International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

A Novel Approach for Improving Efficiency of Agglomerative Hierarchical Clustering For Numerical Data Set

Amar S. Chandgude[1], Vijay Kumar Verma[2]

[1] M.Tech IV Sem, Lord Krishna College of Technology Indore M.P. India

[2] Assistant Professor CSE, Loed Krishna College of Technology Indore M.P India

Abstract: ***Hierarchical clustering methods construct the clusters by recursively partitioning the instances in either a top-down or bottom-up fashion. These methods can be subdivided into Agglomerative hierarchical clustering and Divisive hierarchical clustering. The result of the hierarchical methods is a dendrogram, representing the nested grouping of objects and similarity levels at which groupings change. A clustering of the data objects is obtained by cutting the dendrogram at the desired similarity level.Single linkage method is based on similarity of two clusters that are most similar (closest) points in the different clusters. Complete linkage method based on similarity of two clusters that are least similar (most distant) points in the different clusters. Average linkage method based on average of pairwise proximity between points in the two clusters. In this paper we proposed an ensemble based technique to decide which methods is most suitable for a given dataset.***

Keyword: Data Mining, Diagnosis, Heart Attack, Symptoms, Classification, Prediction

1. Introduction

1.1 Cluster and Clustering

Some common definitions are collected from the clustering literature and givenbelow

1. ”A Cluster is a set of entities which are alike, and entities from different clusters are not alike.”
2. ”A cluster is an aggregation of points in the space such that the *distance* between two points in the cluster is less than the distance between any point in the cluster and any point not in it.”
3. ”Clusters may be described as connected regions of a multidimensional space containing a relatively high density of points, separated from other such regions by a region containing a relatively low density of points.”

Although the cluster is an application dependent concept, all clusters are comparedwith respect to certain properties: density, variance, dimension, shape, and separation. The cluster should be a tight and compact high-density region of data pointswhen compared to the other areas of space. From compactness and tightness, it followsthat the degree of dispersion (variance) of the cluster is small. The shape of thecluster is not known a priori. It is determined by the used algorithm and clusteringcriteria. Separation defines the degree of possible cluster overlap and the distanceto each other[1,3,4].

1.2 Main characteristics of a cluster

Defining the characteristics of a cluster, similar to giving a single, unique andcorrect definition, is not an exact science (Cory right, 2006). Although differentauthors emphasize on different characteristics, they do however agree on themain dimensions.

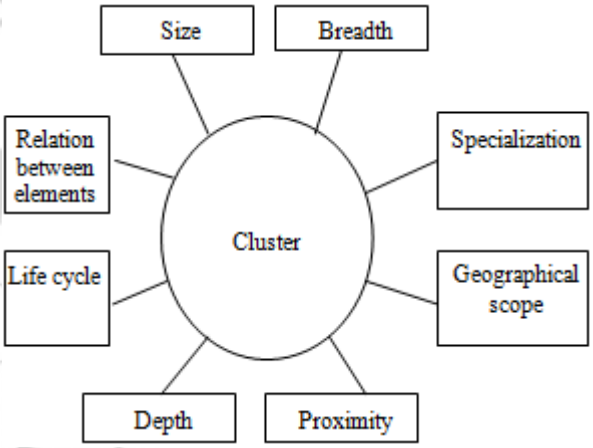

Figure 1 Main characteristics of a cluster

Boundaries of a cluster are not exact. Clusters vary in size, depth and breadth. Some clusters consist of small and some of medium and some of large in size. The depth refers to the range related by vertically relationships. Furthermore, a cluster is characterized by its breadth as well. The breath is defined by the range related by horizontally relationships[2,5,8].

2. Clustering Methods

There are many clustering methods have been developed, each of which uses a different induction principle. Farley and Raftery suggest dividing the clustering methods into two main groups: hierarchical and partitioning methods. Han and Kamber (2001) suggest categorizing the methods into

additional three main categories: density-based methods, model-based clustering and grid based methods. An alternative categorization based on the induction principle of the various clustering methods is presented in (Estivill-Castro, 2000). We discuss some of them here[6,7].

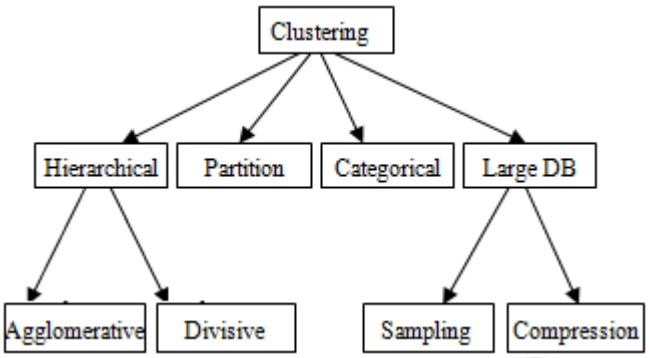

Figure 2: Clustering methods

3. Problem Statement

After having chosen the distance or similarity measure, we need to decide which clustering algorithm to apply. There are several agglomerative procedures and they can be distinguished by the way they define the distance from a newly formed cluster to a certain object, or to other clusters in the solution. The most popular agglomerative clustering procedures include the following:

1) Single linkage (nearest neighbor): The distance between two clusters corresponds to the shortest distance between any two members in the two clusters.
2) Complete linkage (furthest neighbor): The oppositional approach to single linkage assumes that the distance between two clusters is based on the longest distance between any two members in the two clusters.
3) Average linkage: The distance between two clusters is defined as the average distance between all pairs of the two clusters' members.
4) Centroid: In this approach, the geometric center (centroid) of each cluster is computed first. The distance between the two clusters equals the distance between the two centroids.

Each of these linkage algorithms can yield totally different results when used on the same dataset, as each has its specific properties. So it is very difficult to decide which method is to best for select data set. The complete-link clustering methods usually produce more compact clusters and more useful hierarchies than the single-link clustering methods, yet the single-link methods are more versatile[9,10,11].

4. Literature Review

In 2009 Xiaoke Su, Yang Lan, Renxia Wan, and Yuming Qin proposed "A Fast Incremental Clustering Algorithm". In this paper, a fast incremental clustering algorithm is proposed by changing the radius thresholdvalue dynamically. The algorithm restricts the number of the final clusters and reads the original dataset only once.At the same time an inter-cluster dissimilarity measuretaking into account the frequency information of theattribute values is introduced. It can be used for thecategorical data[12].

In 2010 ParulAgarwal, M. AfsharAlam, RanjitBiswasproposed "Analysing the agglomerative hierarchical Clustering Algorithm for Categorical Attributes". In this paper provide in depth explanation of implementation adopted for k-pragna, an agglomerative hierarchical clustering technique for categorical attributes[11].

In 2011 Hussain Abu-Dalbouh1 and NoritaMdNorwawi proposed "Bidirectional Agglomerative Hierarchical Clustering using AVL Tree Algorithm".Proposed Bidirectional agglomerative hierarchical clustering to create a hierarchy bottom-up, by iteratively merging the closest pair of data-items into one cluster. The result is a rooted AVL tree. The *n* leafs correspond to input data-items (singleton clusters) needs to *n/2* or *n/2+1* steps to merge into one cluster, correspond to groupings of items in coarser granularities climbing towards the root. One of the advantages of the proposed bidirectional agglomerative hierarchical clustering algorithm using AVL tree and that of other similar agglomerative algorithm is that, it has relatively low computational requirements. The overall complexity of the proposed algorithm is O(*logn*) and need (*n/2* or *n/2+1*) to cluster all data points in one cluster whereas the previous algorithm is O(n^2) and need (*n-1*) steps to cluster all data points into one cluster[13].

In 2012 Dan Wei, Qingshan Jiang, Yanjie Wei and Shengrui Wang proposed "A novel hierarchical clustering algorithm for gene Sequences" .The proposed method is evaluated by clustering functionally related gene sequences and by phylogenetic analysis. In this paper, they presented a novel approach for DNA sequence clustering, mBKM, based on a new sequence similarity measure, DMk, which is extracted from DNA sequences based on the position and composition of oligonucleotide pattern. Proposed method can be applied to study gene families and it can also help with the prediction of novel genes[14]

.

In 2013 Yuri Malitsky Ashish Sabharwal, Horst Samulowitz, Meinolf Sellmann proposed "Algorithm Portfolios Based on Cost-Sensitive Hierarchical Clustering". Different solution approaches for combinatorial problems often exhibit incomparable performance that depends on the concrete problem instance to be solved. Algorithm portfolios aim to combinethe strengths of multiple algorithmic approaches by training a classifier that selects or schedules solvers dependent on the given instance. Proposed algorithm devises a new classifier that selects solvers based on a cost-sensitive hierarchical clustering model. They devised a cost-sensitive hierarchical clustering approach for building algorithm portfolios. The empirical analysis showed that adding feature combinations can improve performances lightly, at the cost of increased training time, while merging cluster splits based on cross-validation lowers prediction accuracy [15].

5. Proposed Method

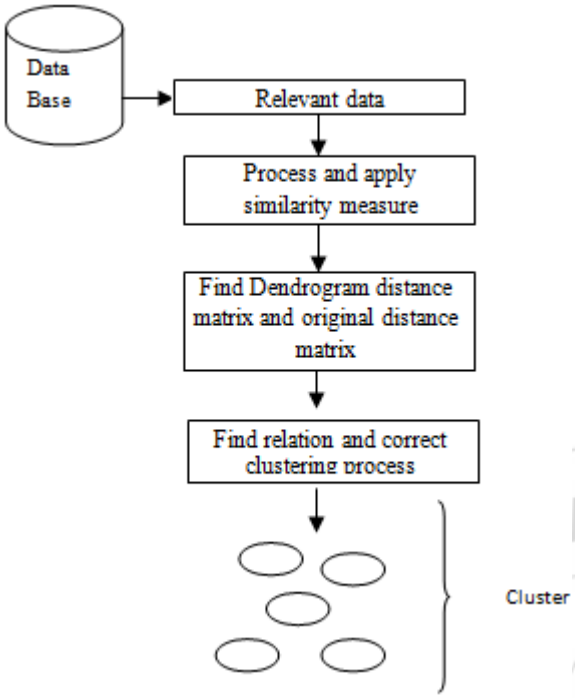

Figure 3: Architecture of proposed method

6. Proposed Algorithm

1) Assign each object as individual cluster like c_1, c_2, c_3, ..c_n where n is the no. of objects
2) Find the distance matrix D, using any similarity measure
3) Find the closest pair of clusters in the current clustering, saypair (r), (s), according to d(r, s) = mind (i, j) { i, is an object in cluster r and j in cluster s}
4) Merge clusters (r) and (s) into a single cluster to form amerged cluster. Store merged objects with its correspondingdistance in Dendrogram distance Matrix.
5) Update distance matrix, D, by deleting the rows and columnscorresponding to clusters (r) and (s). Adding a new row andcolumn corresponding to the merged cluster(r, s) and old cluster (k) is defined in this way:d[(k), (r, s)] = min d[(k),(r)], d[(k),(s)].For other rows and columns copy the corresponding data fromexisting distance matrix.
6) If all objects are in one cluster, stop. Otherwise, go to step 3.
7) Find relational value with single, complete and average linkage methods.
8) Generate correct clusters.

7. Experimental Analysis

We evaluate the performance of proposed algorithm and compare it with single linkage, complete linkage and average linkage methods. The experiments were performed on Intel Core i5-4200U processor 2GB main memory and RAM: 4GB Inbuilt HDD: 500GB OS:Windows 8.The algorithms are implemented in using C# Dot Framework Net language version 4.0.1. Synthetic datasets are used to evaluate the performance of the algorithms. For comparing the performance of the proposed algorithmswe implement the single linkage and complete linkage method. Our first comparison is based on execution time and number of objects.

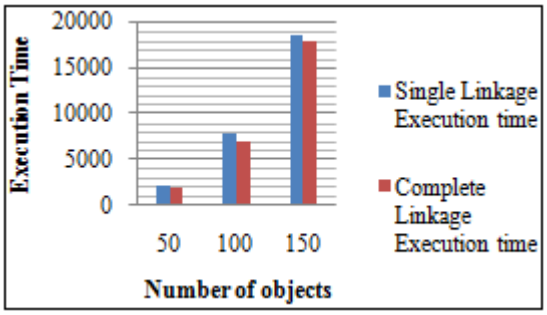

Figure 4: Comparison graph withExecution time and number of objects

8. Conclusion

There are various classification techniques that can be used for the identification and prevention of heart disease. The performance of classification techniques depends on the type of dataset that we have taken for doing experiment. Classification techniques provide benefit to all the people such as doctor, healthcare insurers, patients and organizations who are engaged in healthcare industry. These techniques are compared on basis of Sensitivity, Specificity, Accuracy, ErrorRate, True Positive Rate and False Positive Rate. The objective of each techniques is to predict more accurately the presence of heart disease with reduced number of attributes.

References

[1] DivyaTomar and SonaliAgarwal " A survey on Data Mining approaches for Healthcare" International Journal of Bio-Science and Bio-Technology Vol.5, No.5 (2013), pp. 241-266.
[2] BangaruVeeraBalaji andVedulaVenkateswaraRao"Improved Classification Based Association Rule Mining" International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 5, May 2013
[3] V.Krishnaiah, Dr.G.Narsimha and Dr.N.SubhashChandra"Diagnosis of Lung Cancer Prediction SystemUsing Data Mining Classification Techniques" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (1) 2013, 39 – 45
[4] ShamsherBahadur Patel, Pramod Kumar Yadav and Dr. D. P.Shukla" Predict the Diagnosis of Heart Disease Patients Using Classification Mining Techniques" IOSR Journal of Agriculture and Veterinary Science (IOSR-JAVS) e-ISSN: 2319-2380, p-ISSN: 2319-2372. Volume 4, Issue 2 (Jul. - Aug. 2013),
[5] N.Suneetha,Ch.V.M.K.Hari and Sunil Kumar " Modified Gini Index Classification: A Case Study Of Heart Disease Dataset" (IJCSE) International Journal on Computer

Science and Engineering Vol. 02, No. 06, 2010, 1959-1965

[6] JyotiSoni, Uzma Ansari, Dipesh Sharma and SunitaSoni "Intelligent and Effective Heart Disease Prediction System using Weighted Associative Classifiers" JyotiSoni et al. / International Journal on Computer Science and Engineering (IJCSE) ISSN : 0975-3397 Vol. 3 No. 6 June 2011

[7] SunitaSoni and O.P.VyasUsing Associative Classifiers for Predictive Analysis in Health Care Data Mining "International Journal of Computer Applications (0975 – 8887) Volume 4 – No.5, July 2010

[8] Mai Shouman, Tim Turner, Rob Stocker "Using Decision Tree for Diagnosing Heart Disease Patients" Proceedings of the 9-th Australasian Data Mining Conference (AusDM'11), Ballarat, Australia

[9] M. Akhiljabbar, Dr B.L Deekshatulu andDrPritiChandra "Heart Disease Classification Using Nearest Neighbor Classifier With Feature Subset Selection"Computer Science and Telecommunications 2013|No.3(39)

[10] SunitaSoni andO.P.Vyas "Fuzzy Weighted Associative Classifier: APredictive Technique For Health Care Data Mining" International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.1, February 2012

[11] Chaitrali S. Dangare and Sulabha S. Apte, PhD. "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques"International Journal of Computer Applications (0975 – 888) Volume 47– No.10, June 2012

[12] M. AkhilJabbar, B.L Deekshatulu&Priti Chandra "Classification of Heart Disease using Artificial Neural Networkand Feature Subset Selection" Global Journal of Computer Science and TechnologyNeural & Artificial IntelligenceVolume 13 Issue 3 Version 1.0 Year 2013Type: Double Blind Peer Reviewed International Research Journal Publisher: Global Journals Inc. (USA)Online ISSN: 0975-4172 & Print ISSN: 0975-4350

[13] N SNithyaand K DuraiswamyGain ratio based fuzzy weighted association rule miningclassifier for medical diagnostic interface Vol. 39, Part 1, February 2014, pp. 39–52. Indian Academy of Sciences

[14] M.Akhiljabbar, Dr.PritiChandrab, Dr.B.LDeekshatuluc "Heart Disease Prediction System using Associative Classificationand Genetic Algorithm"International Conference on Emerging Trends in Electrical, Electronics and CommunicationTechnologies-ICECIT, 2012

[15] A. Anushyaand A. Pethalakshmi "A Comparative Study of Fuzzy ClassifiersWith Genetic On Heart Data" International Conference on Advancement in Engineering Studies & Technology, ISBN : 978-93-81693-72-8, 15th JULY, 2012, Puducherry