www.ijiarec.com

ISSN:2348-2079

Volume-7 Issue-4

International Journal of Intellectual Advancements and Research in Engineering Computations

Prediction of Cardiac Disease using I-Birch Algorithm

Dr.N.Hemageetha[1], V.Priya Dharshini[2]

[1]Associate Professor & Head, Department of Computer Science, Government Arts College For women, Salem – 636008.

[2]M.Phil Research Scholar, Department of Computer Science, Government Arts College for Women, Salem – 636008.

ABSTRACT

Heart disease is one of the biggest causes for morbidity and mortality among the population of the world. Prediction of cardiovascular disease is one of the important subjects in the section of clinical data analysis. Heart disease encompasses many diseases of the heart and blood pressure, heart attacks, angina pectoris (chest pain or discomfort caused by a reduced blood supply to the heart muscle), stroke and heart failure. Heart disease describes a condition that affects patient heart. In proposed work, Data mining techniques is used to predict the cardiac disease. Patient ID, Patient Name, Age, Gender, Delta Heart Rate and cpu-date are the attributes present in the dataset. Clustering is an important data mining and descriptive task. It has been researched deeply by various researchers for diverse application areas and is applied in multiple working domains such as data classification and image processing. In proposed work Clustering algorithm, Improved Balanced Iterative Reducing and clustering using Hierarchies (I-BIRCH) are used. To establishing threshold value to the central point where its data formed a cluster dynamically, have verified the advantage on close ratio of the data of I-BIRCH algorithm of establishing threshold value dynamically through the experiment, put forward improve algorithm apply to in the loss analysis of Health Analysis System.

Keywords: Data Mining Techniques, I-BIRCH Algorithm, CF TREE.

INTRODUCTION

Data Mining Approach to detect Heart diseases is an interesting research field. The objective of the dissertation is to predict exactly the presence of Heart Disease with less number of attributes. The data set has been taken from this research work. Patient ID, Patient Name, Age, Gender, Delta Heart Rate and CPU-date are the available attributes in the dataset. Heart Disease is a collection of diseases and conditions that cause cardiovascular problem. Cardiovascular diseases are the highest-flying disease in modern world. According to the world health organization about more than 12 million deaths occur worldwide, every year due to heart problem. It is also one of the dangerous diseases in India which causes maximum death. The diagnosis of this disease is intricate process. It should be diagnosed accurately and correctly. High risk occurs due to limitation of the medical experts and their unavailability. Normally, it is diagnosed using intuition of the medical specialist. It would be vastly advantageous if the techniques will be included with the medical information system. Heart disease means any disorder of the heart. Cardiovascular disease describes problems with the blood vessels and circulatory system, heart disease refers to problem that affects heart itself. In this thesis we have taken dataset from www.datasciencecentral.com. Patient ID, Patient Name, Age, Gender, Delta Heart Rate and CPU-date are used as dataset. The frame work of the research is shown in the figure.1.

Author for correspondence:
M.Phil Research Scholar, Department of Computer Science, Government Arts College for Women, Salem – 636008.

LITERATURE REVIEWS

Fuzzy Geographically Weighed Clustering (FGWC) is introduced by ArieWahyuWijayanto [1] based on Particle Swarm Optimization (PSO). Dr.C.A.Dhote [2] identified hybrid Swarm Intelligence based technique for data clustering using Particle Swarm Optimization and Bee Algorithm. ShafiqAlam [3] recommended Evolutionary Particle Swarm Optimization (EPSO)-clustering algorithm which is based on PSO. JaskaranjitKaur [4] planned a new hybrid method which combines the features of K-means clustering algorithm and BIRCH (a hierarchical clustering algorithm). LászlóKovács [5] specified BIRCH pre-clustering method. The pre-clustering is capable of data reduction method in the instance of large data sets. In the pre-clustering process, an important feature is to make available a good intra-cluster similarity. Du Haizhou [6] proposed a D-BIRCH algorithm founding threshold value to the essential point where its data shaped a cluster dynamically. YizhouYang [7] proposed a parallel spatial index called Hilbert R tree index, which can be carried on multi core CPU and computer cluster for parallel spatial queries and data retrieval. Yu Tu [8] specified a new kind of OSBC algorithm based on BIRCH clustering features for time series division. NidalIsmael [9] recommended BIRCH clustering algorithm for very large data sets. In the algorithm, a CF-tree is constructed whose all entries in each leaf protuberance must please a uniform threshold T, and the CF-tree is reconstructed at each stage by different threshold. TianZhang [10] proposed BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies), and exposes that it is specifically appropriate for very large databases.

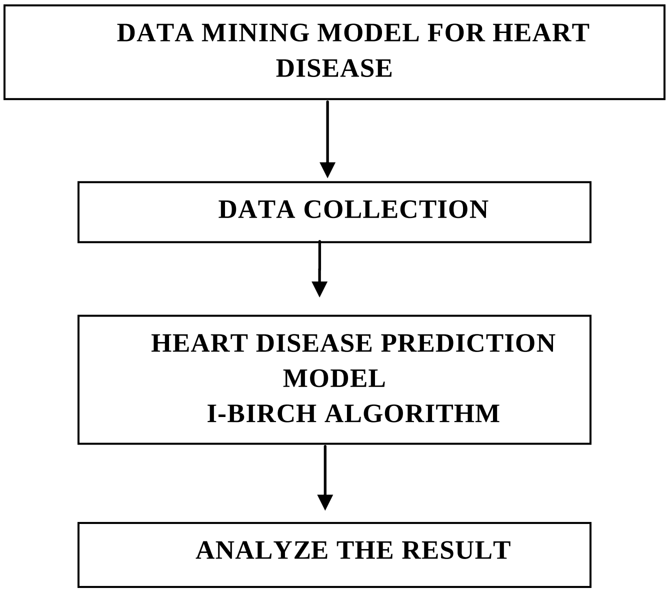
DATA MINING MODEL FOR HEART DISEASE
DATA COLLECTION
HEART DISEASE PREDICTION MODEL
I-BIRCH ALGORITHM
ANALYZE THE RESULT

Figure.1 Framework of the research

METHODOLOGY

Clustering in Data Mining is made possible by an unsupervised data mining algorithm used to perform hierarchical clustering called I-BIRCH(Improved Balanced Iterative Reducing and Clustering using Hierarchies). It uses the modified leaf clustering features (MLCF) entry and modified leaf cluster feature tree (MLCF Tree) two concepts for the general cluster description. Modified leaf

Clustering feature tree outlines the clustering of useful information. Space is much smaller than the meta-data collection and can be stored in memory, which can improve the algorithm in clustering large data sets on the speed and scalability. It is very suitable for handling discrete and continuous attribute data clustering problem.

A node in the I-BIRCH tree is called a Clustering Feature. I-BIRCH builds on the idea that, points that are close enough should always be considered as a group.

There are four phases in BRICH algorithm. In the first phase, the database based on the branching factor B and the threshold value T, the initial CF is built. Phase two is an optional phase in which the initial CF tree would be reduced in size to obtain a smaller CF tree. Global clustering is performed in the third phase from the initial CF tree or the smaller tree of phase two.

Good clusters can be obtained from phase3 of the algorithm. Phase4 of the algorithm would be needed in the clustering process, if it is required to improve the quality of the clusters. The execution of Phase1 of I-BIRCH begins a threshold value.

I-BIRCH ALGORITHM

I-BIRCH algorithm contains the major CF Tree and the parameters such as memory, disk, outlier handling.

CF tree

- A height balancedtree with two parameters:
- ▪ branching factor B
- ▪ threshold T
- Each non-leaf node contains at most B entries [CF_i,$child_i$], where $child_i$ is a pointer to its i-the child 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 sub clusters represented by its entries.
- A leaf node holds at maximum 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 for efficient scans.
- A leaf node also represents a cluster made up of all the sub clusters by its entries.
- With respect to a threshold value T, all entries in a leaf node must satisfy a threshold requirement,
- The tree size is said to be 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).
- A leaf node is not a single data point but a subcluster for each entry.
- The leaf contains actual clustersbecause ofcompact representation of the dataset.
- In a leaf, any cluster is not larger than T.

Algorithm

- Phase 1: Build an initial in-memory CF tree, scan all data using the given amount of memory and recycling space on disk.
- Phase 2: Change into desirable length by building a smaller CF tree.
- Phase 3: Global clustering.
- Phase 4: Cluster refining – this is optional, 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 outliers 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 sub clusters represented 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).
- STEP4: Option of discarding outliers.

Modified leaf CF Entry

In original I-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, CF = {N, LS, SS} in the formula. 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},Where:

- N: is the number of data set points.
- LS: in the data set it is the linear sum of points.
- SS: in the data set it is the square sum of points.
- T: is the threshold value in the leaf CF entry.

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.

Clustering features and CF tree

Clustering feature (CF) entry is triple summarizing the information. A CF tree is a height-balanced tree with 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.

RESULT AND DISCUSSION

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. Using this dataset, java tool is used to develop the model.

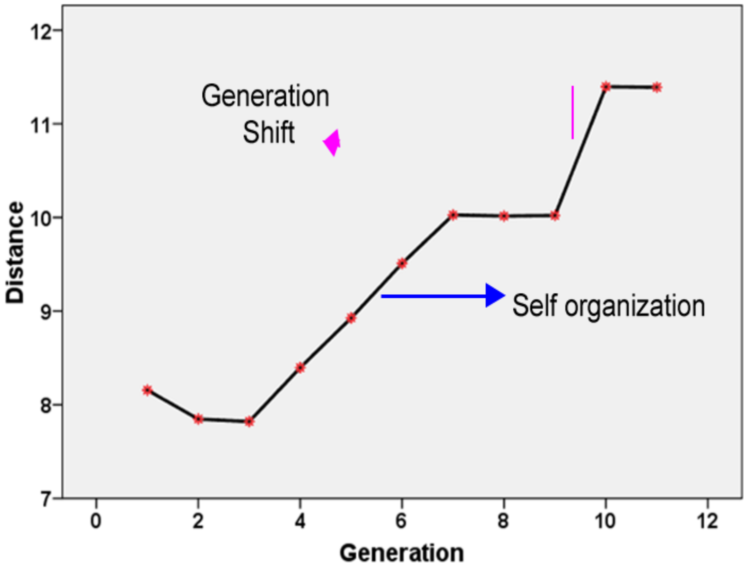

Figure.2 Generation evolution

The plot regarding the intra-cluster distance and number of generations showing the swarm evolution is given in Figure.2 The Figure.2 shows that in any generation if there is not a particle weak enough to be consumed, the I-BRICH adjusts itself by adjusting its intra-cluster distance until a weaker particle is formed and consumed. The figure also reveals the quick self organization of the particle to the lower intra cluster distance in the initial generations. The initial low intra-cluster distance is due the large number of clusters.

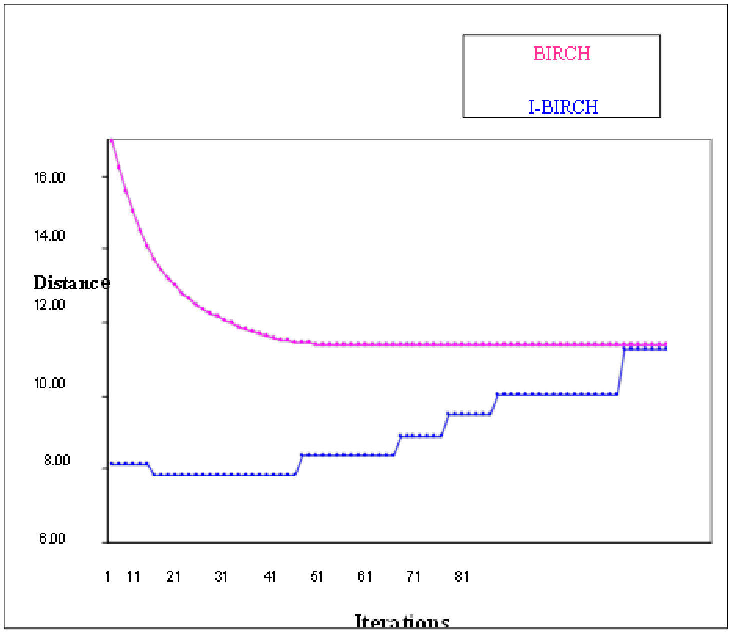

Figure. 3 BIRCH V/S I-BIRCH

The plot of the performance of I-BIRCH-clustering against BIRCH clustering is given in Figure. Due to large number of clusters I-BIRCH clustering start from lower. Intra-cluster distance as compared to BIRCH clustering where the number of clusters are fixed and start from higher intra-cluster distance in the Figure.

In this Table.1 the intra cluster distance and the number of particle is explained. Then, in which generation the particle consumed is described.

Table.1 Self Organizations and Intra-Cluster Distance

Generation_ Id	Number of particles	Intra cluster distance	Particle Consumed
	10	8.156419	
1	10	7.847273	1
2	9	7.820662	1
3	8	8.396219	0
4	8	8.926592	1
5	7	9.51014	1
6	6	10.02707	1
7	5	10.01508	0
8	5	10.02071	0
9	5	11.39571	1
Result	4	11.39009	-----

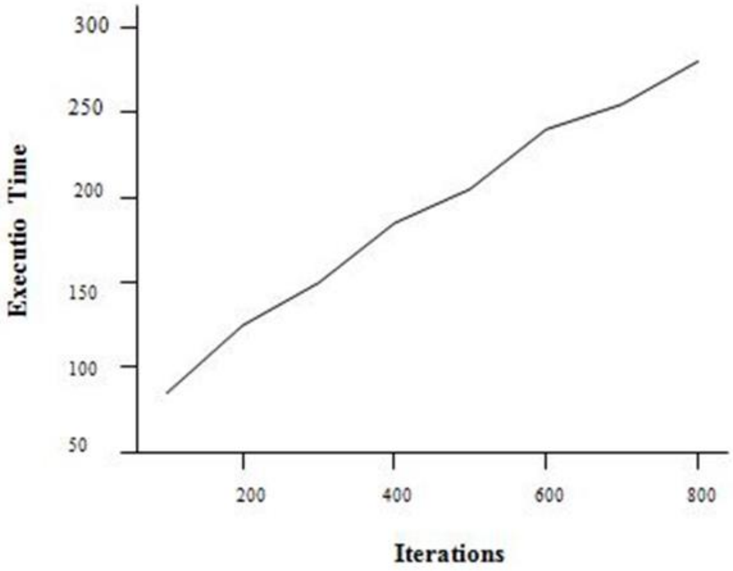

Figure.4 Iterations v/s time line

In terms of time of execution there is a linear relationship between number of iterations and the time of execution. Figure4 shows the relationship between the number of iterations of I-BIRCH and the execution time in the Figure.4.

CONCLUSION

Clustering is used in many fields such as data mining, knowledge discovery, statistics and machine learning. In this dissertation, the I-BIRCH algorithm used to predict the Heart Disease with the help of java environment. The challenges that has been undertaken for heart disease prediction is the proper selection of the machine learning technique to get accurate prediction using only minimum number of input variables. The result show the accuracy of I-BIRCH is high as compared to the BIRCH Algorithm. I-BIRCH algorithm gives us 91.5% which gives accuracy percentage than BIRCH. To predict exactly the presence of the disease based on the Heart rate. The data is collected from the I-BIRCH algorithm plays an important role in predicting the patient using the range value. Application of the data mining techniques in medicine is an evolving field of research and a lot of work has to be done. The main aim of this research work is to guide the patient to predict the disease and to help the society with all standards of people in India.

REFERENCES

[1]. ArieWahyuWijayanto,AyuPurwarianti, “ Improvement of Fuzzy Geographically Weighted Clustering using Particle Swarm Optimization”,DOI:10.13140/2.1.3920.8645

[2]. Chandrashekhar A. Dhote, Anuradha D. Thakare, Shruti M. Chaudhari, “ Data clustering using particle swarm optimization and bee algorithm”, doi: 10.1109/iccnt.2013.6726828

[3]. ShafiqAlam, “An Evolutionary particle swarm optimization algorithm for data clustering”, doi:10.1109/SIS.20084668294.

[4]. Jaskranjitkaur, “Performance evaluation of enhanced hierarchical and partitioning based clustering algorithm in data mining”, DOI:10.1109/ICATCCT.2015.7456993.

[5]. A. Hossen and B. Al-Ghunaimi, ``A wavelet-based soft decision techniquefor screening of patients with congestive heart failure," *Biomed.Signal Process. Control*, 2(2), 2007, 135_143. doi: 10.1016/j.bspc.2007.05.008.

[6]. Y. Isler, A. Narin, M. Ozer, and M. Perc, ``Multi-stage classification of congestive heart failure based on short-term heart rate variability,"*Chaos, Solitons Fractals*, 118, 2019, 145_151.
doi: 10.1016/j.chaos.2018.11.020.

[7]. M. Brennan, M. Palaniswami, and P. Kamen, ``Do existing measures ofPoincare plot geometry re_ect nonlinear features of heart rate variability?"*IEEE Trans. Biomed. Eng.*, 48(11), 2001, 1342_1347.
doi: 10.1109/10.959330.

[8]. Yu Tu, "Online segmentation algorithm for time series based on I-BIRCH clustering feature. 978-0-7695-4297-3/10$2600@2010 IEEE. DOI:10.1109/CIS.2010.19.

[9]. NidalIsmael, MohamoudAlzaalan and wesamAshour, "Improved multi threshold I-BIRCH Clustering algorithm 2(1), 2014, 1-10, http://dx.doi.org/10.14257/ijaiasd.2014.2.1.01.

[10]. Tian Zhang, "A Efficient data clustering method for very large database – ISBN: 0-8979/-794-4 doi: 10.1145/233269.233324.