Automatic BIRCH thresholding with features transformation for hierarchical breast cancer clustering

Ahmad Alzu'bi[1], Maysarah Barham[2]
[1]Department of Computer Science, Jordan University of Science and Technology, Irbid, Jordan
[2]Department of Computer Science, Middle East University, Amman, Jordan

Article Info

Article history:

Received Apr 29, 2021
Revised Jul 19, 2021
Accepted Aug 12, 2021

Keywords:

Automatic thresholding
Balanced iterative reducing and clustering using hierarchies
Breast cancer
Computer-aided diagnosis
Hierarchical clustering

ABSTRACT

Breast cancer is one of the most common diseases diagnosed in women over the world. The balanced iterative reducing and clustering using hierarchies (BIRCH) has been widely used in many applications. However, clustering the patient records and selecting an optimal threshold for the hierarchical clusters still a challenging task. In addition, the existing BIRCH is sensitive to the order of data records and influenced by many numerical and functional parameters. Therefore, this paper proposes a unique BIRCH-based algorithm for breast cancer clustering. We aim at transforming the medical records using the breast screening features into sub-clusters to group the subject cases into malignant or benign clusters. The basic BIRCH clustering is firstly fed by a set of normalized features then we automate the threshold initialization to enhance the tree-based sub-clustering procedure. Additionally, we present a thorough analysis on the performance impact of tuning BIRCH with various relevant linkage functions and similarity measures. Two datasets of the standard breast cancer wisconsin (BCW) benchmarking collection are used to evaluate our algorithm. The experimental results show a clustering accuracy of 97.7% in 0.0004 seconds only, thereby confirming the efficiency of the proposed method in clustering the patient records and making timely decisions.

Corresponding Author:

Ahmad Alzu'bi
Department of Computer Science, Jordan University of Science and Technology
Irbid 22110, Jordan
Email: agalzubi@just.edu.jo

1. INTRODUCTION

Extracting meaningful information from the medical records to make proper early decisions is a demanding task and should be investigated meticulously. Many challenges are usually encountered in the procedure of diseases diagnosis and treatment due to the large amount of medical data generated by health monitoring systems and equipments. Among the most challenging factors are the diversity of disease characteristics, heterogeneity of treatment, complexity of data collection and processing, and interpretation of medical diagnostics generated from various media [1]–[3], i.e., audio, visual, image, and text content.

Clustering is a simple and yet efficient unsupervised approache that assigns the data subjects into high similar groups, i.e., clusters. However, handling the underlying diversity of clustering analysis, objectives, terms, and assumptions of various clustering algorithms can be daunting [4], [5]. Therefore, there is a demand to neatly determine a correct congruence between the aggregation algorithms and the biomedical applications. Additionally, an adequate approach of data selection and clustering is crucial in the medical diagnosis, which usually requires a relevant knowledge and prior domain expertise.

Clustering feature tree (CF-tree) is one of the efficient and scalable data clustering methods based on a memory data structure and serves as a summary of data distribution. The CF-tree is the core mechanism of the hierarchical balanced iterative reducing and clustering using hierarchies (BIRCH) [6]. BIRCH can handle multi-dimensional data points dynamically or incrementally, and it ordinarily produces good clustering results in few data scans. Among the common hierarchical clustering approaches, BIRCH is effective in solving many real-life applications such as constructing iterative and interactive classifiers and forming codebooks for image retrieval and segmentation [7]–[9]. A clustering feature (CF) is represented as a node in BIRCH clustering tree, which demonstrates the underlying cluster of a specific point or multiple points. BIRCH considers the closeset points as one group where the CFs demonstrates this scale of abstraction. Generally, BIRCH method includes scanning the subjects to construct an in-memory features tree, rebuilding smaller CF trees, performing a global clustering, and clusters refinement.

However, the downside of BIRCH algorithm is the sensitivity to the order of data records in the numerical attributes. Its performance also depends on several parameters including the branching factor Br, threshold T, and cluster count k. In BIRCH, a height-balanced CF tree of hierarchical clusters is built. A cluster is represented as a node where the leafs are the actual clusters. The branching factor Br limits the number of node's children. A new data point is added to the leaf cluster if the cluster radius does not exceed a defined threshold T. Otherwise, the new data point is assigned into a new empty cluster.

A proper threshold selection is necessary to improve the accuracy of BIRCH, which also affects the size of clusters. Moreover, the BIRCH performance is largely influenced by the linkage methods, that used to construct the sub-clusters tree, and by the distance measures used to calculate the distance between the data points and the cluster centroids. Zhang *et al.* [6] have shown the superiority of BIRCH compared to the clustering large applications based on RANdomized search (CLARANS) [10] method. Ismael *et al.* [11] have also attempted to address the shortcomings of BIRCH using a single threshold initialization. The CF-tree is built with the restriction that the leaf entries must use a uniform threshold T while different thresholds are used to reconstruct the CF tree. Several studies [12]–[18] have also highlighted the impact of using multiple thresholds or single threshold either in BIRCH or other hierarchical clustering. Many research efforts have been devoted for clustering the breast cancer records. Vijayarani and Jothi [19] have evaluated the clustering performance and the outlier detection accuracy. They implemented the aggregation process in data flows and examined the extreme values in data flows using BIRCH with CLARANS and BIRCH with k-means. Chowdhary *et al.* [20] have investigated a hybrid fuzzy method to diagnose the breast cancer using the C-means clustering and support vector machines (SVM) algorithm. Lavanya and Palaniswami [21] have proposed assigning the data subjects to different classes using the principle of majority weighted minority oversampling technique.

In this paper, an improved BIRCH variant is proposed by a three-fold paradigm: attributes preprocessing, threshold initialization, and evaluating several linkage and similarity measures. We aim at building an efficient hierarchical clustering to diagnose the patients of breast cancer, which maintains the time and storage constraints. We also investigate the impact of outlier patterns on the performance of BIRCH in terms of clustering accuracy and runtime complexity. The standard benchmarking datasets, breast cancer wisconsin [22] and breast cancer wisconsin (diagnostic) [23], are used to evaluate the proposed approach. The remaining part of this paper is organized as follows: section 2 illustrates this work methodology and the proposed algorithms; section 3 presents the experimental results with detailed discussion and comparisons; and section 4 concludes this paper.

2. RESEARCH METHOD

This section presents the conventional basic BIRCH algorithm, the proposed BIRCH-based clustering framework, datasets and performance evaluation protocol.

2.1. The hierarchical birch

The basic BIRCH algorithm consists of four main phases [6], [24]: i) loading data points into a CF tree to conduct an initial scanning on the dataset; ii) optionaly, building a smaller CF tree by condensing any resizable data or merging the crowded sub-clusters; iii) applying a global clustering on the CF data points through another clustering method, e.g., k-means; and iv) refining the clusters by correcting any inaccuracies in the CF tree. BIRCH requires initializing the number of branches on the CF-leaf and CF-non leaf. The location of a data point, i.e., patient record, is compared to the location of each clustering feature at the root node and passes it to the closest root node. The following are the essential parameters that largely influence the performance of BIRCH:

- CF features: the number of data points (N) for a given data point (x), the linear sum of data points (LS), and the square sum of data points (SS). The latter two parameters are defined as (1) and (2):

$$LS = \sum_{I=1}^{N} x_i \tag{1}$$

$$SS = \sum_{I=1}^{N} x_i^2 \tag{2}$$

- Centroid: It is derived from a CF and defined as (3):

$$x_0 = \sum_{i=1}^{N} x_i = \frac{LS}{N} \tag{3}$$

- Radius (R): the *average* distance from any cluster data point to its centroid, and it is defined as (4):

$$R = \sqrt{\frac{\sum_{i=1}^{N}(x_i - x_0)^2}{N}} = \sqrt{\frac{N * SS - 2 * LS^2 + N * LS}{N^2}} \tag{4}$$

- Diameter (D): the square root of the average mean squared distance between all pairs of the cluster datapoints, and it is defined as (5):

$$D = \sqrt{\frac{\sum_{i=1}^{N}\sum_{J=1}^{N}(x_i - x_j)^2}{N}} = \sqrt{\frac{2N*SS-2LS^2}{N(N-1)}} \tag{5}$$

If two clusters, *C1* and *C2*, are merged then the constructed CF would be the summation of corresponding parameters in the clusters, which is defined as (6):

$$CF = CF1 + CF2 = (N1 + N2, LS1 + LS2, SS1 + ss2 \tag{6}$$

2.2. The framework of improved BIRCH

Figure 1 demonstrates the sequence of phases involved in the proposed BIRCH for breast cancer clustering, and each phase is consecutively illustrated throughout this paper. Firstly, we will use the benchmarking medical datasets to preprocess the patient records and features by selecting the most relevant features and fitting them to the corresponding clusters labels (benign and malignant). Secondly, the threshold value is automatically initialized using a three-steps function that select a random subset of features. Any data outliers are also eliminated by rescaling the patient features. Data features are rescaled, i.e., normalized, into a new data space using the minimum/maximum values of all patients' records. Thirdly, we apply an ablation study on numerous linkage methods and similarity distance metrics. Finally, all the patients' records are predicted and assigned into a proper cluster.

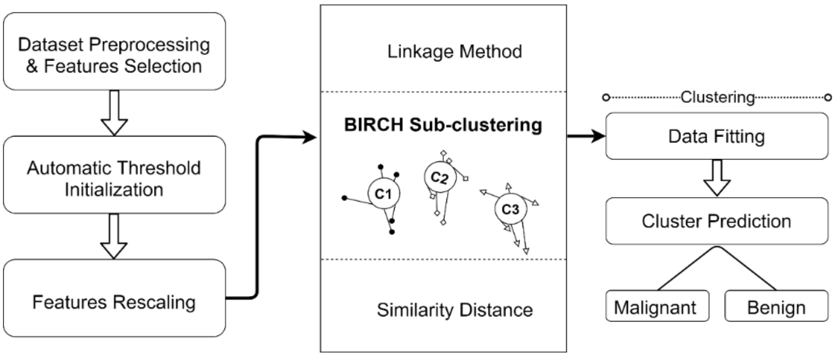


Figure 1. A graphical depiction of the main phases involved in the improved BIRCH algorithm

2.3. Data preprocessing

Data records are preprocessed by selecting the most relevant features and fitting them into the corresponding clusters labels, i.e., benign and malignant. Additionally, any outliers are detected and eliminated using features rescale. Our procedure of data preprocessing consists of two main phases: features selection and features rescale. A proper features selection facilitates the construction of clusters and reduces the data space, hence requiring less processing and storage. In our framework, the patient record is formulated into a vector of features $x=[x_1,...,x_i]$. However, the redundant records are omitted using a min-max normalization. Then, the data are split into two groups where $x=[x_1,...,x_i]$ represents the patient features and $y=[y_1,...,y_i]$ represents the cluster, i.e., benign or malignant.

We use the random sampling to collect data from the patient dataset in which all the records have an equal opportunity of being chosen. The size of selected data is empirically set to 50% of the whole records. Then, we pass this randomly selected sample to the automatic thresholding function. Finally, the matrix elements (F) are rescaled to generate the normalized features (F_n) as (7):

$$Fn = scale\ [F, inputmin, V(Min), inputmax, V(Max)] \qquad (7)$$

where, F represents the input features, $V(Min)$ is the vector of minimum feature values, $V(Max)$ is the vector of maximum feature values, $input_{max}$ is the upper bounding limit of normalization interval, and $input_{min}$ is the lower bounding limit of normalization interval. This procedure projects the features into a new space within $V(Min)$ and $V(Max)$. Therefore, it rescales according to the size of input features that corresponds to the bounding limits, i.e., $input_{min}$ and $input_{max}$.

2.4. Automatic threshold initialization

BIRCH clustering builds the CF-tree in which the leaf entries must meet a fixed threshold, but this usually produces a poor clustering quality. In our work, the threshold value is initialized automatically to improve the clustering accuracy and speed. Therefore, the threshold T is used in the CF-Leaf to store any changes on the used threshold. Our thresholding algorithm is inspired by the work introduced by Ridler and Calvard [25] in which they assign a threshold to separate the image pixels into classes. Correspondingly, we construct a matrix of patient features and generate a random optimal threshold. In BIRCH, each data point is assigned to the closest CF-leaf if the radius does not exceed the threshold T. Otherwise, this point is assigned to a new empty leaf. In contrast, we propose that the new data point that exceeds the threshold should be initialized automatically, thereby enlarging the radius scale on the leaf nodes and reducing the parent split. This process includes three steps.

Step 1: Segments the feature matrices into two parts using an initial random threshold, i.e., $T(1)$, as shown in algorithm 1.

Algorithm 1. Threshold initialization and features split

```
Input: sample points from dataset (I) selected randomly
Output: initial threshold
Begin
   N: random sample of features, I: features
   Counts: summation of elements, T: threshold
   cuSum1: cumulative summation, i=1    //counter for T
   1.1  Find the mean of N features
          T(1)=mean(I)
          Counts=features matrix(I)
          calculate the cuSum1 of counts
   1.2   Round the result
          T(i)=sum(N.*counts)/cuSum1(end).
end
```

Step 2: Calculates a new threshold by averaging the means of two samples, as shown in algorithm 2.

Algorithm 2. Calculating the mean values.

```
Input: the mean of features
Output: a new updated threshold
begin
  MBT: mean below the current threshold
  MAT: mean above the current threshold
  Counts: summation of elements
  N: random sample of features, T: threshold
  2.1 calculate MBT
      MBT=sum(N(N<=T(i))*counts(N<=T(i)))/cuSum2(end)
  2.2 calculate MAT
      MAT=sum(N(N>T(i))*counts(N>T(i)))/cuSum3(end)
  2.3 T(i)=(MAT+MBT)/2                       //new threshold
end
```

Step 3: Repeats step 2 until the threshold value does not change anymore, as shown in algorithm 3.

Algorithm 3. Threshold selection

```
Input: threshold, Output: optimal threshold
begin
  3.1 repeat Algorithm 2
       while T(i)~=T(i-1)
       T(i)=features matrix
       While ABS(newT(i)-oldT(i-1))=1 do:
  3.2  cuSum2=cumsum(counts(N<=T(i)))
       MBT=sum(N(N<=T(i))*counts(N<=T(i)))/cuSum2(end)
       cuSum3=cumsum(counts(N>T(i)))
       MAT=sum(N(N>T(i))*counts(N>T(i)))/cuSum3(end)
       i=i+1
  3.3 if T(i)~=T(i-1), repeat step 3.2
       T(i)=(MAT+MBT)/2
       T(i)=features matrixes
     end while
end
```

2.5. Linkage methods and similarity distances

BIRCH calculates the distance between data points to join them into clusters iteratively. In binary clustering, each cluster is shaped by many observations and join methods on the data points and clusters. Therefore, we consider various linkage methods in our experiments as defined in Table 1. The cluster *r* is a join of clusters *p* and *q*, n_r is the number of subjects in *r*, and x_{ri} is the *i*th subject in *r*. Table 2 also summarizes all the standard similarity distance metrics studied in this work.

Table 1. The linkage methods examined in the proposed approach

Method	Description
Single	It is known as nearest neighbor, and employs the smallest distance between objects in two clusters. $$d(r,s) = min\left(dist\left(x_{ri}, x_{sj}\right)\right), i\epsilon(i,\ldots,n_r), j\epsilon(1,\ldots,n_s) \quad (8)$$
Complete	It is known as farthest neighbor, and employs the largest distance between objects in two clusters. $$d(r,s) = max\left(dist\left(x_{ri}, x_{sj}\right)\right), i\epsilon(i,\ldots,n_r), j\epsilon(1,\ldots,n_s) \quad (9)$$
Ward	It calculates the weighted squared Euclidean distance between the centroids of two clusters $$d(r,s) = \sqrt{\frac{2n_r n_s}{(n_r - n_s)}}\,\|\overline{x_r} - \bar{x}_s\|_2 \quad (10)$$ Where: $\|\overline{x_r} - \bar{x}_s\|_2$ is the eculidean distance, $\bar{x}_r$ and $\bar{x}_s$ are the centroids of clusters *r* and *s*. n_r and n_s are the number of elements in clusters *r* and *s*.
Centroid	It calculates the square of Euclidean distance between the centroids of two clusters $$d(r,s) = \left\|\bar{x}_r - \bar{x}_s\right\|_2 \quad (11)$$ where $$\bar{x}_r = \frac{1}{n}\sum_{i=1}^{n_r} \bar{x}_{ri} \quad (12)$$
Average	It calculates the average distance between all pairs of objects in two clusters. $$d(r,s) = \frac{1}{n_r n_s}\sum_{i=1}^{n_r}\sum_{j=1}^{n_s} dist(x_{ri}, x_{sj}) \quad (13)$$
Median	It employs the Euclidean distance between the weighted centroids of the two clusters $\tilde{x}_r$ and $\tilde{x}_s$. $$d(r,s) = \left\|\tilde{x}_r - \tilde{x}_s\right\|_2 \quad (14)$$

Table 2. Similarity distance metrics

Metric	Description	
Euclidean (p=2) Cityblock (p=1) Chebychev (p=∞)	$d(r,s)=\sqrt[p]{\sum_{i=1}^{n}\lvert x_{ri}-x_{si}\rvert^{p}}$	(15)
Squared Euclidean	Squared Euclidean that is usually used for regression analysis. $d(r,s)=\sum_{i=1}^{n}\lvert x_{ri}-x_{si}\rvert^{2}$	(16)
StdEuclidean	Standardized Euclidean that divides each squared discrepancy between attributes by the sample size. $d(r,s)=\sqrt{\sum_{i=1}^{n}\frac{\left(x_{ri}-x_{si}\right)^{2}}{n}}$	(17)
Mahalanobis	The distance between the data point and the sample distribution using the covariance matrix, where s_i^2 is the standard deviation. $d(r,s)=\sqrt{\sum_{i=1}^{n}\frac{\left(x_{ri}-x_{si}\right)^{2}}{s_i^{2}}}$	(18)

2.6. Datasets and performance metrics

Breast cancer wisconsin dataset (BCW) [22] consists of 11 attributes and 699 instances divided into different partitions. It includes the following features: record ID, clump thickness, the uniformity of cell, shape and size, marginal adhesion, normal nuclei, bare nuclei, epithelial cell size, bland chromatin, mitoses, and cluster label, i.e., 2 for benign and 4 for malignant. The patient ID is excluded from our experiments. Breast cancer wisconsin (diagnostic) dataset [23] consists of 31 attributes and 569 instances divided into different partitions. It includes the cluster label, i.e., M for malignant and B for benign, and 10 features calculated for each cell nucleus as follows: perimeter, area, radius (mean of distances), texture, smoothness (radius variation), compactness (*perimeter*^{2}/*area*-1.0), concavity, concave points, symmetry, and fractal dimension.

The proposed BIRCH variant is evaluated by the following performance metrics: true positives (TP), false positives (FP), false negatives (FN), true negatives (TN), accuracy, precision and recall. We also use F-measure (F-score) to make the precision and recall comparable in place of arithmetic mean by punishing the extreme values more. Additionally, fowlkes-mallows index (F*m*-index) is used to find the dissimilarity between the final clusters.

3. RESULTS AND DISCUSSION

In this section, we demonstrate and discuss the experimental results obtained by the improved BIRCH clustering. Thresholds are automatically initialized after processing the features of medical records, and we also present the results obtained by the basic and improved BIRCH with relevant comparisons.

3.1. Clustering results on BCW dataset

Firstly, we discuss the clustering results obtained by the original BIRCH using a range of fixed thresholds: 0.2, 0.5, and 0.9. These thresholds are manually assigned within the range {0-1}. Table 3 summarizes the best result recorded using a range of linkage and distance measures under a thorough experiments. It can be observed that the basic BIRCH achieved the best clustering performance using the ward linkage and Euclidean similarity distance. It is also performing with a threshold 0.2 better than other threshold values considered in our experiments, i.e., 0.5 and 0.9.

On the other hand, our BIRCH variant outperforms the basic BIRCH over all methods using a randomly initialized threshold. Table 4 shows the clustering results of improved BIRCH on the BCW dataset. The improved BIRCH achieves 97.7% of clustering accuracy and improves the accuracy of the basic BIRCH by 4% and the recall by 6%. The accuracy results confirm the superiority of the improved BIRCH clustering using various linkage and similarity distances. The basic BIRCH only outperforms the improved version using the centroid linkage with Seuclidean similarity distance. However, both BIRCH versions reported the best performance using ward linkage and Euclidean. In terms of speed, the improved BIRCH is obviously faster than the basic BIRCH under all the experimental configurations. It takes an average time of 0.0006 seconds to complete the clustering process on the BCW dataset compared to an average time of 0.3723 seconds in the basic BIRCH, which is also fast on the BCW (diagnostic) dataset.

Table 3. Clustering results on BCW dataset using the basic BIRCH

Linkage	Distance	Time (s)	Recall	TP	TN	FP	FN	F*m*	Accuracy	Threshold
Ward	Euclidean	0.13	0.93	0.51	0.43	0.02	0.04	0.94	0.936	0.2
	Seuclidean	0.32	0.93	0.51	0.42	0.03	0.04	0.94	0.931	0.9
	SqrEuclidean	0.10	0.92	0.50	0.38	0.08	0.04	0.90	0.884	0.2
Centroid	Euclidean	0.38	0.99	0.55	0.00	0.45	0.00	0.74	0.548	0.2
	Seuclidean	0.11	0.43	0.51	0.42	0.03	0.03	0.93	0.928	0.2
	SqrEuclidean	0.98	0.93	0.50	0.41	0.05	0.04	0.92	0.915	0.2
Average	Euclidean	0.58	0.92	0.52	0.01	0.06	0.01	0.90	0.889	0.2
	Seuclidean	0.007	0.94	0.52	0.07	0.38	0.03	0.74	0.585	0.5
	SqrEuclidean	0.94	0.92	0.50	0.38	0.07	0.05	0.90	0.886	0.2
Single	Euclidean	0.01	0.99	0.54	0.01	0.44	0.01	0.74	0.548	0.2
	Seuclidean	0.001	0.99	0.55	0.01	0.42	0.02	0.74	0.548	0.2
	SqrEuclidean	0.91	0.99	0.54	0.01	0.44	0.01	0.74	0.548	0.2

Table 4. Clustering results on BCW dataset using the improved BIRCH

Linkage	Distance	Time (s)	Recall	TP	TN	FP	FN	F*m*	Accuracy	Threshold
Ward	Euclidean	0.0004	0.99	0.52	0.44	0.02	0.04	0.96	0.977	0.38
	Seuclidean	0.0002	0.99	0.51	0.43	0.03	0.03	0.95	0.949	0.48
	SqrEuclidean	0.0002	1.00	0.47	0.40	0.05	0.07	0.89	0.937	0.47
Centroid	Euclidean	0.0006	1.00	0.55	0.00	0.45	0.00	0.74	0.656	0.44
	Seuclidean	0.0009	1.00	0.54	0.01	0.45	0.01	0.74	0.655	0.48
	SqrEuclidean	0.0010	0.98	0.51	0.43	0.03	0.03	0.94	0.967	0.47
Average	Euclidean	0.0007	0.98	0.51	0.42	0.04	0.03	0.93	0.962	0.44
	Seuclidean	0.0009	0.99	0.52	0.43	0.02	0.03	0.95	0.969	0.38
	SqrEuclidean	0.0005	0.99	0.51	0.43	0.02	0.04	0.94	0.969	0.45
Single	Euclidean	0.0006	1.00	0.53	0.01	0.45	0.01	0.74	0.656	0.44
	Seuclidean	0.0007	1.00	0.55	0.00	0.45	0.00	0.74	0.656	0.44
	SqrEuclidean	0.0008	1.00	0.55	0.00	0.45	0.00	0.74	0.656	0.37

3.2. Clustering results on BCW (diagnosis) dataset

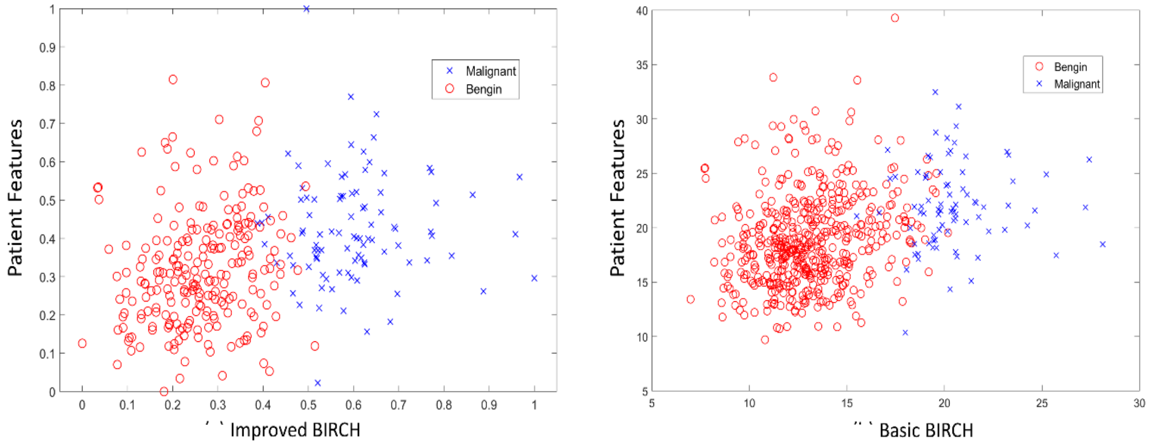
Table 5 summarizes the clustering results obtained after applying the best configuration of the basic and improved BIRCH on the BCW (diagnostic) dataset. Obviously, our BIRCH variant outperforms the basic one by an accuracy of 93.3% compared to 65.5% under the same setups. Also, the average clustering time of the improved BIRCH is about 0.0008 second compared to 0.6424 second taken by the basic BIRCH.

Table 5. Clustering results on the BCW (diagnosis) dataset

Method	Time (s)	Recall	TP	TN	FP	FN	F*m*	Accuracy	Threshold
Basic BIRCH	0.6420	0.873	0.465	0.189	0.278	0.067	0.739	0.655	0.200
Improved BIRCH	0.0008	0.969	0.478	0.398	0.070	0.054	0.884	0.933	0.561

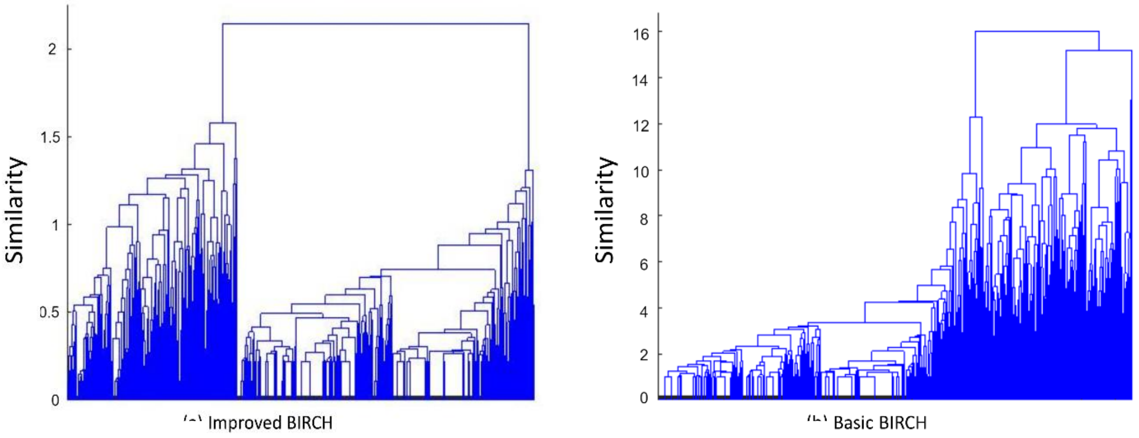
3.3. Clustering hierarchical relationship

Figures 2(a) and 2(b) depict the patients' clusters of breast cancer using the improved BIRCH compared to the basic version obtained by the best configuration, i.e., ward linkage and Euclidean distance. As shown in Figure 2(a), two clusters (benign and malignant) of breast cancer records are represented by rescaled features in the improved BIRCH and optimally predicted using a random threshold of 0.38. It can be observed that the overlapping features at the cluster borderlines are minimized by our BIRCH variant. The BIRCH clusters are also visualized using the dendrogram [26] which depicts the hierarchical relationship between the dataset records, i.e., cluster objects. It is used as common representation of the hierarchical clustering, as shown in Figures 3(a) and 3(b). All the data points are shown at the bottom of the dendrogram. Each point or subject is assigned to separate clusters and any two close clusters are merged to shape a final cluster at the top. The height in the dendrogram is the similarity distance between two clusters in the data space. The highest mean and median *Fm* scores were obtained for the basic BIRCH and improved BIRCH using a threshold 0.2 and a random threshold 0.38, respectively. It can be observed that the clusters merge in the improved BIRCH is better than the basic one in showing which clusters are very similar.



(a) (b)

Figure 2. A depiction of the refined clusters, (a) improved BIRCH and (b) basic BIRCH



(a) (b)

Figure 3. Th dendrogram plots of clustering, (a) improved BIRCH and (b) basic BIRCH

3.4. Precision and recall evaluation

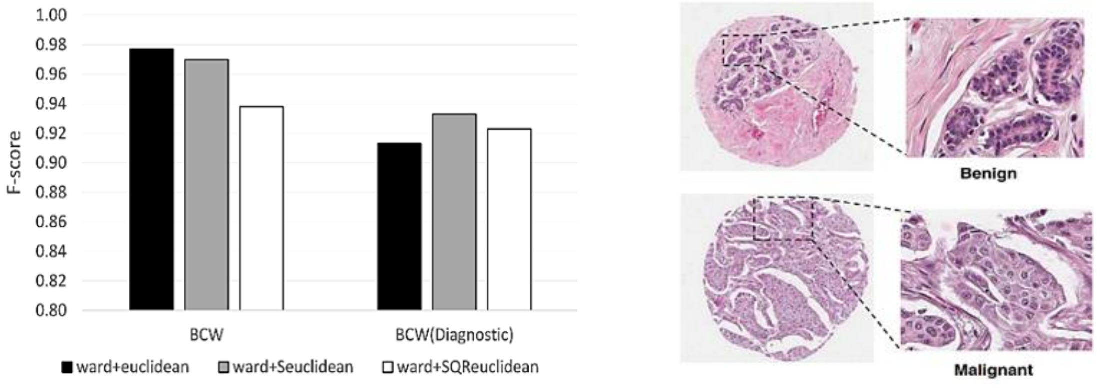
We consider here how precision determines the clinical sensitivity, i.e., fraction of true positives to all with breast cancer, and the clinical specificity, i.e., fraction of true negatives to all without breast cancer. Table 6 summarizes the results of improved BIRCH on the datasets. The reported results are approaching 100% precision and 100% recall on both datasets, which confirms the stability of clustering algorithm. We also underline the importance of measuring the recall and precision at the same time using the F-score [27], as shown on Figure 4(a). Obviously, the improved BIRCH achieves higher F-scores than the basic BIRCH. A sample of breast tumors diagnosed as benign or malignant is demonstrated in Figure 4(b).

Table 6. Precision-recall results

	Precision		Recall	
	BCW	BCWD	BCW	BCWD
Ward+Euclidean	0.992	0.977	0.996	0.989
Ward+Seuclidean	0.992	0.944	0.995	0.969
Ward+ SqrEuclidean	1.000	0.985	1.000	0.997

3.5. Comparisons with related works

As shown in Table 7, we compare the performance of our proposed BIRCH algorithm in terms of accuracy, precision, and recall with the most two related clustering works examined on the same dataset, i.e., BCW. It can be obviously observed that our BIRCH algorithm outperforms the other approaches in all the performance metrics, which emphasizes its high capability in clustering the breast cancer records.



(a) (b)

Figure 4. The performance of improved BIRCH in terms of F-score, (a) F-score results on BCW and diagnostic and (b) diagnosed breast tissues [28]

Table 7. Performance comparison with the related approaches

Method	Accuracy	Precision	Recall
BIRCH+K-mean [19]	0.704	0.748	0.768
BIRCH+CLARNS [19]	0.764	0.760	0.760
BIRCH+Mwmote [21]	0.969	0.940	0.970
This paper	0.977	0.995	0.991

4. CONCLUSION

In this paper, we have improved the capability of the hierarchical BIRCH aggregation algorithm in clustering the medical records of breast cancer patients. The experimental results emphasize the superiority of the improved BIRCH over the basic BIRCH with efficient features selection, data rescaling, automatic threshold initialization, linkage methods and distances metrics. We demonstrated that a proper data preprocessing improves the BIRCH performance. Additionally, our proposed automatic thresholding largely increases the quality of generated clusters. Also, the impact of binding methods on the complexity of tree subgroups, i.e., subclustering, is highlighted. We achieved a clustering accuracy of 97.7% with discriminating clusters better than the original BIRCH. In future, the proposed BIRCH could be further optimized by passing the cluster centroids to another clustering algorithm, e.g., *k*-means. This procedure could be adopted in a sequential or parallel manner, i.e., various representations.

REFERENCES

[1] M. Z. Nayef AL-Dabagh, "Automated tumor segmentation in MR brain image using fuzzy c-means clustering and seeded region methodology," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 10, no. 2, pp. 284–290, Jun. 2021, doi: 10.11591/ijai.v10.i2.pp284-290.

[2] S. Tongbram, B. A. Shimray, and L. S. Singh, "Segmentation of image based on k-means and modified subtractive clustering," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 22, no. 3, pp. 1396–1403, Jun. 2021, doi: 10.11591/ijeecs.v22.i3.pp1396-1403.

[3] A. Alzu'bi and A. Abuarqoub, "Deep learning model with low-dimensional random projection for large-scale image search," *Engineering Science and Technology, an International Journal*, vol. 23, no. 4, pp. 911–920, Aug. 2020, doi: 10.1016/j.jestch.2019.12.004.

[4] A. Pourkashani, A. Shahbahrami, and A. Akoushideh, "Copy-move forgery detection using convolutional neural network and K-mean clustering," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 3, pp. 2604–2612, Jun. 2021, doi: 10.11591/ijece.v11i3.pp2604-2612.

[5] Q. Duan, Y. L. Yang, and Y. Li, "Rough K-modes clustering algorithm based on entropy," *IAENG International Journal of Computer Science*, vol. 44, no. 1, pp. 13–18, 2017.

[6] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An efficient data clustering method for very large databases," *ACM SIGMOD Record*, vol. 25, no. 2, pp. 103–114, Jun. 1996, doi: 10.1145/235968.233324.

[7] J. Park *et al.*, "Automatic segmentation of brachial artery based on fuzzy C-Means pixel clustering from ultrasound images," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, no. 2, pp. 638–643, Apr. 2018, doi: 10.11591/ijece.v8i2.pp638-643.

[8] J. Chang, L. Wang, G. Meng, S. Xiang, and C. Pan, “Deep adaptive image clustering,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 5880–5888, doi: 10.1109/ICCV.2017.626.

[9] E. mehdi Cherrat, R. Alaoui, and H. Bouzahir, “Improving of fingerprint segmentation images based on K-MEANS and DBSCAN clustering,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 4, pp. 2425–2432, Aug. 2019, doi: 10.11591/ijece.v9i4.pp2425-2432.

[10] E. Schubert and P. J. Rousseeuw, “Fast and eager k-medoids clustering: O(k) runtime improvement of the PAM, CLARA, and CLARANS algorithms,” *Information Systems*, vol. 101, no. 2–3, Art. no. 101804, Nov. 2021, doi: 10.1016/j.is.2021.101804.

[11] N. Ismael, M. Alzaalan, and W. Ashour, “Improved multi threshold BIRCH clustering algorithm,” *International Journal of Artificial Intelligence and Applications for Smart Devices*, vol. 2, no. 1, pp. 1–10, 2014.

[12] B. Lorbeer, A. Kosareva, B. Deva, D. Softić, P. Ruppel, and A. Küpper, “A-BIRCH: Automatic threshold estimation for the BIRCH clustering algorithm,” in *Advances in Intelligent Systems and Computing*, Springer International Publishing, 2017, pp. 169–178.

[13] R. K. Owen, N. J. Cooper, T. J. Quinn, R. Lees, and A. J. Sutton, “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*, vol. 99, pp. 64–74, Jul. 2018, doi: 10.1016/j.jclinepi.2018.03.005.

[14] E. P. Barracchia, G. Pio, D. D’Elia, and M. Ceci, “Prediction of new associations between ncRNAs and diseases exploiting multi-type hierarchical clustering,” *BMC Bioinformatics*, vol. 21, no. 1, Art. no. 70, Dec. 2020, doi: 10.1186/s12859-020-3392-2.

[15] G. Santoni, U. Zander, C. Mueller-Dieckmann, G. Leonard, and A. Popov, “Hierarchical clustering for multiple-crystal macromolecular crystallography experiments: the ccCluster program,” *Journal of Applied Crystallography*, vol. 50, no. 6, pp. 1844–1851, Dec. 2017, doi: 10.1107/S1600576717015229.

[16] M. Hu, K. Zeng, Y. Wang, and Y. Guo, “Threshold-based hierarchical clustering for person re-identification,” *Entropy*, vol. 23, no. 5, Art. no. 522, Apr. 2021, doi: 10.3390/e23050522.

[17] W.-B. Xie, Y.-L. Lee, C. Wang, D.-B. Chen, and T. Zhou, “Hierarchical clustering supported by reciprocal nearest neighbors,” *Information Sciences*, vol. 527, pp. 279–292, Jul. 2020, doi: 10.1016/j.ins.2020.04.016.

[18] K. V. Rajkumar, A. Yesubabu, and K. Subrahmanyam, “Fuzzy clustering and fuzzy c-means partition cluster analysis and validation studies on a subset of citescore dataset,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 4, pp. 2760–2770, Aug. 2019, doi: 10.11591/ijece.v9i4.pp2760-2770.

[19] S. Vijayarani and M. P. Jothi, “An efficient clustering algorithm for outlier detection in data streams,” *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 2, no. 9, pp. 3657–3665, 2013.

[20] C. L. Chowdhary, M. Mittal, K. P., P. A. Pattanaik, and Z. Marszalek, “An efficient segmentation and classification system in medical images using intuitionist possibilistic fuzzy C-Mean clustering and fuzzy SVM algorithm,” *Sensors*, vol. 20, no. 14, Art. no. 3903, Jul. 2020, doi: 10.3390/s20143903.

[21] S. Lavanya and S. Palaniswami, “Hierarchical sampling techniques for imbalanced datasets,” *Asian Journal of Information Technology*, vol. 15, no. 16, pp. 2887–2896, 2016.

[22] W. H. Wolberg and O. L. Mangasarian, “Multisurface method of pattern separation for medical diagnosis applied to breast cytology.,” *Proceedings of the National Academy of Sciences*, vol. 87, no. 23, pp. 9193–9196, Dec. 1990, doi: 10.1073/pnas.87.23.9193.

[23] W. H. Wolberg, W. N. Street, and O. L. Mangasarian, “Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates,” *Cancer Letters*, vol. 77, no. 2–3, pp. 163–171, Mar. 1994, doi: 10.1016/0304-3835(94)90099-X.

[24] J. Dong, F. Wang, and B. Yuan, “Accelerating BIRCH for clustering large scale streaming data using CUDA dynamic parallelism,” in *International Conference on Intelligent Data Engineering and Automated Learning*, 2013, pp. 409–416, doi: 10.1007/978-3-642-41278-3_50.

[25] T. W. Ridler and S. Calvard, “Picture thresholding using an iterative selection method,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 8, no. 8, pp. 630–632, 1978, doi: 10.1109/TSMC.1978.4310039.

[26] A. Ciaramella, D. Nardone, and A. Staiano, “Data integration by fuzzy similarity-based hierarchical clustering,” *BMC Bioinformatics*, vol. 21, no. S10, Art. no. 350, Aug. 2020, doi: 10.1186/s12859-020-03567-6.

[27] N. Sevani, I. Hermawan, and W. Jatmiko, “Feature selection based on F-score for enhancing CTG data classification,” in *2019 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom)*, Aug. 2019, pp. 18–22, doi: 10.1109/CYBERNETICSCOM.2019.8875656.

[28] H. Majeed *et al.*, “Breast cancer diagnosis using spatial light interference microscopy,” *Journal of Biomedical Optics*, vol. 20, no. 11, Art. no. 111210, Aug. 2015, doi: 10.1117/1.JBO.20.11.111210.