https://doi.org/10.5281/zenodo.12707510

ISSN: 2096-3246
Volume 55, Issue 03, Aug, 2023

HEART DISEASE IDENTIFICATION METHOD USING MACHINE LEARNINGIN E-HEALTHCARE

TALARI SIVALAKSHMI, KUMMARA RANGA SWAMY, BEDUDHURI.HIMAVANI
Assistant Professor[1,2,3]
raghava.digala@gmail.com,
shivalakshmidinesh@gmail.com, rangaswamy.kumara@gmail.com

department of CSE, Sri Venkateswara Institute of Technology,
N.H 44, Hampapuram, Rapthadu, Anantapuramu, Andhra Pradesh 515722

Keywords:

Clustering ,Distance Matrix , Unsupervised data mining , I-BIRCH

ABSTRACT

Clustering is a crucial step in descriptive statistics and data mining. It is used in many different fields of work, including data categorization and image processing, and has been the subject of much study by many different academics. We present I-BIRCH, an improved balanced iterative reducing and clustering technique that makes use of hierarchies. It works well with massive datasets and is an unsupervised data mining technique. The algorithm begins by clustering data points with a single dimension, and then it moves on to cluster data points with many dimensions in order to get the optimal clustering with a single view of the data. The "noise" (data points that do not form part of the underlying pattern) is something it can manage. Clustering calculations take O(n2) time and use a distance matrix that is O(n2) huge. When mining complicated or massive datasets, this kind of grouping is a necessary component. When there is information about the heart, such an ID, a name, an age, a gender, a delta heart rate, and the CPU date, it is utilised in the population dataset. If the dataset can provide immediate results for mining using I-BIRCH, then the experimental results will demonstrate quadratic time scalability.

https://doi.org/10.5281/zenodo.12707510

ISSN: 2096-3246
Volume 55, Issue 03, Aug, 2023

Introduction

BIG DATA MINING

Big data is a term that describes the large volume of data – both structured and unstructured – that is a major cause for business on a day-to-day basis. Big data can be analyzed for insights that lead to better decisionsand strategic business moves. Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate. Challenges include analysis, capture, search, sharing, storage, transfer, visualization, querying, and information privacy. The term often refers simply to the use of predictive analytic or certain other advanced methods to extract value from data, and seldom to a particular size of data set as shown in fig 1.1. Accuracy in big data may lead to more confident decision making, and better decisions can result in greater operational.

Data mining is the capability of extracting useful information from large datasets or streams of data, due to its volume, variability, and velocity, it was not possible before to do it.

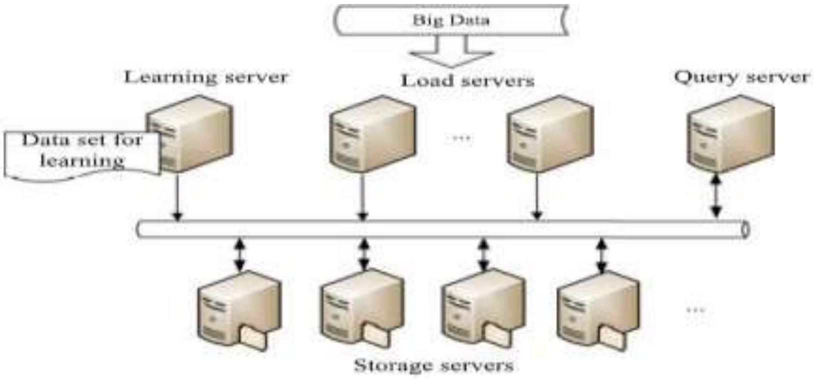


Fig 1.1: Big Data

CHARACTERISTICS OF BIG DATA

Big data can be described by the following characteristics and is shown in table 1.1.

Volume

Variety

It is the quantity of generated and stored data. The size of the data determines the value and potentialinsight- and whether it can actually be considered big data or not.

It is the type and nature of the data. This helps people who analyze it to effectively use the resulting insight.

Velocity

In this context, the speed at which the data is generated and processed to meet the demands and challenges that lie in the path of growth and development.

Variability

Inconsistency of the data set can hamper processes to handle and manage it.

Veracity

The quality of captured data can vary greatly, affecting accurate analysis.

Big data analysis is the process of applying advanced analytics and visualization techniques to large datasets to uncover hidden patterns and unknown correlations for effective decision making.

1.2.1 CHALLENGES

- ComplexityTimeliness
- Privacy
- Space
- Heterogeneity

HETEROGENEITY AND INCOMPLETENESS

Data can be both structured and unstructured. 80% of the data generated by organizations are unstructured. They are highly dynamic and does not have particular format. It may exists in the form of email attachments, images, pdf documents, medical records, X rays, voice mails, graphics, video, audio etc. and they cannot be stored in row/ column format as structured data. Transforming this data to structured format for later analysis is a major challenge in big data mining. So new technologies have to be adopted for dealing with such data.

TIMELINESS

As the size of the data sets to be processed increases, it will take more time to analyse.

In some situations results of the analysis is required immediately. Sopartial result need to developin advance so that a small amount of incremental computation with new data can be used to arrive at a quick determination.

SCALE AND COMPLEXITY

Managing large and rapidly increasing volumes of data is a challenging issue. Traditional software tools are not enough for managing the increasing volumes of data. Data analysis, organization, retrieval and modeling are also challenges due to scalability and complexity of data that needs to be analysed.

LITERATURE REVIEW

EPIDEMIOLOGY AND RISK PROFILE OF HEART FAILURE

With an estimated 5.8 million cases in the US and 23 million cases globally (and increasing), heart failure (HF) is a huge problem in public health. One in five people will get HF at some point in their lives. The age-adjusted incidence of HF may have plateaued, which is encouraging news. However, heart failure still causes a lot of harm and death; in fact, the 5-year mortality rate is comparable to that of several malignancies. The high rates of hospitalisations, readmissions, and outpatient visits caused by HF are a major financial and logistical strain on the healthcare system. In the United States alone, HF costs about $39 billion each year. Heart failure (HF) is a clinical condition rather than a discrete disease; its manifestations might vary according to factors such as patient age, gender, race/ethnicity, LVEF status, and the cause of HF. It is becoming more clear from epidemiological research that there are pathophysiological differences between people with HF with maintained LVEF and those with diminished EF. Both the prevalence and severity of HF may be predicted by a variety of known risk factors, including ischemic heart disease, hypertension, smoking, obesity, diabetes, and many more. Important aspects of HF's epidemiology and risk profile are covered in this Review.

"ANALYSIS AND COMPARISON OF VISUAL APPROACHES TO DATA PREPROCESSING FOR FERTILITY SUCCESS RATE PREDICTION,"

Despite the use of vast information databases for medical diagnostics system evaluation, data extraction from these databases often fails. To find the main connection between the data, there isn't a good enough tool. In such a scenario, data mining techniques are used to extract the essential insights from healthcare data. The subtracted information may be used for an accurate diagnosis and subsequent therapy. Fertility has been a source of mental distress for people all over the globe in recent years due to infertility. Procedures such as IUI, IVF, ICSI, and GIFT are part of the treatment arsenal for infertility. By using the preprocessed data from the database, one may forecast the success rate of the infertility therapy. The current preprocessing approach is described and the accuracy of the post-preprocessing prediction rate is examined in this research. Preprocessing the raw data using the known approaches clearly increases the accuracy by up to 90%. The next step for this project is to find ways to combine several strategies to reach a better outcome.

"AN NEW AI-BASED DECISION SUPPORT SYSTEM FOR THE DIAGNOSIS OF HEART DISEASES,"

Unfortunately, not all hospitals have the technology necessary to diagnose cardiac conditions. This is particularly true in less-populated rural regions, where residents often get less medical attention and assistance. Furthermore, excellent outcomes from medical operations are not always guaranteed by relying just on a physician's intuition and expertise. Unconventional computer-based diagnostic systems are necessary because of medical mistakes and their unfavourable outcomes; these methods cut down on medical deadly errors, improve patient safety, and ultimately save lives. The suggested approach uses ANNs to build a decision-support system that can detect the three most common cardiac conditions: mitral stenosis, aortic stenosis, and ventricular septal defect. In addition, the system offers a promising chance to create a functional screening and testing tool for the detection of cardiac disease, which may greatly aid doctors in making precise diagnoses. The performance and accuracy of the suggested method have been examined via a number of tests that used actual medical data. With a 92% accuracy rate for heart disease categorization, the system's performance and accuracy were found to be satisfactory when compared.

"Cardiovascular disease (CVD) is the leading cause of death in the United States and is responsible for 17% of national health expenditures," said the American Heart Association in its policy statement on the future of CVD in the US. There will likely be a dramatic rise in these expenses due to the ageing population. The American Heart Association has devised a system to forecast the future expenses of treating hypertension, coronary heart disease, heart failure, stroke, and all other CVDs from 2010 to 2030. This is done in order to be prepared for the potential demands of future cardiovascular treatment. For individuals with more than one cardiovascular problem, this approach prevented the duplication of expenses. Projections show that 40.5% of Americans will have a cardiovascular disease by 2030.In order to get clinically useful averages, nonlinear speech analysis algorithms are mapped to a standard metric. The Unified Parkinson's Disease Rating Scale (UPDRS) is the gold standard clinical score for assessing the average severity of Parkinson's disease (PD) symptoms. Currently, UPDRS is established based on the subjective clinical assessment of the patient's capacity to sufficiently manage various activities. Here, we build on earlier research showing that short, self-administered speech tests may objectively evaluate UPDRS to a clinically meaningful degree, all without the patient having to be physically present in the clinic. We process a massive database using a variety of well-known methods for voice signal processing. Current Setup For the purpose of solving the feature selection issue, we also have an existing, innovative, and quick conditional mutual information feature selection technique. In order to improve the classification accuracy and decrease the classification system's execution time, features selection methods are used. It has also been utilised to acquire the best practices of model evaluation and hyper parameter tweaking using the leave one subject out cross-validation approach. The classifiers' performances are evaluated using the performance measurement measures. The features chosen by the features selection techniques have been used to test the classifiers' performance. With classifier support vector machine, the experimental findings demonstrate that the current feature selection technique (FCMIM) is possible for creating a high-level intelligent system to diagnose heart disease. When compared to other approaches that have been used before, the proposed diagnostic system (FCMIM-SVM) demonstrated high accuracy. Also, healthcare providers may quickly and simply use the current method to detect cardiac problems.

CONS: • Suggested diagnostic system; • Hyperparameter tweaking; • Decreased execution time.

SUGGESTED METHODS

As an unsupervised data mining technique, I-BIRCH (Improved Balanced Iterative Reducing and Clustering utilising Hierarchies) enables clustering in Big Data. To describe clusters in general, it makes use of two ideas: the modified leaf clustering features (MLCF) entry and the modified leaf cluster feature tree (MLCF Tree). The grouping of valuable information is shown by the modified leaf clustering feature tree. The algorithm's performance and scalability in clustering big data sets may be enhanced since space is much lower than meta-data collecting and can be kept in memory. It works well for clustering data with both discrete and continuous attributes.

A Clustering Feature is a node in the BIRCH tree. A cluster of one or more points is represented by this little image. The foundational principle of BIRCH is that any set of points that are sufficiently near together should be treated as a single entity.

At this level of abstraction, Clustering Features come through. A vector of three values, CF = (N;LS; SS) is used to hold clustering features. The product of its linear and square sums, as well as the number of points it encompasses (N).

Two factors, the branching factor (B) and the threshold (T), determine the height balance of a CF tree. You may express a non-leaf node as {CFi, childi}, where,i = 1, 2,..., B, and Childi is a reference to the ith child node. The colour of the ith child's subcluster, denoted as Cfi.

The contents of the non-leaf node stand in for all of the sub-clusters, while the node itself represents a cluster. Similarly, for T to be valid, the contents of a leaf node must reflect all of its sub clusters.

A four-stage process is used to implement the BIRCH clustering method. Phase 1 involves retrieving the database and constructing the first CF using the branching factor B and threshold value T. Phase2, which is not mandatory, is when

In order to get a smaller CF tree, the original CF tree would be shrunk. Using either the larger tree from step 2 or the original CF tree, data points are clustered globally in phase 3.

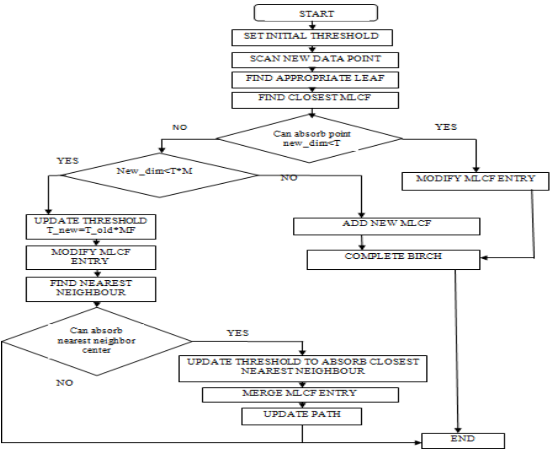
Phase 3 of the method yields good clusters. To enhance the clusters' quality, the clustering procedure would need phase 4 of the method. With a threshold value T, BIRCH Phase 1 execution starts.

ADVANTAGES

- It is local in that each clustering decision is made without scanning all data points and currently existing clusters.
- It exploits the observation that data space is not usually uniformly occupied and not every data point is equally important.
- It makes full use of available memory to derive the finest possible sub-clusters while minimizing I/O costs.
- It is also an incremental method that does not require the whole dataset in advance.

https://doi.org/10.5281/zenodo.12707510

ISSN: 2096-3246
Volume 55, Issue 03, Aug, 2023



YSTEM ARICHTECTURE

IMPLEMENTATION

MODULES

- Data Sets
- Clustering features and CF tree
- Secure CF Tree Insertion
- Time Scalability

MODULES

DESCRIPTION DATA

SETS

To evaluate improved birch algorithm implement the basic birch algorithm and the multi threshold birch algorithm. Values of the generated artificial dataset are used to assess the level of the algorithm success. The real data sets are used in the experiment. Any type of dataset can be dynamically included. The given dataset is read by the reader as shown in fig 4.1 and specified attributes can be mined using BIRCH clustering technique. The default accuracy is about 80% which means there is a big dominant cluster and that is suitable for this work.

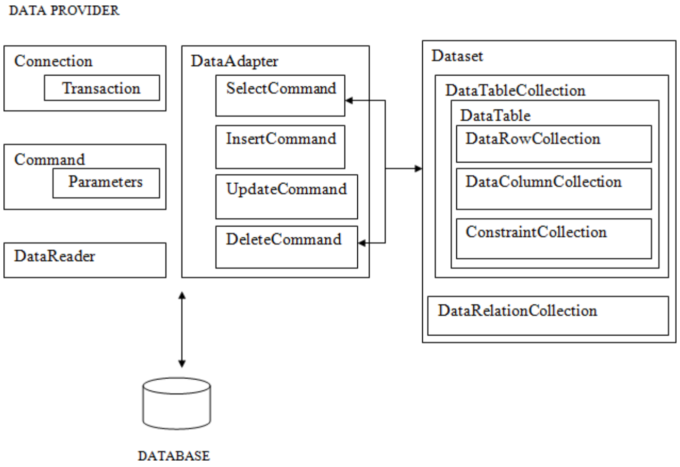
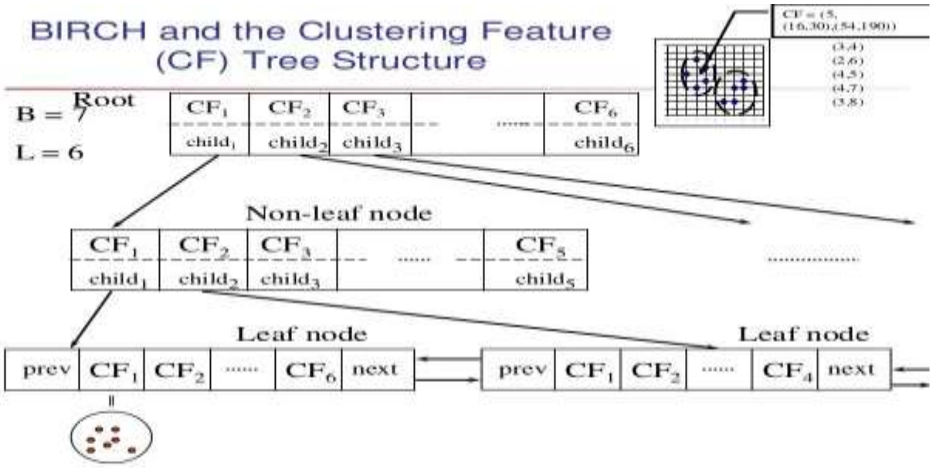


Fig 4.1: Dataset Reader

CLUSTERING FEATURES AND CF TREE

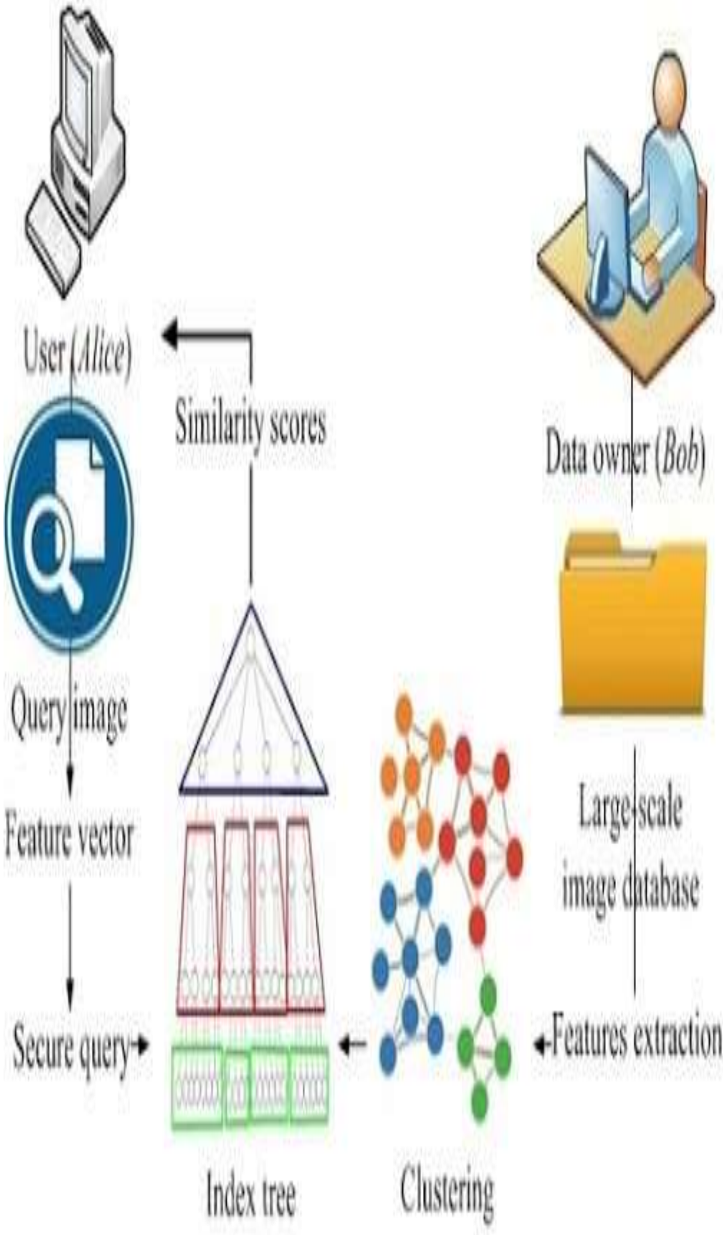
Clustering feature (CF) entry is triple summarizing the information. A CF tree is a height-balanced treewith two parameters: branching factor B and threshold T. Each non leaf node contains at most B entries of the form and it is a pointer to its i-th child node, and CF, is the CF of the sub cluster, represented by this child as in fig



SECURE CF TREE INSERTION

Insert a new Entry securely in the CF tree in the partitioned case. The secure CF Tree insertion procedure is similar to the insertion procedure in single data base case but with invocation of secure protocols. Both the user and data provider learn the tree structure but the entries in the tree will be different for each user.

So user inserts its partitionedshare of *CF* in its tree with a secure query as given



in fig 4.3.

Fig 4.3: Secure CF Tree

TIME SCALABILITY

Two distinct ways of increasing the dataset size are used to test the scalability of I-BIRCH. Increasing Number of Points per Cluster: For each of DS1, DS2 and DS3, create a range of datasets by keeping the generator settings the same except for changing nl and nk to change n, and hence N. The running time for first 3 phases, as well as for all 4 phases are plotted against the dataset size N.N is consistent for all three patterns.

I –BIRCH ALGORITHM

1. Before scanning any data point from database, initialize the initial CF tree threshold, this threshold will be used as initial threshold value for every new created MLCF entry and will not be changed during the clustering process.
2. For each new scan of data point, Start from the root, recursively descend the CF-tree by choosing the closest child node, Upon reaching a leaf node, find the closest leaf entry and then test whether it can absorb the new point without violating the local MLCF threshold condition, If so, update the CF entry to reflect the insertion of the new data point and complete normally as origin birch algorithm.
3. If the chosen closest MLCF entry can't absorb the new data point (local threshold condition violated) then:

   a) Try to increase the local threshold by multiplying with threshold Modifying Factor (MF), this Modifying Factor value will depend on the value of MLCF threshold, if the threshold is small the Modifying Factor will be relatively large and if the threshold is large the Modifying Factor will be relatively small.

   b) If the chosen MLCF entry can absorb the new data point with the modified threshold update the CF entry to reflect the insertion of the new data point, and update the threshold to the new threshold value. Find the nearest neighbor entry to the current MLCF entry and test whether it can absorb the nearest neighbor entry centroid with the new threshold, if so merge the two MFLC entries and updatethe path to the root.

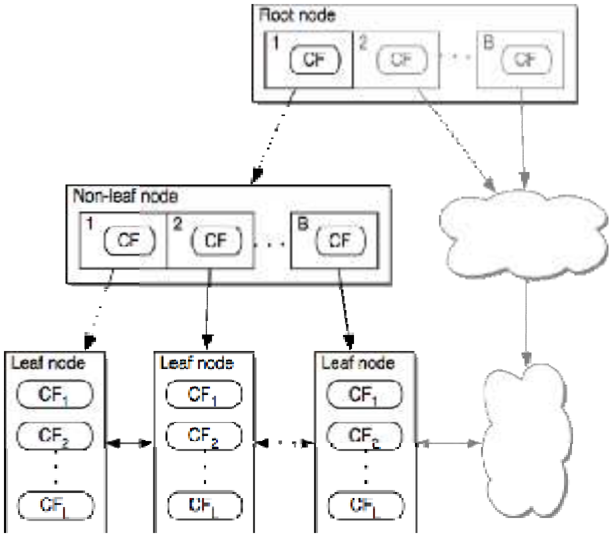
c) If the chosen MLCF entry can't absorb the new data point with the modified threshold, keep the oldthreshold and add a new MLCF entry and complete normally as origin birch algorithm .

RESULTS AND DISCUSSION

BIRCH algorithm contains the major CF Tree and the parameters such as memory, disk, outlierhandling are

CF TREE

- A height balancedtree with two parameters:
  - branching factor B
  - threshold T
- Each non-leaf node contains at most B entries of the form [CF_i,$child_i$], where $child_i$ is a pointer to its i-thchild node and CF_i is the CF of the subcluster represented by this child.
- Hence, a non-leaf node represents a cluster made up of all the subclusters represented by its entries asshown in fig 3.5.
- A leaf node contains at most L entries, each of them of the form [CF_i], where i = 1, 2, …, L .
- It also has two pointers,prev and next, which are used to chain all leaf nodes together for efficient scans.
- A leaf node also represents a cluster made up of all the sub clusters represented by its entries.
- But all entries in a leaf node must satisfy a threshold requirement, with respect to a threshold value T: thediameter (or radius) has to be less than T.
- The tree size is a function of T (the larger the T is, the smaller the tree is).
- A node is required to fit in a page of size of P.
- B and L are determined by P (P can be varied for performance tuning).
- Very compact representation of the dataset because each entry in a leaf node is not a single data point buta subcluster.
- The leaf contains actual clusters.
- The size of any cluster in a leaf is not larger than T.



The Structure of CF Tree I-BIRCH parameters

SCOPE	PARAMETER	DEFAULT VALUE
Global	Memory(M)	80*1024bytes
	Disk(R)	20% M
	Distance def	D2
	Threshold def	Threshold for D
Phase 1	Initial Threshold	0.0
	Delay-Split	On
	Page size(P)	1024 bytes
	Outlier handling	On
Phase 3	Input range	1000
	Algorithm	Adapted HC
Phase 4	Refinement pass	1
	Discard Outlier	Off

ISSN: 2096-3246
Volume 55, Issue 03, Aug, 2023

Phase 1: Scan all data and build an initial in-memory CF tree, using the given amount of memory andrecycling space on disk.

Phase 2: Condense into desirable length by building a

smaller CF tree. Phase 3: Global clustering.

Phase 4: Cluster refining – this is optional, and requires more passes over the data to refine the results.

Phase1

STEP 1: Starts with initial threshold, scans the data and inserts points into the tree.

STEP 2: If it runs out of memory before it finishes scanning the data, it increases the threshold value and rebuilds a new, smaller CF tree, by re-inserting the leaf entries from the older tree and then resuming the scanning of the data from the point at which it was interrupted.

STEP 3: Good initial threshold is important but

hard to figure out. STEP 4: Outlier removal (when

rebuilding tree).

Phase 2

STEP 1: Preparation for Phase 3.

STEP 2: Potentially, there is a gap between the size of Phase 1 results and the input range of Phase 3.

STEP 3: It scans the leaf entries in the initial CF tree to rebuild a smaller CF tree, while removing more outliners and grouping crowded sub clusters into larger ones.

Problems after Phase 1:

- Input order affects results.
- Splitting triggered by node size.

Phase 3

STEP 1: It uses a global or semi-global algorithm to cluster all leaf entries.

STEP 2: Adapted agglomerative hierarchical clustering algorithm is applied directly to the subclustersrepresented by their CF vectors.

Phase 4

STEP 1: Additional passes over the data to correct inaccuracies and refine the clusters further.

STEP 2:It uses the centroids of the clusters produced by Phase 3 as seeds, and redistributes the data points to its closest seed to obtain a set of new clusters.

STEP 3: Converges to a minimum (no matter how many time is repeated). STEP 4: Option of discarding outliners.

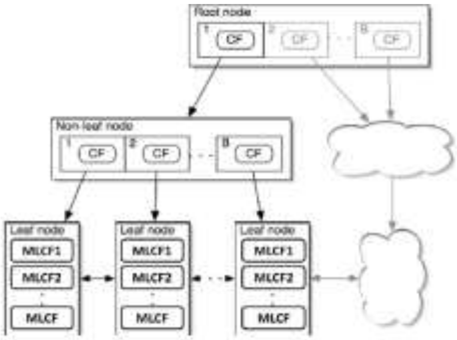
Overviews of these phases are given in fig 3.7.

Modified leaf CF Entry

In original birch algorithm a Clustering Feature (CF) entry is a triple summarizing the information that maintains about a sub cluster of data points, as described in previous sections the structure CF entry is described by the following formula, CF = {N, LS, SS}. In the modified leaf CF entry (MLCF), add a fourth value to represent the threshold value of the leaf CF entry, MLCF entry is described by the following formula. MLCF =
{N, LS, SS, T} as shown in fig 3.6, Where:

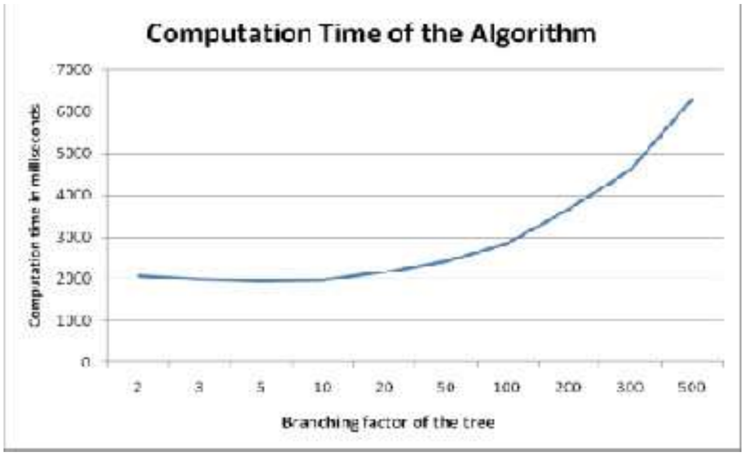
N: is the number of points in the data set. LS: is the linear sum of points in the data set.

SS: is the square sum of points in the data set. T: is the threshold value of the leaf CF entry.



The Structure of Modified Leaf CF Tree

I - BIRCH Algorithm



Computation Graph

I-BIRCH algorithm works with the statement, “Branching factor is directly proportional to the computation time”. Branching factor of the tree indicates that there is an increase in the size of the tree. As the branching factor increases the computation time also increases respectively.

CONCLUSION

Data mining, statistics, knowledge discovery, and machine learning are just a few of the several domains that make use of clustering. Using numerous thresholds as opposed to the original birch algorithm's single threshold is an improvement that will be included in this study. It will fix several problems with the basic birch algorithm and seem to provide decent performance. Reduced CF tree size improves birch algorithm efficiency across the board, and the enhanced multi-threshold birch algorithm makes sure that this efficiency gain doesn't come at the expense of clustering accuracy.

REFERENCES

In the proceedings of the 2016 IEEE Computer Cardiology Conference (CinC), D. J. Cornforth and H. F. Jelinek presented a paper titled "Detection of congestive heart failure Using Renyi entropy" (pp. 669_672).
2."Conditional mutual information-based featureselection for congestive heart failure recognition using heart rate variability," published in "Comput. Methods Programmes Biomed." in 2012 with the DOI 10.1016/j.cmpb.2011.12.015, is written by S.-N. Yu and M.-Y. Lee.

Y. I³ler and M. Kuntalp, "'Improving performance in detecting congestive heart failure by combining traditional HRV indicators with wavelet-tropy measures,"Comput. Biol. Med., vol. 37, no. 10, pp. 1502_1510, 2007, doi: 10.1016/j.compbiomed.2007.01.012.
"A wavelet-based soft decision approach for screening of patients with congestive heart failure," published in Biomed.Signal Process. Control, 2007, doi: 10.1016/j.bspc.2007.05.008. The authors are A. Hossen and B. Al-Ghunaimi. 3.
Y. Isler, A. Narin, M. Ozer, and M. Perc, "Multi-stage classi_cation of congestive heart failure based on short-term heart rate variability," Chaos, Solitons Fractals, vol. 118, pp. 145_151, Jan. 2019, doi: 10.1016/j.chaos.2018.11.020.
"Do current measurements of Poincare plot geometry reflect nonlinear characteristics of heart rate variability?" was asked by M. Palaniswami, P. Kamen, and M. Brennan. "IEEE Transactions on Biomedical Engineering," volume 48, issue 11, pages 1342–1347, November 2001, doi:

10.1109/10.959330.
"Astudy on the optimum order of autoregressive models for heart rate variability," published in Physiol.Meas. in 2002, with the DOI 10.1088/0967-3334/23/2/308. The authors of the article are A. Boardman, F. S. Schlindwein, A. Leite, and A. P. Rocha.

In their 2007 article "Use of sample entropy approach To study heart rate variability in obstructive sleep apnea syndrome," H. M. Al-Angari and A. V. Sahakian discuss the use of this method to research OSA. The article is published in the IEEE Trans. Biomed. Eng. journal and has the DOI: 10.1109/TBME.2006.889772. 6.

"A survey on pattern recognition applications of support vector machines," published in the International Journal of Pattern Recognition and Artificial Intelligence in 2003, with the DOI 10.1142/S0218001403002460, was written by S.-W. Lee and H. Byun.

"Comparison of the effects of crossvalidation methods on determining performances of classi_ers used in diagnosing congestive heart failure" (Meas. Sci. Rev., vol. 15, no. 4, pp. 196_201, 2015, doi: 10.1515/msr-2015-0027), written by Y. Isler, A. Narin, and M. Ozer (10.).

"Linear and non-linear 24 h heart rate variability in chronic heart failure," published in the journal Autonoma Neurosci. in 2000 with the DOI 10.1016/S1566-0702(00)00239-3, was the work of S. Guzzetti et al. 11.In their 2014 study, Narin, Isler, and Ozer investigated how feature selection approaches based on backward elimination and statistical significance may enhance the performance of HRV indices in congestive heart failure (CHF). The results were published in the journal Computer Biology and Medical Imaging (doi: 10.1016/j.compbiomed.2013.11.016).Classification tree for risk assessment in patients suffering from congestive heart failure by long-term heart rate variability, "IEEE Journal of Biomedical and Health Informatics," 17, no. 3, 727–733, May 2013, doi: 10.1109/JBHI.2013.2244902. Authors: P. Melillo, N. De Luca, M. Bracale, and L. Pecchia.