm such as K-Means clustering algorithm modified by applying the same principles of Manhattan Distance calculating method of centroid of CF sub-clusters instead of data points will used to cluster the whole CF-Hierarchical Tree which is built by MD-BIRCH algorithm.
- 12- Optional step: Additional scans using MD-BIRCH can be applied in order to refine the CF-Tree, remove more outliers and overall produce a better quality clustering results.

Figure 3.1 will represent the proposed methodology design and how MD-BIRCH clustering algorithm would execute under real-world scenarios on a specific data set:



**Figure 3.1 MD-BIRCH Clustering Algorithm**

## **Chapter Four**

### **Design, Analysis & Implementation**

#### **4.1 SEER Dataset**

The main objective will be on cancer patient's data since cancer is one the most common deadly diseases all around the world, cancer data are usually complex since there are too many types of cancer but fortunately with the progressing of medical science day by day physicians can diagnose cancer faster so they can treat the patients before they reach more severe levels of spreading the cancer cells in larger areas of their body, physicians utilize the common patterns among accumulated data over the past decades to diagnose the diseases in more efficient ways, our study will also utilize data mining algorithms to find patterns over all available cancer's big data in more accurate and efficient way.

The Surveillance, Epidemiology, and End Results (SEER) program provides information on cancer statistics in an effort to reduce the cancer burden among the U.S. population. SEER is supported by the Surveillance Research Program (SRP) in NCI's Division of Cancer Control and Population Sciences (DCCPS)<sup>xxx</sup>( National Cancer Institute, Surveillance, epidemiology and end results (seer) program ([www.seer.cancer.gov](http://www.seer.cancer.gov)) SEER Research data file(1975-2017)), which is one of the most valuable resources to access cancer databases; this is why the study will depend on SEER database and use the extracted data as a main dataset for our study.



A. SEERStat is a statistical software downloaded from SEER web application after signing the data agreement, it's a tool to view individual cancer records in order to produce statistics for studying the impact of cancer on a population, the version used in this study is SEER\*Stat 8.3.6 (Figure 4.1):

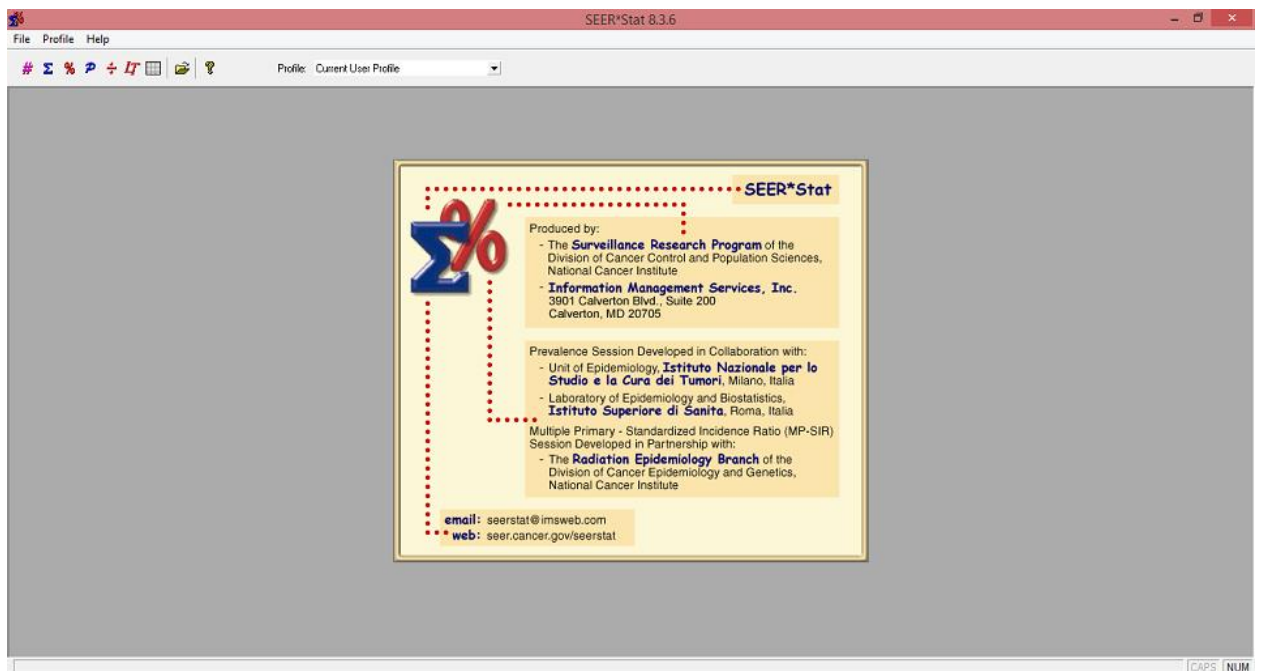


Figure 4.1: SEER\*Stat 8.3.6 is used as the main tool to provide cancer statistics

- B. By accessing the online database server and choosing the last updated database from year 1975 to 2017 (Figure 4.2), This study will focus on three types of Cancers Breast cancer, Leukemia cancer and Stomach cancer and the study will also focus on selecting a sample of ages between 20 to 64 as well as the year of diagnosis from 1990 and above (Figure 4.3)

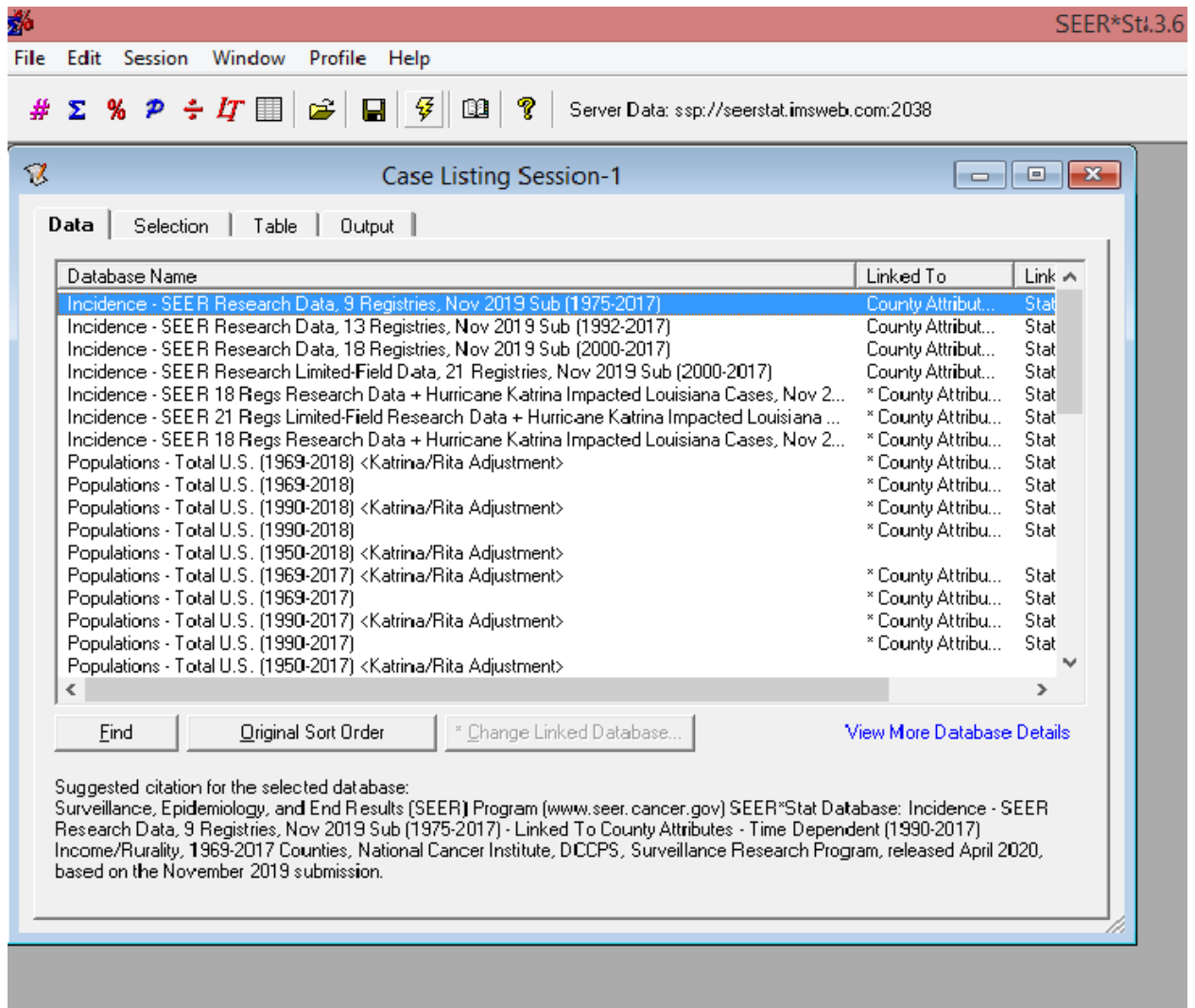
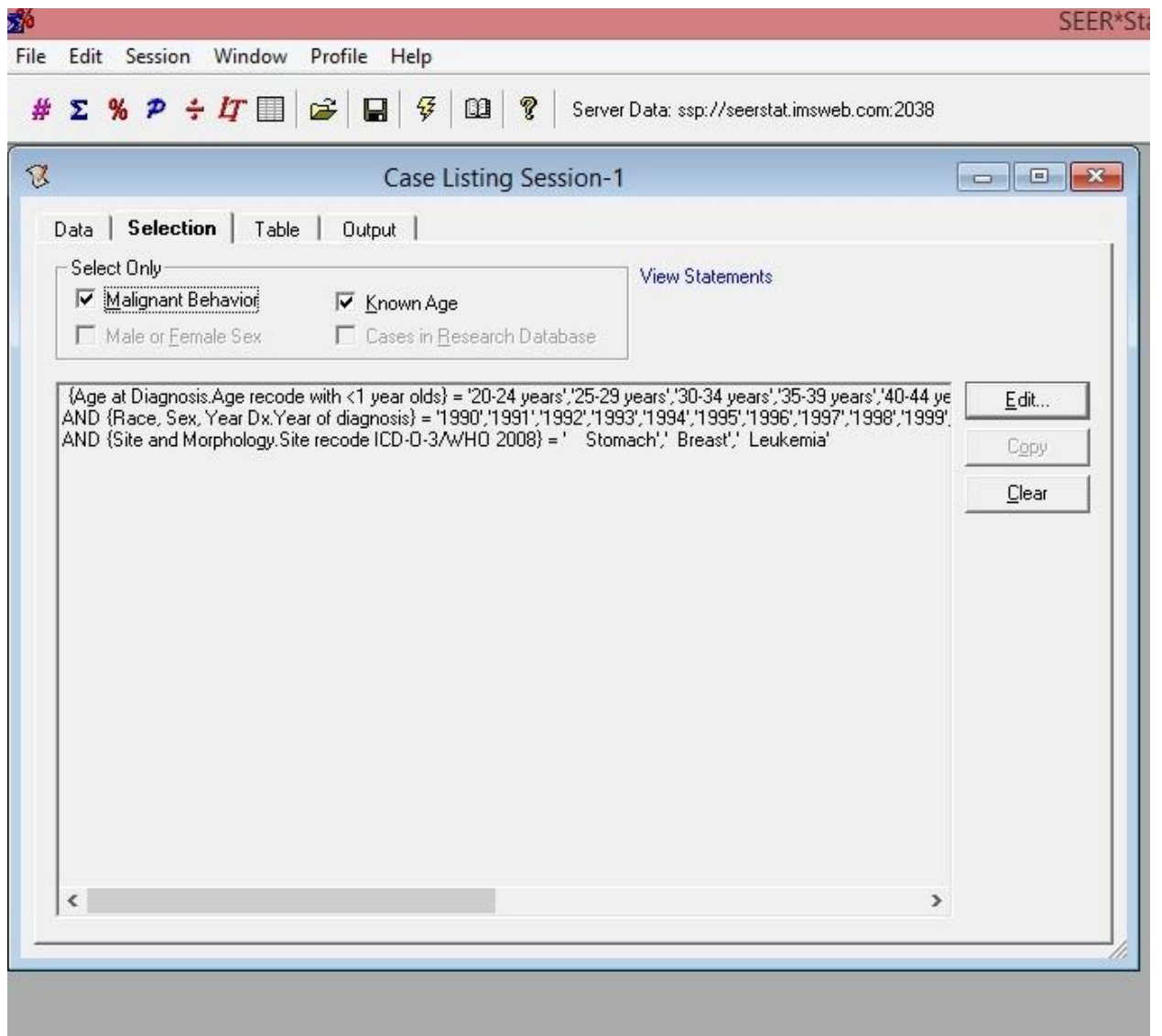
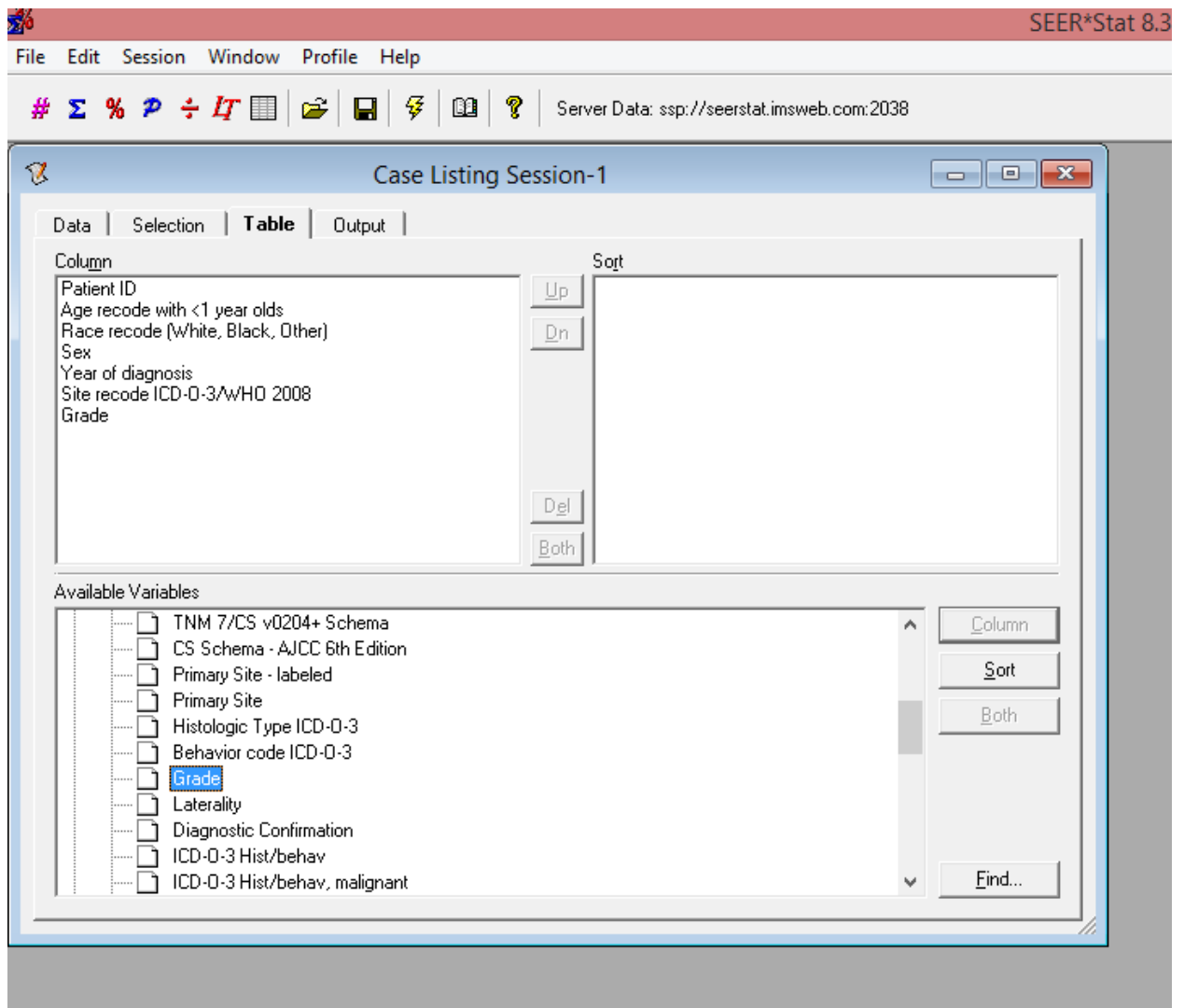


Figure 4.2: SEER Research Data



**Figure 4.3 Case Selections of Variables**

C. The columns sample needed to be represented (Patient ID, Age, Race, Sex, Year of Diagnosis, Cancer Type and the Grade of Cancer) Figure 4.4, then the program is executed in order to produce the cancer patients sample matrix of which has resulted of 367,780 patient records:



**Figure 4.4: Columns Selection**

**D.** The results has been saved in a temporary Excel sheet, the main database in this study is Oracle database 10.2 g which will be used for data cleansing, preprocessing and transformation, Oracle SQL Developer is the main tool for data inserting, updating, deleting and any other data manipulation processes, to prepare the database environment and to create a Cancer Patients Table.

The data from the Cancer\_Patient\_Info is imported into Excel sheet to the table.

#### Data cleansing and transformation process

Unknown data is cleaned the by removing any data that is ambiguous or empty (Null values)

Resulted in 77159 Rows to be deleted

The total remaining data rows count after data cleansing process is 290620 patient records

Data is transformed to numeric values since BIRCH and the newer version MD-BIRCH clustering algorithms accept data only in its numeric form, so a set of commands is executed in Oracle DB by using Oracle SQL Developer tool in order to run a set of SQL Data Manipulations that is responsible to transform the data such as the following:

Transforming Cancer Types of “Breast Cancer” into a value of ‘1’, Cancer Types of “Stomach Cancer” into a value of ‘2’ and Cancer Types of “Leukemia Cancer” including any keyword of “Aleukemic Cancer ” into a value of ‘3’.

Transforming Cancer Grades of “Grade I” into a value of ‘1’, Cancer Grades of “Grade II” into a value of ‘2’, Cancer Grades of “Grade III” into a value of ‘3’ and Cancer Grades of “Grade IV” into a value of ‘4’

Transform Male Patients into value of 1 and Female Patients into value of 2.

Transform White Race into value of 1 and Black Race into value of 2 and any other Races into value of 3.

Dividing Cancer Patients into a set of age groups where:

Patients between ages of 20 to 29 are transformed into age group 20, Patients between ages of 30 to 39 are transformed into age group of 30, Patients between ages of 40 to 49 are transformed into age group of 40, Patients between ages of 50 to 59 are transformed into age group of 50 and Patients between ages of 60 to 64 are transformed into age group of 60

After performing all the previous processes on Cancer Patients dataset the dataset will be clean and transformed into the right format, which will enable both the standard BIRCH clustering algorithm as well as the proposed version MD-BIRCH clustering algorithm to absorb the dataset and execute on it normally without facing any problems during their stages of execution.

This study is about executing BIRCH clustering algorithm as well as the proposed MD-BIRCH clustering algorithm on SEER data set that have been prepared in order to fit both algorithms, multiple clusters will be performed to see how much time each algorithm would take to execute on big data as well as see what kind of cluster figures each will produce at N number and compare the results in order to conclude if MD-BIRCH clustering algorithm worth it. Python 3.7 programming language is used in order to implement the proposed methodology, Spyder IDE from Anaconda 3 is used as the main Integrated Development Environment, Scikit-Learn machine learning library is used in order to execute the open source code of BIRCH clustering algorithm, the open source code of BIRCH clustering algorithm use Euclidean distance by default, the source code will be modified and Manhattan Distance functions will be applied in multiple stages of the algorithm as shown in the proposed methodology of MD-BIRCH clustering algorithm, A timer will be used in order to calculate the time difference between the start of execution and the end of execution of both BIRCH and MD-BIRCH clustering algorithms.

Oracle database table “tb\_Cancer\_P\_Info” that has been created will be converted into an Excel file by extracting its data in order to optimize both BIRCH and MD-BIRCH clustering algorithms then the file will be imported directly into an array inside the program the cancer array will be able to absorb cancer patients data in the right format by scanning Cancer Excel data file, the array is passed as an argument into both parameters of BIRCH and MD-BIRCH clustering algorithms.

The program is aimed to produce two types of results:

- 1- The time taken of executing both BIRCH and MD-BIRCH clustering algorithms on a set of clusters; this will produce quantitative type results.
- 2- The graphical dimensional plots results of executing both BIRCH and MD-BIRCH clustering algorithms on a set of clusters; this will produce qualitative type results.

The two types of results can be used on evaluating how well MD-BIRCH clustering algorithm has performed over medical data set (SEER) both in performance using the quantitative type results as well as the accuracy using the qualitative type results in comparison to BIRCH clustering algorithm.

## **4.2 Experiment**

The data set is formed of 290620 records of total number of patients cancer data taken from SEER database after it have been cleansed and transformed, SEER\*Stat tool is used to query a sample of patients ages between 20 to 64 as well as the year of diagnosis from 1990 and above, the study will focus on three types of cancers:

1. Breast cancer
2. Stomach cancer
3. Leukemia cancer

A set four dimensional data that are influential to the studying of multiple factors of cancer patient's data will be injected into BIRCH and MD-BIRCH clustering algorithms:



- 1- Age Group
- 2- Year of Diagnosis
- 3- Cancer Type
- 4- Cancer Degree

The injected data will be scaled at a specific weight where dimension x represents age group scaled by direct correlation with cancer type and cancer degree while inverse correlation with year of diagnosis and dimension y represents year of diagnosis scaled by direct correlation with cancer type and cancer degree also while inverse correlation with age group.

The main program is prepared to execute multiple times from 2-9 clusters on each BIRCH and MD-BIRCH clustering algorithms which in consequence will produce different graphs with each set of graphs can have similarity as well as differences, the graphs can be used to evaluate on the quality of both clustering approaches and the quantitative results produced from the time output of executing the clustering algorithms. Scikit-Learn Machine Learning Library is the main source to get to execute the Standard BIRCH clustering algorithm as well as it's the main source to extract the code of BIRCH algorithm in order to be able to modify it including all of its parts such as the global clustering algorithm in order to produce MD-BIRCH clustering algorithm by modifying it with the Manhattan distance including the global cluster algorithm will be modified to a Manhattan version that embed Manhattan distance when calculating the similarity measures, the global clustering algorithm used here is Agglomerative clustering algorithm, which will be

changed instead of using its default similarity measure - Euclidean distance- the source code will be modified in order to apply Manhattan distance into its *affinity* parameter.

For determining the accuracy of MD-BIRCH clustering algorithm, there are multiple validation cluster indexes to measure the quality of a produced cluster, since previous knowledge about how the dataset should be clustered does not exist internal validation cluster index will be used and to determine which one could be more suited for the clustered data resulted from this experiment, Liu et al. presented “Understanding of Internal Clustering Validation Measures” which focused on a common used 11 internal clustering validation indexes, five aspects of clustering investigated to test their validation properties, which has concluded that SDbw index is the only measure that performs well in all five aspects<sup>xxxi</sup>( Bhardwaj, 2017, 183-186) .

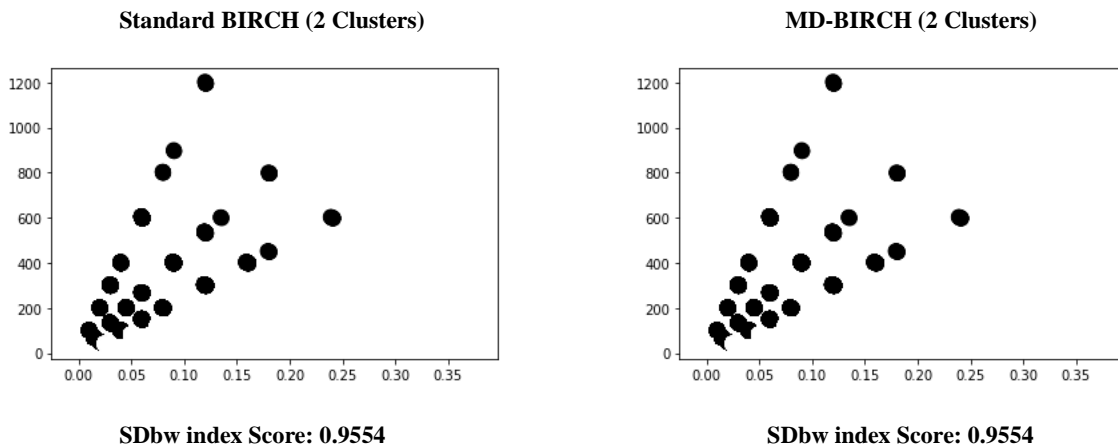
The clustering results will be measured by SDbw index in order to compare the quality of both BIRCH and MD-BIRCH clustering algorithms.

- ❖ The experiment will be applied on an Intel® Core i5 with 4 GB Memory on Windows 8.1 64bit
- ❖ The experiment will be executed by using Python 3.7 programming language

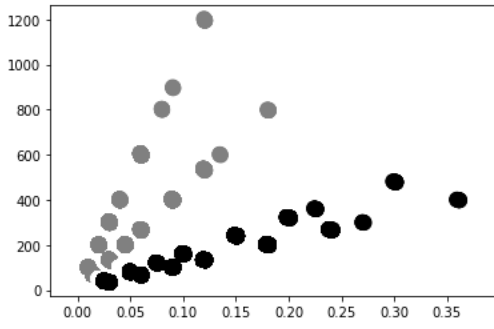
# Chapter Five

## Results

In this chapter results of executing both Standard BIRCH as well as MD-BIRCH clustering algorithms are displayed in two types, qualitative and quantitative results, the qualitative results of executing the proposed clustering algorithm on SEER big medical data from (2 to 9) required clusters and comparing it to the results of the Standard BIRCH algorithm are shown in Figure 5.1 and Table 5.1 as clustering plots measured by SDbw cluster validation index and the quantitative results are shown in Table 5.3 in order to compare the performance of both the proposed MD-BIRCH clustering algorithm with Standard BIRCH clustering algorithm, lastly summery table for qualitative results shown in Table 5.2 and quantitative results shown in Table 5.4 in order to be able to evaluate and draw a conclusion on the end results of the proposed MD-BIRCH algorithm compared with the Standard BIRCH algorithm by performance and quality.

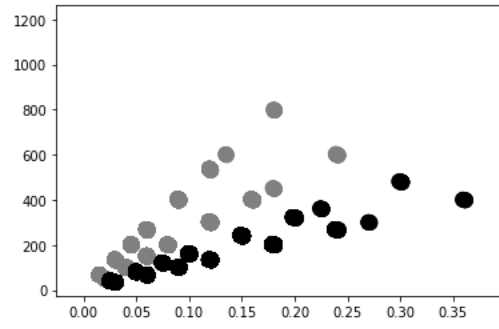


**Standard BIRCH (3 Clusters)**



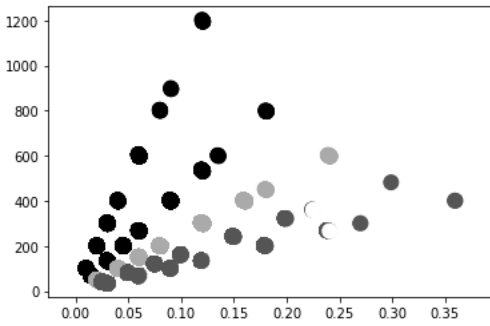
**SDbw index Score: 0.9505**

**MD-BIRCH (3 Clusters)**



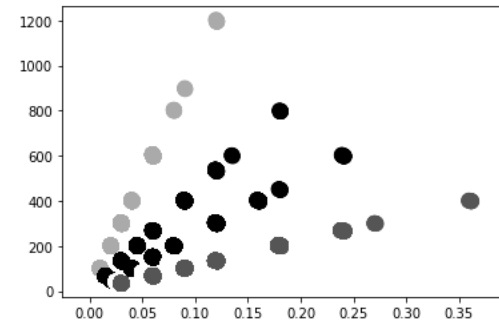
**SDbw index Score: 0.9300**

**Standard BIRCH (4 Clusters)**



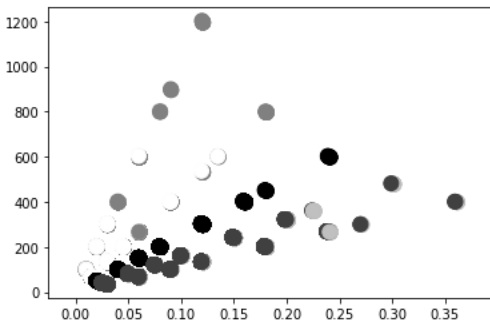
**SDbw index Score: 0.9302**

**MD-BIRCH (4 Clusters)**



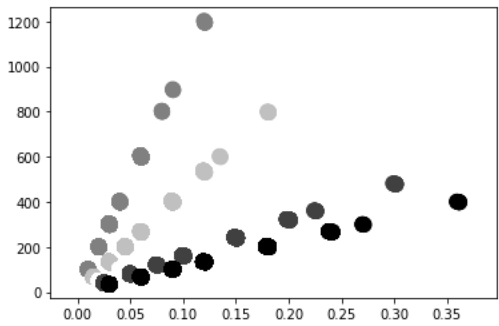
**SDbw index Score: 0.7887**

**Standard BIRCH (5 Clusters)**



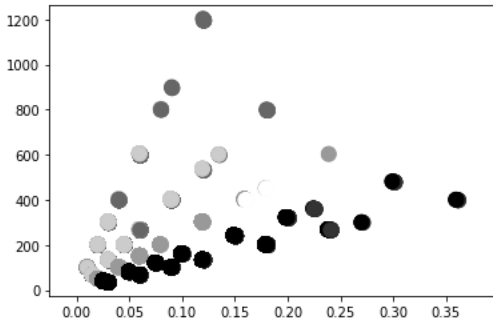
**SDbw index Score: 0.9089**

**MD-BIRCH (5 Clusters)**



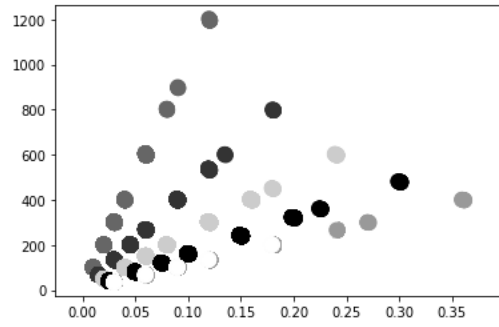
**SDbw index Score: 0.6786**

**Standard BIRCH (6 Clusters)**



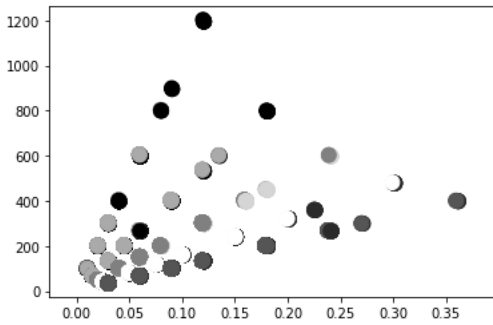
**SDbw index Score: 0.7771**

**MD-BIRCH (6 Clusters)**



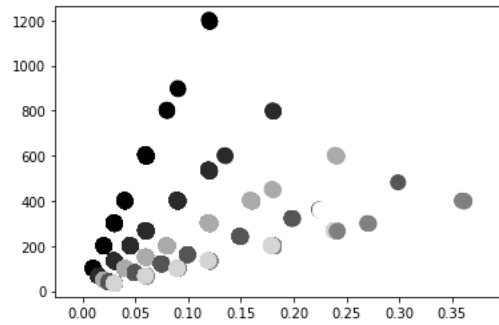
**SDbw index Score: 0.6195**

**Standard BIRCH (7 Clusters)**



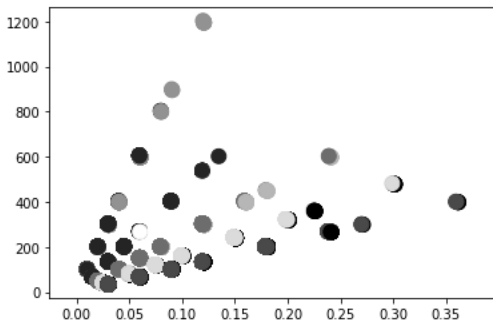
**SDbw index Score: 0.7089**

**MD-BIRCH (7 Clusters)**



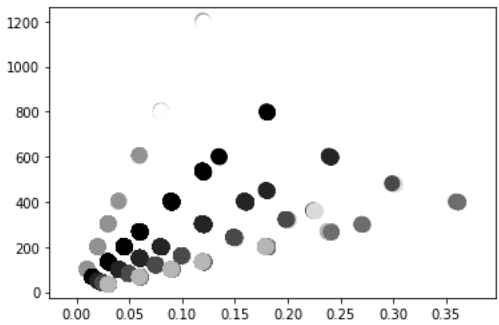
**SDbw index Score: 0.6194**

**Standard BIRCH (8 Clusters)**

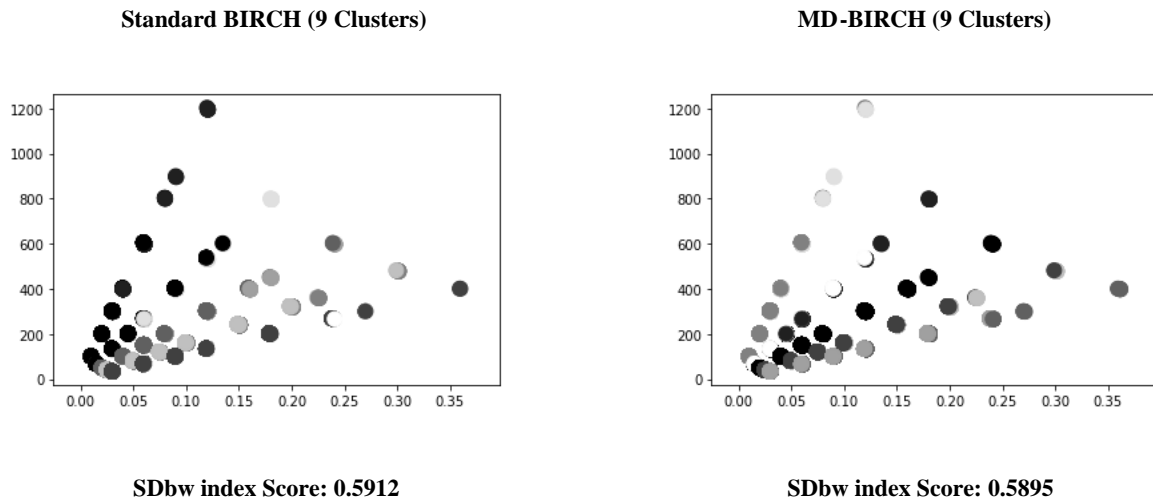


**SDbw index Score: 0.6458**

**MD-BIRCH (8 Clusters)**



**SDbw index Score: 0.6372**



**Figure 5.1: Cluster results of Standard BIRCH vs. MD-BIRCH quality measured by validation through SDbw index over multiple clustering iterations from 2 to 9 clusters**

<b>Number of Clusters/Algorithm</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	<b>7</b>	<b>8</b>	<b>9</b>
<b>Standard BIRCH</b>	0.9554	0.9505	0.9302	0.9089	0.7771	0.7089	0.6458	0.5912
<b>MD-BIRCH</b>	0.9554	0.9300	0.7887	0.6786	0.6195	0.6194	0.6372	0.5895

**Table 5.1: SDbw cluster validation score for Standard BIRCH and MD-BIRCH clustering algorithms under certain number of clusters (lower value > better quality)**

<b>Method/ SDbw Score</b>	<b>Standard BIRCH</b>	<b>MD-BIRCH</b>
<b>Total Cluster Validation Score</b>	6.4680	5.8183
<b>Average Cluster Validation Score</b>	0.8085	0.7273

**Table 5.2: Summary quality table of execution of Standard BIRCH and MD-BIRCH clustering algorithms on number of clustering iterations from 2 to 9 clusters**

<b>Number of Clusters/Algorithm</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	<b>7</b>	<b>8</b>	<b>9</b>
<b>Standard BIRCH</b>	56.52s	56.33s	57.45s	56.72s	55.94s	56.04s	56.69s	56.44s
<b>MD-BIRCH</b>	55.45s	56.25s	56.43s	56.34s	55.77s	55.12s	56.69s	55.45s

**Table 5.3: Time of execution of Standard BIRCH and MD-BIRCH clustering algorithms under certain number of clusters**

<b>Method/Time</b>	<b>Standard BIRCH</b>	<b>MD-BIRCH</b>
<b>Total Time</b>	452.13s	447.50s
<b>Average Time</b>	56.51s	55.93s

**Table 5.4: Summary performance table of execution of Standard BIRCH and MD-BIRCH clustering algorithms on number of clustering iterations from 2 to 9 clusters**



## Chapter Six

### Conclusion & Future Work

#### 7.1 Conclusion

The proposed methodology of modified BIRCH clustering algorithm which is already good, fast and reliable over big data such as SEER medical data has introduced an enhanced version of BIRCH clustering algorithm, which is called MD-BIRCH an algorithm that reliant on utilizing Manhattan distance similarity measure in multiple phases during the execution of MD-BIRCH clustering algorithm including its global clustering algorithm.

MD-BIRCH has outperformed BIRCH algorithm in 2 approaches:

- 1- The qualitative approach: which measure the quality of the resulted clusters of BIRCH vs. MD-BIRCH over multiple iterations of clusters by utilizing the unique SDbw index which is one of the most reliable internal cluster validation index; which has shown that MD-BIRCH outperformed Standard BIRCH on every clustering iteration from 2 to 9 clusters, the average quality of MD-BIRCH was (0.7272) SDbw index vs. the average quality of Standard BIRCH (0.8085), in SDbw index the lower the score the better; MD-BIRCH algorithm has an enhanced quality over BIRCH clustering algorithm with 10.04% over big medical data.
- 2- The quantitative approach: which measure the performance of BIRCH vs. MD-BIRCH clustering algorithm over multiple iterations of clusters by calculate the time difference between the start time of execution and the end time of execution, the average execution time of MD-BIRCH was (55.93) seconds vs. the Standard BIRCH (56.51) seconds, MD-

BIRCH has slightly an enhanced performance over BIRCH clustering algorithm with difference of 1.03% over big medical data.

MD-BIRCH clustering algorithm has an impressive enhanced quality over big data, MD-BIRCH has just slightly enhanced performance over mid to big dataset samples, however in much bigger data the performance can increase as well as over bigger data any slightly performance improvement can have such big impact when the data consume much time while performing clustering or any machine learning execution over it.

## **7.2 Future Work**

This work has utilized Manhattan distance similarity measurement inside multiple phases in BIRCH clustering algorithm, Manhattan distance used at early, middle and late stages of executing BIRCH algorithm, it has modified the structure of BIRCH including its global clustering algorithm, which has produced MD-BIRCH clustering algorithm.

This work can be extended in order to utilize more similarity measures such as: Minkowski distance, Jaccard similarity, Cosine Similarity...etc.

MD-BIRCH is an unsupervised machine learning clustering algorithm which use also another global clustering algorithm such as K-means and Agglomerative algorithms in order to perform universal scan on the whole Clustering Feature (CF) Tree, this work can be extended by doing a comparison between global clustering algorithms used in MD-BIRCH and in order to find which global clustering algorithm can perform well over big data in terms of both quality and the performance, then tune the algorithm using Manhattan

distances or other similarity measures Minkowski distance, Jaccard similarity and modify the algorithm by utilizing different machine learning techniques such as Reinforcement Learning.

## References:

Andritsos, P.(2002). Data Clustering Techniques.

Bhardwaj, S. (2017). Data mining clustering techniques – A review, International Journal of Computer Science and Mobile Computing.

Chayadevi, M.L., &Raju, G.T. (2012). Data mining, classification and clustering with morphological features of microbes. International Journal of Computer Applications.

Dong, J., Wang, F., & Yuan, B. (2013). Accelerating BIRCH for Clustering Large Scale Streaming Data Using CUDA Dynamic Parallelism, [International Conference on Intelligent Data Engineering and Automated Learning](#).

Garg, A., Mangla, A., Gupta, N., &Bhatnagar, V. (2006). PBIRCH: A Scalable Parallel Clustering algorithm for Incremental Data, International Database Engineering and Applications Symposium(IDEAS'06).

Han, J., &Kamber, M. (006). Data mining: Concepts and techniques (2ed Ed.). Beijing: China Machine Press.

Han,J. &Kamber, M. (2001). Data Mining: Concepts and Techniques, Morgan Kaufmann Publisher, San Francisco, USA

Ismael, N., Alzaalan, M., &Ashour, W. (2014). Improved multi threshold birch clustering algorithm. International Journal of Artificial Intelligence and Applications for Smart Devices.

Liu, Y., Li, Z., Xiong, H., Gao, X., &Wu, J. (2010). Understanding of Internal Clustering Validation Measures. 2010 IEEE International Conference on Data Mining

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 2<sup>nd</sup> INNS Conference on Big Data*, October 23-25, 2016, Thessaloniki, Greece.

Ramadhani, F., Zarlis, M., & Suwilo, S. (2019). Improve BIRCH algorithm for big data clustering.

Sajana, T., Rani, C.M.S., & Narayana, V. (2016). A survey on clustering techniques for big data mining, *Indian Journal of Science and Technology*.

Sinwar, D., & Kaushik, R. (2014). Study of Euclidean and Manhattan Distance Metrics using Simple K-Means Clustering. *International Journal for research in applied science & engineering technology*, 2321-9653.

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.

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

Zhang, T., Ramakrishnan, R., & Linvy, M. (1997). BIRCH: A New Data Clustering Algorithm and Its Application.s, *SIGMOD '96 6/96 Montreal, Canada IQ 1997 ACM 0-89791 -794-4/96/0006*

National Cancer Institute, Surveillance, epidemiology and end results (seer) program

([www.seer.cancer.gov](http://www.seer.cancer.gov)) SEER Research data file(1975-2017).

## Appendix

In order to prepare a cancer patients table the following command is executed in Oracle database by using Oracle SQL Developer tool:

```
CREATE TABLE "BI"."TB_CANCER_P_INFO"  
("P_ID" NUMBER(8,0),  
 "AGE_GROUP" VARCHAR2(25),  
 "RACE" VARCHAR2(100),  
 "SEX" VARCHAR2(10),  
 "DIAGNOSIS_YEAR" NUMBER(4,0),  
 "CANCER_TYPE" VARCHAR2(100),  
 "GRADE" VARCHAR2(100))
```

Data cleansing is performed by removing unknown data which is executed as the following command:

```
Delete from tb_Cancer_P_Info where C_Grade not like '%Grade%'
```

Data transformation is performed by transforming the data into numerical ones by executing the following commands:

```
Update tb_Cancer_P_Info set Cancer_Type='1' where Cancer_Type='Breast';  
Update tb_Cancer_P_Info set Cancer_Type='2' where Cancer_Type='Stomach';  
Update tb_Cancer_P_Info set Cancer_Type='3' where Cancer_Type like '%Leukemia%'  
or Cancer_Type like '%Aleukemic%';  
Update tb_Cancer_P_Info set C_Grade ='1' where C_Grade like '%Grade I%';  
Update tb_Cancer_P_Info set C_Grade ='2' where C_Grade like '%Grade II%';  
Update tb_Cancer_P_Info set C_Grade ='3' where C_Grade like '%Grade III%';  
Update tb_Cancer_P_Info set C_Grade ='4' where C_Grade like '%Grade IV%';  
Update tb_Cancer_P_Info set Sex=1 where Sex='Male';  
Update tb_Cancer_P_Info set Sex=2 where Sex='Female';  
Update tb_Cancer_P_Info set Race=1 where Race='White';  
Update tb_Cancer_P_Info set Race=1 where Race='Black';  
Update tb_Cancer_P_Info set Race=3 where Race like '%Other%';
```

After modification of BIRCH algorithm to the proposed MD-BIRCH algorithm (3.3 Proposed Methodology (Steps + Figure 3.1 MD-BIRCH Clustering Algorithm) and extraction of the prepared SEER dataset, the following commands are being executed in Python programming language:

```
import time

for r1 in range(2,10):

    start_time = time.time()

    BIRCH_clust_algo = BIRCH(n_clusters=r1)

    BIRCH_clust_algo.fit(Data_Val)

    labels = BIRCH_clust_algo.predict(Data_Val)

    print((time.time() - start_time))

for r2 in range(2,10):

    start_time = time.time()

    MD_BIRCH_clust_algo = MD_BIRCH(n_clusters=r2)

    MD_BIRCH_clust_algo.fit(Data_Val)

    labels = MD_BIRCH_clust_algo.predict(Data_Val)

    print((time.time() - start_time))
```

---

<sup>i</sup> Han, J., &Kamber, M. (2006). **Data mining: Concepts and techniques** (2ed Ed.). Beijing: China Machine Press.

<sup>ii</sup> Han, J., &Kamber, M. (2006). **Data mining: Concepts and techniques** (2ed Ed.). Beijing: China Machine Press.

<sup>iii</sup> Han, J., &Kamber, M. (006). **Data mining: Concepts and techniques** (2ed Ed.). Beijing: China Machine Press.

---

<sup>iv</sup> National Cancer Institute, Surveillance, epidemiology and end results (seer) program ([www.seer.cancer.gov](http://www.seer.cancer.gov)) SEER Research data file(1975-2017).

<sup>v</sup> 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.

<sup>vi</sup> 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.

<sup>vii</sup> 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.

<sup>viii</sup> Bhardwaj, S. (2017). Data mining clustering techniques – A review. **International Journal of Computer Science and Mobile Computing**, 6(5), 183-186.

<sup>ix</sup> 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.

<sup>x</sup> 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.

<sup>xi</sup> 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.

<sup>xii</sup> 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.

<sup>xiii</sup> Zhang, T., Ramakrishnan, R., &Linvy, M. (1997).BIRCH: A New Data Clustering Algorithm and ItsApplications



---

<sup>xiv</sup> Zhang, T., Ramakrishnan, R., & Linvy, M. (1997). BIRCH: A New Data Clustering Algorithm and Its Applications

<sup>xv</sup> Garg, A., Mangla, A., Gupta, N., & Bhatnagar, V. (2006). PBIRCH: A Scalable Parallel Clustering algorithm for Incremental Data. **International Database Engineering and Applications Symposium (IDEAS'06)**.

<sup>xvi</sup> Garg, A., Mangla, A., Gupta, N., & Bhatnagar, V. (2006). PBIRCH: A Scalable Parallel Clustering algorithm for Incremental Data. **International Database Engineering and Applications Symposium (IDEAS'06)**.

<sup>xvii</sup> 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.

<sup>xviii</sup> 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.

<sup>xix</sup> Dong, J., Wang, F., & Yuan, B. (2013). Accelerating BIRCH for Clustering Large Scale Streaming Data Using CUDA Dynamic Parallelism.

<sup>xx</sup> Dong, J., Wang, F., & Yuan, B. (2013). Accelerating BIRCH for Clustering Large Scale Streaming Data Using CUDA Dynamic Parallelism.

<sup>xxi</sup> 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 2<sup>nd</sup> INNS Conference on Big Data**, October 23-25, 2016, Thessaloniki, Greece (pp.169-178).

<sup>xxii</sup> 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 2<sup>nd</sup> INNS Conference on Big Data**, October 23-25, 2016, Thessaloniki, Greece (pp.169-178).

<sup>xxiii</sup>

<sup>xxiv</sup> 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 2<sup>nd</sup> INNS Conference on Big Data**, October 23-25, 2016, Thessaloniki, Greece (pp.169-178).

<sup>xxv</sup> Ramadhani, F., Zarlis, M., & Suwilo, S. (2019). Improve BIRCH algorithm for big data clustering.

---

<sup>xxvi</sup> 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.

<sup>xxvii</sup> 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.

<sup>xxviii</sup> 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.

<sup>xxix</sup> Sinwar, D., & Kaushik, R. (2014). Study of Euclidean and Manhattan Distance Metrics using Simple K-Means Clustering. **International Journal for research in applied science & engineering technology**, 2321-9653.

<sup>xxx</sup> National Cancer Institute, Surveillance, epidemiology and end results (seer) program ([www.seer.cancer.gov](http://www.seer.cancer.gov)) SEER Research data file(1975-2017).

<sup>xxxi</sup> Bhardwaj, S. (2017). Data mining clustering techniques – A review. **International Journal of Computer Science and Mobile Computing**, 6(5), 183-186.