International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

An Enhancement of Hierarchical Clustering Algorithms

Sultan Abdullah Alatawi[1], Nazar Elfadil Mohamed[2]

Department of Computer Science,Fahad Bin Sultan University, Saudi Arabia

Abstract: ***Nowadays, with large and great expansion of the Internet, worldtechnology, 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 [1]. 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.***

Keywords: Clustering BIRCH Hierarchical Classification Data Mining

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 [4].So, data representation will be simplified by classifying them into known classes [3]. Classification in data mining has some requirements like pre-defined classes, discrete data domain, sufficient amount of training data, and attribute values [3]. The significant objective of order is to anticipate the objective class for each case in the information base [3]. For instance, arrangement is utilized when a bank credit official needs to dissect information to realize which advance candidates are sheltered and who one undependable to loan [2].There are many algorithms and techniques that can be employed in classification 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 [4]. That is, expectation models foresee consistent worth capacities. It is used to find a numerical output. For instance, prediction models are used by marketing managers to predict how much the customer will spend during a sale. 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. In like manner, the group is an assortment of information protests that are like one another and disparate with objects of different bunches [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.

There are certain requirements of clustering such as scalability (highly scalable clustering algorithms are required to deal with large databases), ability to deal with different kinds of attributes (algorithms need to be applied), clustering any type of data such as interval-based (numerical) data, categorical data, and binary data, discovery of clusters with attribute shape (it is important to develop algorithms that can detect clusters of arbitrary shapes), high dimensionality (the clustering algorithm should be able to handle the high-dimensional space), the ability to deal with noisy data (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 [6].

2. Background

Nowadays, with large and great expansion of the Internet, worldtechnology, 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 [1]. 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 [1][2]. 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 helps us to know what the difference between data and information is, where data are facts that have been collected in raw forms and information is processed data. In general, data mining transforms data into information [3]. It 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 BIRCH clustering algorithm in medical applications may increase the accuracy 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.

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.
Q1: What is the violence of data preprocessing and rescaling on BIRCH clustering performance? When?
Q2: How to Hyper-cubes of BIRCH will be divided into two groups; Left-side group that holds numbers less than or equal to the pivot andRight-side group that holds numbers greater than the pivot?
Q3: Which are the most significant gaps and limitations in the reviewed studies?

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 an authoritative source for cancer statistics in the United States of America. The benefit of linking these two data sources is production of a unique population-based source of information that can be used for an array of epidemiological and health services research. In fact, the linked SEER-Medicare data files are large and complex. 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. The KDD process comprises five major steps (Figure 1): data selection, pre-processing, transformation, data mining, and generation of knowledge [10]. Further details on each of these steps are provided in the following paragraphs.

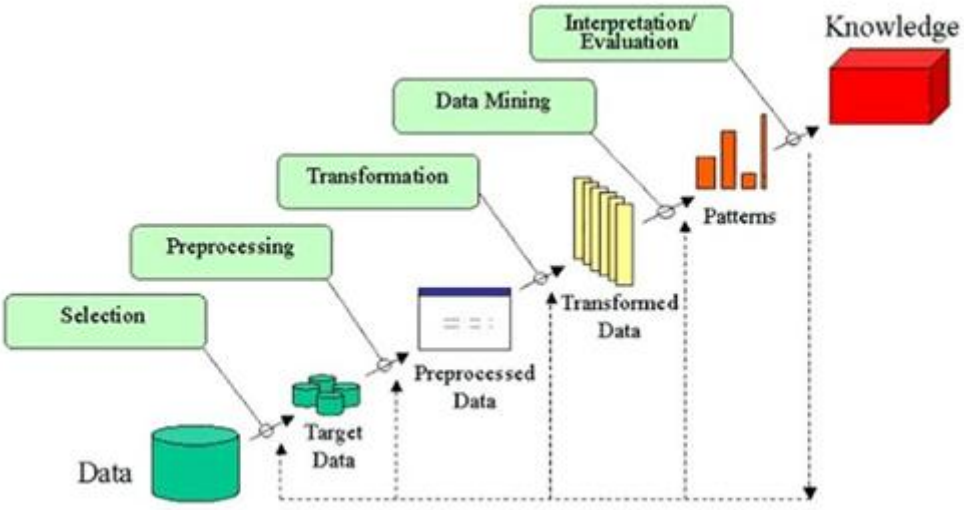

Figure 1: The Basic Steps in the Knowledge Discovery in Databases Process [10]

1) Data selection.
- Data selection is defined as the process where data relevant to the analysis is decided and retrieved from the data collection.
- Data selection using Neural network.
- Data selection using Decision Trees.
- Data selection using Naive bayes.
- Data selection using Clustering, Regression, etc.

2) Data cleaning and pre-processing.
- Data cleaning is defined as removal of noisy and irrelevant data from collection.
- Cleaning in case of Missing values.
- Cleaning noisy data, where noise is a random or variance error.
- Cleaning with Data discrepancy detection and Data transformation tools.

3) Data transformation.

- Data Transformation is defined as the process of transforming data into appropriate form required by mining procedure.
- Data Transformation is a two step process:
- Data Mapping: Assigning elements from source base to destination to capture transformations.
- Code generation: Creation of the actual transformation program.

4) Data mining.

- Data mining is defined as clever techniques that are applied to extract patterns potentially useful.
- Transforms task relevant data into patterns.
- Decides purpose of model using classification or characterization.

5) Interpretation and Evaluation.

- Data integration is defined as heterogeneous data from multiple sources combined in a common source (DataWarehouse).
- Data integration using Data Migration tools.
- Data integration using Data Synchronization tools.
- Data integration using ETL(Extract-Load-Transformation) process.
- Evaluation is defined as as identifying strictly increasing patterns representing knowledge based on given measures.
- Find interestingness score of each pattern.
- Uses summarization and Visualization to make data understandable by user.

6) Knowledge generation and use.

- Knowledge representation is defined as technique which utilizes visualization tools to represent data mining results.
- Generate reports.
- Generate tables.
- Generate discriminant rules, classification rules, characterization rules, etc.

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 [4].So, data representation will be simplified by classifying them into known classes [3]. Classification in data mining has some requirements like pre-defined classes, discrete data domain, sufficient amount of training data, and attribute values [3]. The significant objective of order is to anticipate the objective class for each case in the information base [3]. For instance, arrangement is utilized when a bank credit official needs to dissect information to realize which advance candidates are sheltered and who one undependable to loan [2].There are many algorithms and techniques that can be employed in classification 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 [4]. That is, expectation models foresee consistent worth capacities. It is used to find a numerical output. For instance, prediction models are used by marketing managers to predict how much the customer will spend during a sale. 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. In like manner, the group is an assortment of information protests that are like one another and disparate with objects of different bunches [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.

There are certain requirements of clustering such as scalability (highly scalable clustering algorithms are required to deal with large databases), ability to deal with different kinds of attributes (algorithms need to be applied), clustering any type of data such as interval-based (numerical) data, categorical data, and binary data, discovery of clusters with attribute shape (it is important to develop algorithms that can detect clusters of arbitrary shapes), high dimensionality (the clustering algorithm should be able to handle the high-dimensional space), the ability to deal with noisy data (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 [6].

2.2 Data Mining of Medical Data: Clustering

The key challenge in data mining is the extraction of meaningful information, that is, patterns, from large 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 (Figure 1), which are the hierarchical, partitioning, density-based, grid-based, and model-based algorithms [7]. 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 1 [6].

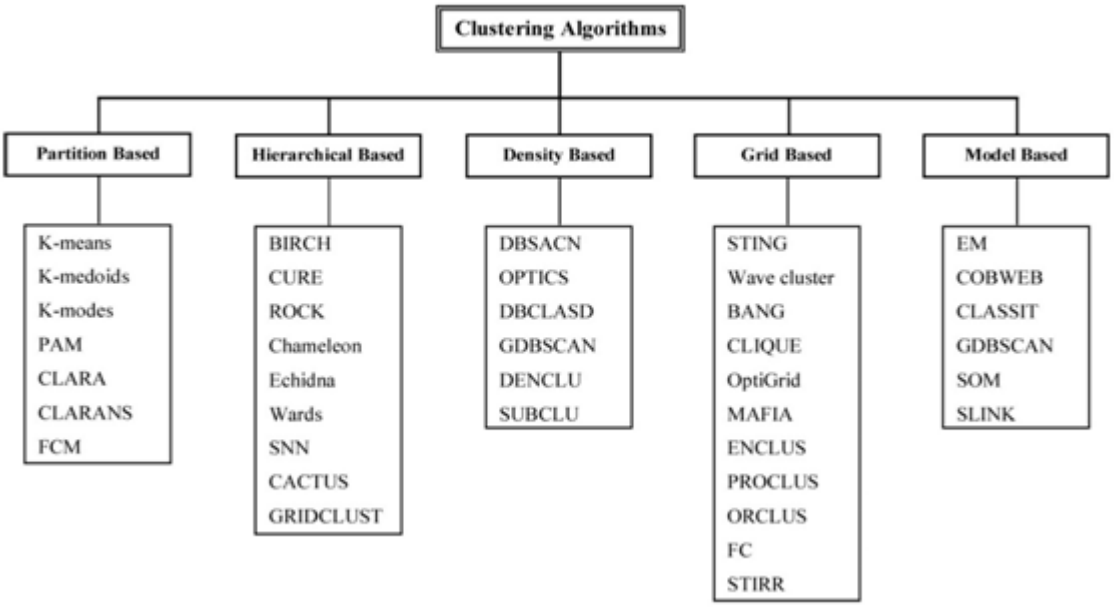

Figure 2: An Overview of Clustering Algorithms for Big Data Mining [6].

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 (Figure 2).

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 methodology considers investigating information at various degrees of granularity. The various leveled bunching calculations construct groups steadily. 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.

Hierarchical clustering has some advantages and disadvantages. The advantages include that this clustering approach is easy to implement and that it, sometimes, achieves the best result. The disadvantages of this clustering approach mainly include that there is no 'undo' in this algorithm and that it is, sometimes, difficult to identify the required number of clusters [8]. Some of the time, when we need to improve the effectiveness and nature of the progressive grouping strategy we blend (incorporate) it with other bunching techniques, for example, BIRCH, which represents Balanced Iterative Reducing and Clustering utilizing Hierarchies; ROCK (Robust Clustering calculation for Categorical ascribes); and Chameleon [5][7].

2.6 The BIRCH Hierarchical Clustering Algorithm

The reasonable iterative decreasing and bunching utilizing orders (BIRCH) calculation is a progressive grouping calculation utilized with exceptionally enormous informational indexes. 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 little portrayal of a basic bunch of one point or numerous focuses. BIRCH expands on the possibility that focuses that are sufficiently close to one the other ought to consistently be considered as a gathering. The CFs give this degree of reflection. In other words, the core of the BIRCH clustering algorithm is the CF. The BIRCHT 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 utilized to take care of two genuine issues:

(I) fabricating an iterative and intelligent pixel classification apparatus and (ii) creating an underlying codebook for picture pressure [9]. BIRCH advances in four stages:
Phase 1: Scanning all information at that point fabricating an underlying CF tree in memory by utilizing the given measure of memory and reusing space on the plate.
Phase 2: Building a smaller CF tree.

Phase 3: Performing global clustering.
Phase 4: Refining the clusters. This progression is discretionary and it requires extra disregards the information to refine the outcomes.

3. Research 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.

3.1 Research Questions (RQs)

This work attempts to answer the following research:
Q1:What is the violence of data preprocessing and rescaling on BIRCHclustering performance? When?
Q2: How to Hyper-cubes of BIRCH will be divided into two groups;Left-side group that holds numbers less than or equal to the pivot and Right-side group that holds numbers greater than the pivot?
Q3: Which are the most significant gaps and limitations in the reviewed studies?

3.2 Defining Search Strategy

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 to provide the best services for them and improve the work quality.

The targeted search strategy of this review included determining the population. Selecting resources, deriving search strings, and the inclusion and exclusion criteria. The study keywords had been determined from the research questions. The chosen keywords were doubIe-checkcd with the study research questions so as to ensure that they were indeed aligned with the research objectives and expectations; as shown in Table ( ). The search string displayed in table 2 was applied to implement search on the five selected online databases using Boolean operator to identify/ the primary' articles.

Table 1: Search Keywords

Approach and features	
Clustering	BIRCH
Hierarchical	Classification
Data Mining	SOM(Self-Organizing Feature Map)
Violence	K-means

Table 2: Search String

Main – Search String
["BIRCH" OR "Clustering Data mining" OR "BIRCH Clustering" OR "Hierarchical" OR "KDD Clustering" OR "BIRCH Algorithm" OR "Enhancement" OR "Classification"] AND ["BIRCH" OR "Clustering"] AND ["Data mining" OR "Knowledge Discovery" OR "Data base"] AND ["BIRCH" OR "Evaluation" OR "Reliability" OR "Performance" OR "Classification"]

3.3 Defining Data Sources

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 an authoritative source for cancer statistics in the United States of America. The benefit of linking these two data sources is production of a unique population-based source of information that can be used for an array of epidemiological and health services research. In fact, the linked SEER-Medicare data files are large and complex. We can create clusters of them using the BIRCH algorithm because this algorithm works for huge databases.

3.4 Defining Search Keywords

Data mining - Clustering K-mean clustering algorithm - Enhancement of Hierarchical.

3.5 Conducting Review Process

3.5.1 Pilot Search

A pilot search was carried out to identify as many results as possible related to embedded systems, readiness' as well as instrument development and validation, based on the formulated search string. As a result, a total of 136 articles were initially identified from the selected databases; 14 from Science Direct, 39 from web of Science. 21 from Scopus. 0 from Emerald, and 65 from Google Scholar, as depicted in Table 2. Lastly. A total of 117 studies were identified and excluded upon implementing the reference Management tool.

3.5.2 Selection of Study

The primary articles selected for this study had been determined based on the two-level of inclusion and exclusion criteria performed upon the selected 139 articles, where 44 studies were excluded for failure to satisfy the criteria set. Next 'some 30 articles were rejected after reading the titles and abstracts of the articles. Finally,46 studies were excluded from the 65 selected articles after reading the full texts.

Next. The remaining 19 articles were grouped into three classes based on the Res, where 8 articles were Linked to Embedded Systems (ES), 7 on Readiness Measurements (RM), And 4 on Instrument Development and validation (IDV), These remaining articles were included as primary studies and SLR was executed.

3.6 Data Synthesis

3.6.1 Primary Studies Overview

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.

3.6.2 Answering the Research Questions

This work has evaluated the performance efficiency for both of the sequential Quicksort and the BIRCH algorithms in three criteria: worst case, average case, and best case. The complexities were founded to be as shown in

Table .

Table 3: The complexities of the sequential and BIRCH algorithms.

Algorithm	Worst case	Average case	Best case
Parallel Quicksort	O(n2/P)	Ø ((n log2 n)/P)	Ω ((n log2 n)/P)
Sequential Quicksort	O(n2)	Ø (n log2 n)	Ω (n log2 n)

The results have revealed that adding more threads should increase the performance of up to 2x of the originaltime.The proposed parallel quicksort algorithm has outperformed the Quicksort method in terms of execution time and speedup on BIRCH . The proposed BIRCH algorithm has taken the advantage of the optic links connectivity.

4. Results Findings

Several cases studies were conducted to reveal the performance of the proposed method. Firstly, a random data elements of size 25 as illustrated in subsection 3.8.1. Secondly, the same data elements but in an ascending order format as introduced in subsection . Thirdly, in subsection 3.8.3 presents the results of a descending ordered data elements is introduced. Finally input arrays of larger data sets were used to show the efficiency of the proposed algorithm.

4.1 Random Data Elements

A random of 64 [2(5)] data elements illustrated in Appendix (A) were used as an input array to the testing environment. The cubic number was set to 8 and each cubic set to 4 [N=8; P=4]. The result after running the sequential Quicksort method was required 60 processes of swapping while 31 processes by the proposed parallel sorting algorithm as shown in Table. This enhancement has been obtained due to utilizing the multiprocessing environment. Furthermore, the maximum computational time in each iteration among all involved parallel processors was computed in order to conduct the comparison. Also the Quicksort algorithm was consumed 0.015 second and the proposed method consumed 0.004 second.Based on the results, it was revealed that the proposed parallel quick sort algorithm which abbreviated BIRCH has outperformed the sequential Quicksort algorithm in term of performance efficiency as shown in Figure.

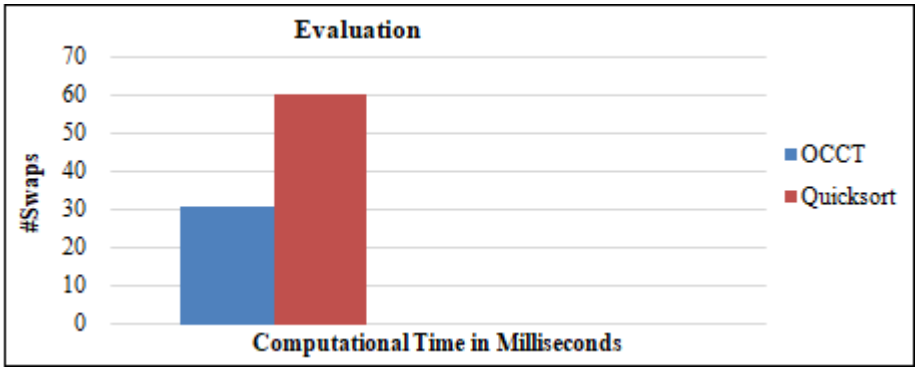

Figure 3: Comparison between BIRCH and Quicksort algorithms in terms of number of swapped buffers and computational time

Table 4: Array Segment (S) in Eachiteration of the Proposed Parallel Sorting ALGORITHAM

(i#)	Array Segment(s)	#Pivot(s) 2(i-1)	Cubic#	[Cubic#]->Processor#	Time second	#of Swapping
1	[64]	1	011	[011]-> 00	0.000	0
2	[32], [32]	2	011	[011]-> 00,01	0.003	16
3	[16] ,[16], [16],[16]	4	011	[011]-> 00,01,10,11	0.001	8
4	[8], [8], [8], [8], [8], [8], [8], [8]	8	011, 110	[011]-> 00,01,10,11 [110]-> 00,01,10,11	0.001	4
5	[4], [4], [4], [4], [4], [4], [4], [4],	16	011,110,101,100	[011]-> 00,01,10,11	0.000	2

	[4], [4], [4], [4], [4], [4], [4], [4]			[110]-> 00,01,10,11 [101]-> 00,01,10,11 [100]-> 00,01,10,11		
6	[2], [2], [2], [2], [2], [2], [2], [2], [2], [2], [2], [2], [2], [2], [2], [2], [2], [2], [2], [2], [2], [2], [2], [2], [2], [2], [2], [2], [2], [2], [2], [2]	32	011,110,101,100 001,010,000,111	[011]-> 00,01,10,11 [110]-> 00,01,10,11 [101]-> 00,01,10,11 [100]-> 00,01,10,11 [001]-> 00,01,10,11 [010]-> 00,01,10,11 [000]-> 00,01,10,11 [111]-> 00,01,10,11	0.000	1
TOTALS					0.005	31

4.2 Ascending Ordered Data Elements

The input array was sorted in an ascending order and sent as a parameter to both algorithms. The results return 0.001 second and 0 swapped buffer. This case is considered as the best case for both algorithms.

4.3 Descending Ordered Data Elements

This case is considered as the worst case scenario for the sequential Quicksort algorithm. The sequential Quicksort algorithm required 48 swapped processes during 0.078 second. The proposed BIRCH algorithms required 36 swapped processes during 0.009 second.

4.4 Sorting Large Data Elements

Several large data sets of sizes 24, 28, 212, 216, 220 and 224 were employed and represented as input arrays. The arrays were sent to both algorithms and the outputs were as follows:

The large datasets of random order were sent to both algorithms. The results revealed that the BIRCH has outperformed the sequential Quicksort algorithm as illustrated in

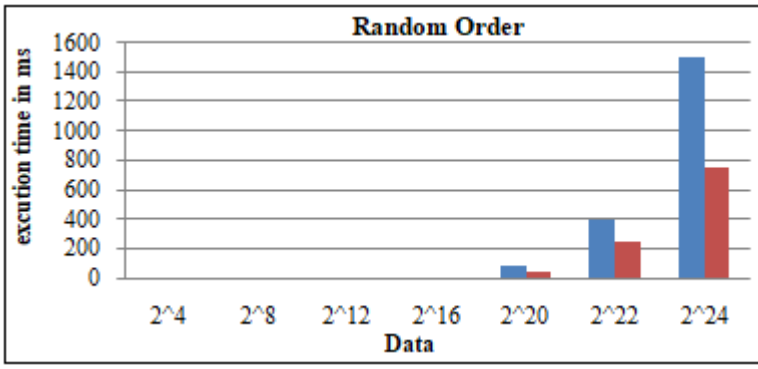

Figure **4**.

The large datasets of descending order were sent to both algorithms. The results revealed that the BIRCH has outperformed the sequential Quicksort algorithm as illustrated in

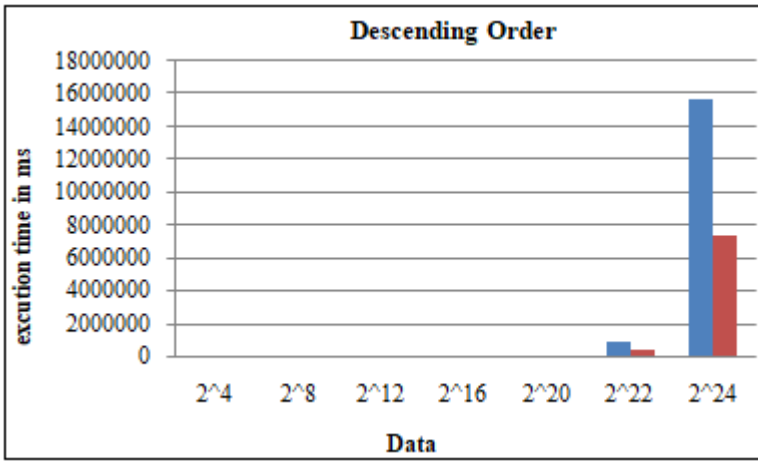

Figure).

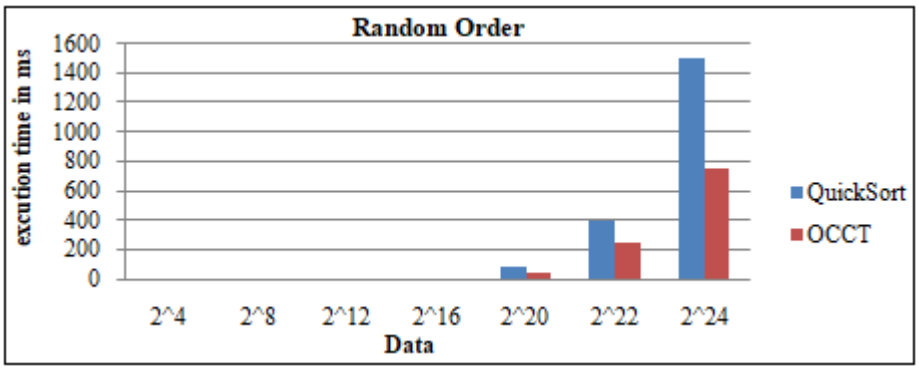

Figure 4: The performance of both Quicksort and BIRCH algorithms for random ordered data sets.

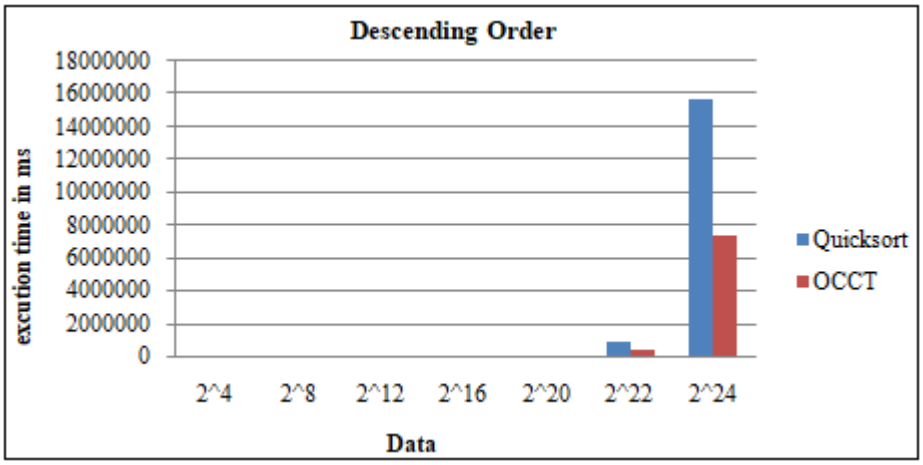

Figure 5: The performance of both Quicksort and BIRCH algorithms for descended ordered data sets.

Therefore, comparisons of random order with sizes of 1MB, 2MB, 4MB, 8MB, 16MB, and 32MB were conducted based on communicational time, computational time, total execution time, speedup, and efficiency as shown in the next figures. Random Order Communicational Time.

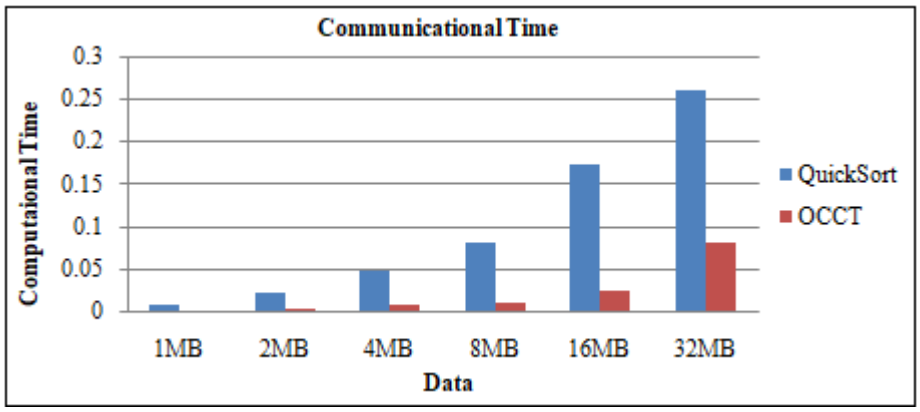

Figure 6: Communicational Time

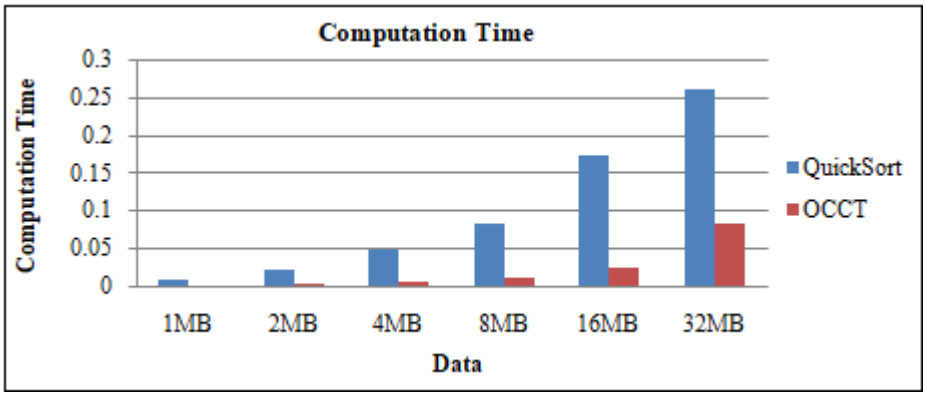

Figure 7: Random Computational Time.

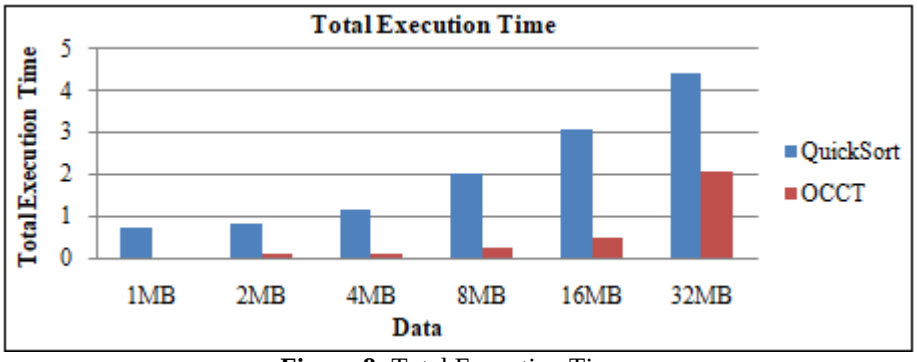

Figure 8: Total Execution Time.

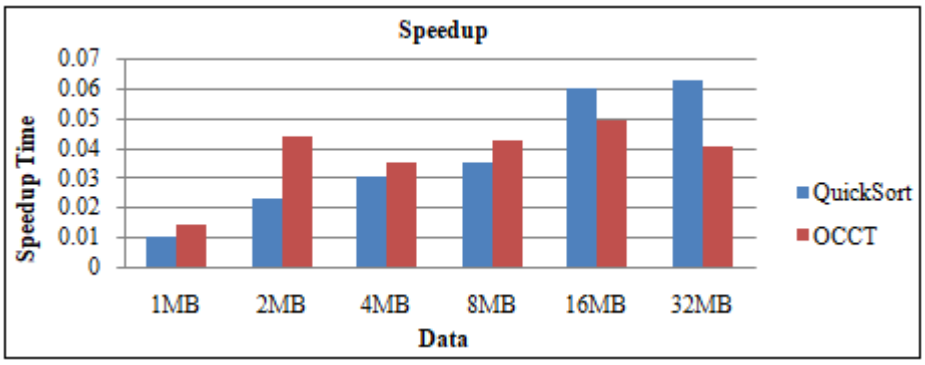

Figure 9: Speedup Time.

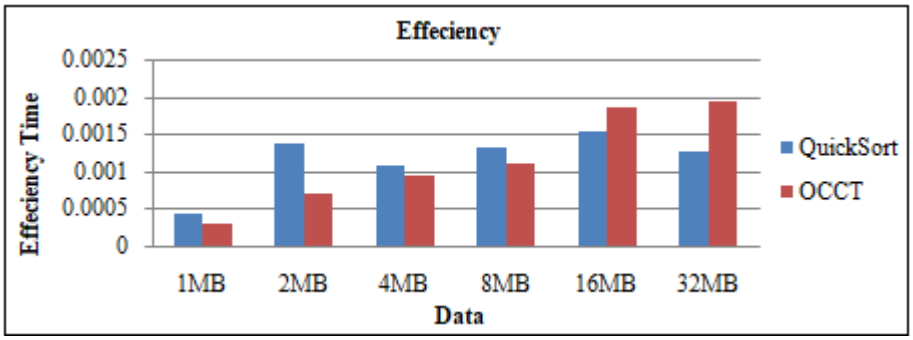

Figure 10: Efficiency

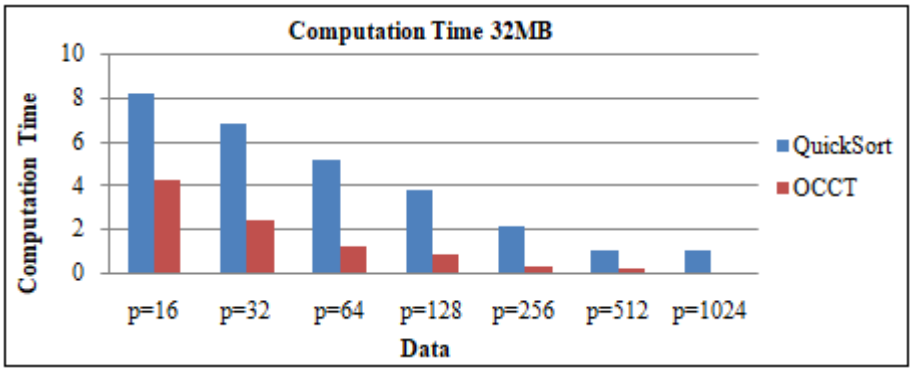

Figure 11: Computation Time 32MB

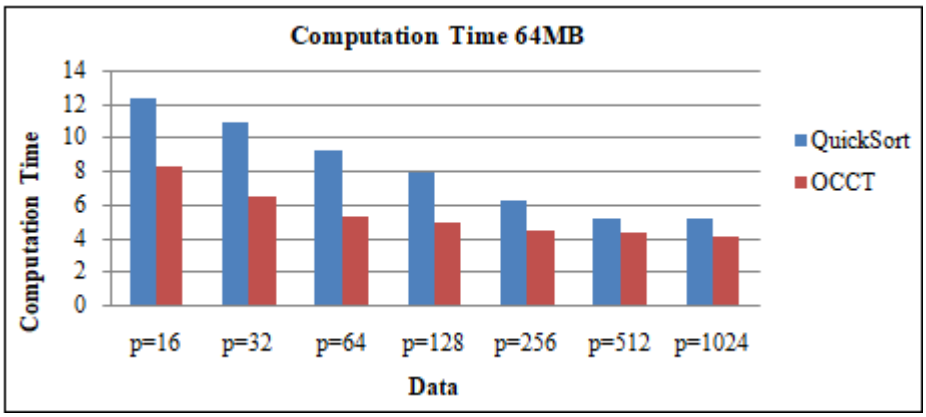

Figure 12: Computation Time 64MB.

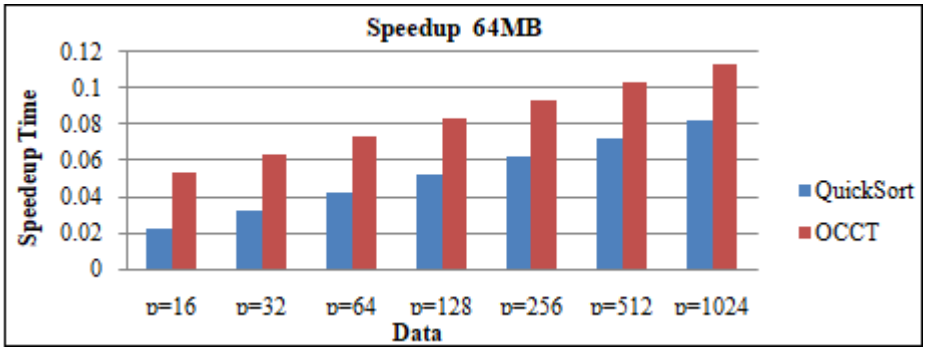

Figure 13: Speedup Time 32MB.

5. Conclusions and Suggestions for Future Works

5.1 Overview

The presented work reveals that utilizing multithreading for sorting large datasets increases the efficiency and performance of sorting algorithms. The datasets of this study comprises three types of ordering: random, ascending, and descending. Memory usages and computational time are the main challenges of sorting algorithms. This thesis presents new algorithms for data sampling (Partitioning & Distribution), traverse data elements in reverse (Gathering), and applying parallel quick sorting method over optical chained cubic tree interconnection network. The network could contains N cubes (N>1) and each cube comprises P processors connected by two types of links: electronic (UTP) and fibre optics.

5.2 Overall Conclusion

In this research study, the focus is to further knowledge in applying parallel sorting algorithms. To answer the research questions that mentioned in section**Error! Reference source not found.**, the following conclusions are founded:

The results have revealed that adding more threads should increase the performance of up to 2x of the original time.

The proposed parallel quicksort algorithm has outperformed the Quicksort method in terms of execution time and speedup on BIRCH . The proposed BIRCH algorithm has taken the advantage of the optic links connectivity.

5.3 Original Contributions

This thesis presents several contributed works as follows:
Optical chain interconnection network using electronic and optical links were designed. This network is divided into N groups of threading holding P processors.

Data sampling technique called partitioning and distributing method to distribute the data sets automatically over the network was proposed. Traverse the distributed data in reverse order to gather the data again from the nodes. A parallel quicksort algorithm in multiprocessors environment was introduced.The drawback of any interruption caused by a process from the OS was overcome by counting the number of swaps in the buffers.

5.4 Suggestions for Further Work

The focus of this study was on sorting a huge amount of datasets using parallel execution. However, there are several techniques for sorting data. Sequential Quicksort algorithm was selected as a benchmark to conduct the evaluation. A further work can be by considering other parallel sorting algorithms. Furthermore, the proposed algorithm can be optimized on selecting the suitable link by considering the speed and the congestion on to obtain better results.The future work can be based on another research question, is it possible to sort video or audio segments using the parallel sorting algorithm in better performance than using single thread.

References

[1] Han, J., & Kamber, M. (006). Data mining: Concepts and techniques (2ed Ed.). Beijing: China Machine Press.
[2] Suh, S.C. (2011). Particular applications of data mining. Massachusetts, USA: Jones & Bartlett Learning.
[3] Jackson, J. (2002). Data mining: A conceptual overview. Communications of the Association for Information Systems, 8, 267-296.
[4] 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.
[5] 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.
[6] 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.
[7] Bhardwaj, S. (2017). Data mining clustering techniques – A review. International Journal of Computer Science and Mobile Computing, 6(5), 183-186.
[8] Berkhin, P. (2006). A survey of clustering data mining techniques. In: Kogan J., Nicholas C., Teboulle M. (eds) Grouping Multidimensional Data. Berlin: Springer.
[9] 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.
[10] Rikhi, N. (2015). Data mining and knowledge discovery in database. International Journal of Engineering Trends and Technology, 23(2), 64-70.
[11] Maimon, O., & Rokach L. (2005). Introduction to knowledge discovery in databases. In: Maimon O., Rokach L. (eds) Data Mining and Knowledge Discovery Handbook (pp. 1-17). Boston, MA: Springer.
[12] 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.
[13] 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.
[14] 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.
[15] 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.
[16] Mitra, S., & Nandy, J. (2011). *KDDClus*: 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).
[17] 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.
[18] 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.
[19] 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).
[20] Xiaona Xia , (2020), " Clustering Analysis of Interactive Learning Activities Based on Improved BIRCH Algorithm" , Faculty of Education, Qufu Normal University, Qufu, Shandong, 273165; School of Computer Science, Qufu Normal University, Rizhao, Shandong, 276826, China; Chinese Academy of Education Big Data, Qufu Normal University, Qufu, Shandong, 273165.
[21] Aro, Abicoye, and Oladipo , (2019), "Enhanced Gabor Features Based Facial Recognition Using Ant Colony Optimization Algorithm", Centre for Research and Development (CERAD) The Federal University of Technology, Akure, Nigeria (www.journal.futa.edu.ng).
[22] Martin C Nwadiugwu, (2020), "Gene-Based Clustering Algorithms: Comparison Between Denclue, Fuzzy-C, and BIRCH", Department of Biomedical Informatics, University of Nebraska Omaha, Omaha, NE, USA., Bioinformatics and Biology Insights, Volume 14: 1–6.
[23] Fanny Ramadhani et al , (2020), "Improve BIRCH algorithm for big data clustering" , IOP Conf. Ser.: Mater. Sci. Eng. 725 012090.
[24] Christian Fischer et al, (2020), "Mining Big Data in Education: Affordances and Challenges" University of Tübingen, Review of Research in Education March, Vol. 44, pp. 130– 160 DOI: 10.3102/0091732X20903304.
[25] Sajana Tiruveedhula and Venkata Narayana, (2016), "A Survey on Clustering Techniques for Big Data Mining", Indian Journal of Science and Technology · February DOI: 10.17485/ijst/2016/v9i3/75971.
[26] S. K. Yadav , S. Pal," Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification", World of Computer Science and Information Technology Journal, Vol. 2, pp. 51-56, 2012.
[27] S. Karthika, N. Sairam, "A Naïve Bayesian Classifier for Educational Qualification", Indian Journal of Science and Technology, Vol. 8, pp. 1-5, 2015.
[28] Dong T, Shang W, Zhu H. An improved algorithm of Bayesian text categorization. Journal of Software. 2011; 6(9):1837–43.
[29] Jiang L, Cai Z, Zhang H, Wang D. Naïve Bayes text classifiers: a locally weighted learning approach. Journal of Experimental and Theoretical Artificial Intelligence. 2013; 25(2):273–86
[30] V. Kamra, Johina, "A Review: Data Mining Technique Used In Education Sector", International Journal of Computer Science and Information Technologies, Vol. 6, pp. 2928-2930, 2015.
[31] Nikita Jain1, Vishal Srivastava2 "DATA MINING TECHNIQUES: A SURVEY PAPER" IJRET: International Journal of Research in Engineering and Technology eISSN: 2319-1163 | pISSN: 2321-7308
[32] Kalyani M Raval "DATA MINING TECHNIQUES" International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 10, October 2012 ISSN: 2277 128X
[33] P. Veeramuthu, R. Periyasamy, V. Sugasini, "Analysis of Student Result Using Clustering Techniques", International Journal of Computer Science and Information Technologies, Vol. 5, pp. 5092-5094, 2014..
[34] A. Dutt, S. Aghabozrgi, M. A. B. Ismail, and H. Mahroeian, "Clustering Algorithms Applied in Educational Data Mining" , in International Journal of Information and Electronics Engineering, Vol. 5, pp. 105-108, 2015.
[35] M. Goyal and R. Vohra, "Applications of Data Mining in Higher Education", International Journal of Computer Science Issues, Vol. 9, pp. 114-120, 2012
[36] K. Prasada Rao, M.V.P. Chandra Sekhara Rao, B. Ramesh, " Predicting Learning Behavior of Students using Classification Techniques", International Journal of Computer Applications, Vol. 139, pp. 15-19, 2016.

[37] S. Anupama Kumar and M. N. Vijayalakshmi "RELEVANCE OF DATA MINING TECHNIQUES IN EDIFICATION SECTOR" International Journal of Machine Learning and Computing, Vol. 3, No. 1, February 2013
[38] Mrs. Bharati M. Ramageri "DATA MINING TECHNIQUES AND APPLICATIONS" Indian Journal of Computer Science and Engineering Vol. 1 No. 4 301-305
[39] Amandeep Kaur Mann, Navneet Kaur" SURVEY PAPER ON CLUSTERING TECHNIQUES" International Journal of Science, Engineering and Technology Research (IJSETR) Volume 2, Issue 4, April 2013
[40] Er. Arpit Gupta 1 ,Er.Ankit Gupta 2,Er. Amit Mishra 3" RESEARCH PAPER ON CLUSTER TECHNIQUES OF DATA VARIATIONS" International Journal of Advance Technology & Engineering Research (IJATER)
[41] Gurpreet Singh. Prof, Karan Jamla. M.T, (2018). Implementation & Analysis of Clus-tering Techniques in Bioinformatics: Cancer Research. International Journal of Engi-neering Science and Computing, Volume 8 Issue No.7, July 2018.
[42] Mirmozaffari, Mirpouya, Alireza Alinezhad, and Azadeh Gilanpour. "Data Mining Apriori Algorithm for Heart Disease Prediction." Int'l Journal of Computing, Communications & Instrumentation Engg (IJCCIE) 4.1 (2017).
[43] Shraddha Shukla and Naganna S. "A Review ON K-means DATA Clustering APPROACH" International Journal of Information & Computation Technology. ISSN 0974-2239 Volume 4, Number 17 (2014), pp. 1847-1860
[44] Sardar, Tanvir Habib and Ansari, Zahid (2018) "An analysis of MapReduce efficiency in document clustering using parallel K-means algorithm," Future Computing and Informatics Journal: Vol. 3 : Iss. 2 , Article 7
[45] Kapil Joshi, Himanshu Gupta, Prashant Chaudhary, Punit Sharma," Survey on Different enhanced K-means Clustering Algorithm", International Journal Of Engineering Trends And Technology, Vol. 27 ,No. 4-September 2015.
[46] Don Kulasiri, Sijia Liu, Philip K. Maini and RadekErban, ―DifFUZZY: A fuzzy clustering algorithm for complex data sets , International Journal of Computational Intelligence in Bioinformatics and Systems Biology vol.1, no.4,pp. 402-417, 2010
[47] Pallavi Purohit ―A new Efficient Approach towards k-means Clustering Algorithm ,International journal of Computer Applications,Vol 65-no 11,march 2013
[48] Ahmed M. Fahim, Abdel-Badeeh M.Salem, Mohamed A. Ramadan, Efficient enhanced k-means clustering algorithm, Article in Journal of Zhejiang University - Science A: Applied Physics & Engineering · January 2006
[49] Mirpouya Mirmozaffari , Alireza Alinezhad , and Azadeh Gilanpour "Data Mining Classification Algorithms for Heart Disease Prediction" Int'l Journal of Computing, Communications & Instrumentation Engg. (IJCCIE) Vol. 4, Issue 1 (2017) ISSN 2349-1469 EISSN 2349-1477
[50] B. Bahrami, and M. H. Shirvani, "Prediction and Diagnosis of Heart Disease by Data Mining Techniques," Journal of Multidisciplinary Engineering Science and Technology (JMEST), ISSN: 3159-0040, Vol. 2, Issue 2, February 2015.
[51] M. Kumari, R. Vohra and A. Arora, "Prediction of Diabetes Using Bayesian Network," International Journal of Computer Science and Information Technologies, Vol. 5 (4), 5174-5178, 2014.
[52] A. K. Sen, S. B. Patel, and D. P. Shukla, "A Data Mining Technique for Prediction of Coronary Heart Disease Using Neuro-Fuzzy Integrated Approach Two Level," International Journal of Engineering and Computer Science, Vol. 2, No. 9, pp. 1663–1671, 2013.
[53] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, pp. 2825–2830 (2011).
[54] Z. Pawlak, "Rough sets", International Journal of Computer and Information Science, vol. 11, no. 5, pp. 341-356. 1982.
[55] Yadav C, Wang S, Kumar M. "Algorithms and approaches to handle large data sets - A survey". International Journal of Computer Science and Network. 2013; 2(3):1–5.