International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 1 | ISSN : 2456-3307

To Assess the Performance of EAHC Algorithm Using Sensor Discrimination Dataset for the Improvement of Data Mining System

K. Thulasiram[1], Dr. S. Ramakrishna[2], Dr. M. Jayakameswaraiah[3]
[1]Research Scholar Department of Computer Science, Sri Venkateswara University, Tirupati, Andhra Pradesh, India
[2]Professor, Department of Computer Science, Sri Venkateswara University, Tirupati, Andhra Pradesh, India
[3]Assistant Professor, School of Computer Science and Applications, Reva University, Bangalore, Karnataka, India

ABSTRACT

The process of grouping a set of physical or intangible objects into classes of similar objects is called clustering. A cluster is a group of data objects that are related to one another within the similar cluster and are dissimilar to the objects in other clusters. It is suitable method for the innovation of data distribution and patterns the fundamental data. There are various clustering methods in data mining system, such as hierarchical clustering method. Most of the approaches to the clustering of variables encountered in the literature are of hierarchical category. This research work represents comprehensive discussion on the performance of our proposed Enhanced Agglomerative Hierarchical Clustering algorithm. This experiential evaluation shows that Enhanced Agglomerative Hierarchical Clustering (EAHC) algorithm contributes decent performance and decreases the runtime of construction by several orders of size, while generating stable and quality hierarchies.

Keywords : Clustering algorithms, Data Mining, Hierarchical Clustering and Enhanced Agglomerative Hierarchical Clustering (EAHC) algorithm

I. INTRODUCTION

A. Data Mining

Data mining comprises the usage of sophisticated data analysis tools to ascertain previously unknown valid patterns and relationships in large dataset. Data mining tools forecast imminent trends and behaviors, helps organizations to take preemptive knowledge driven decision. The queries that were usually tedious to settle can be settled by data mining tools. Data mining is also known as knowledge discovery in Database (KDD) and is the nontrivial extraction of hidden previously unknown and potentially suitable information from data in databases. However, databases are commonly treated data mining and knowledge discovery as synonyms, data mining is actually part of knowledge discovery process[1,4,11]. The data mining techniques, pre-processing of data, classification, clustering and outlier detection plays a major role in the development of data mining system. In our research we are going to focus on some clustering algorithms. The following Figure 1 shows the steps of knowledge discovery process in data mining.

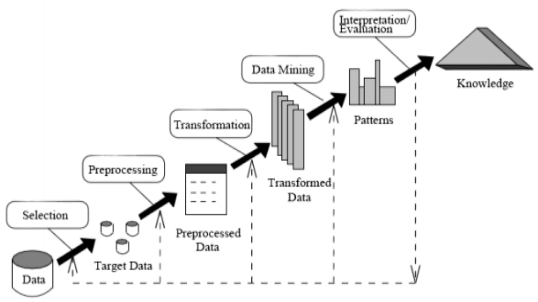

Figure 1. Knowledge Discovery Process

B. Clustering

Clustering is an essential data mining task. It can be defined as the process of organizing objects into groups whose members are similar in some way. Clustering can also be define as the process of grouping the data into classes or clusters, so that Hierarchical objects within a cluster have high similarity in association to one another but are very dissimilar to objects in other clusters[3,12]. Generally clustering can be done by two methods: Hierarchical and Partitioning method. In data mining hierarchical clustering works by grouping data objects into a tree of cluster. Hierarchical clustering methods can be further categorized into agglomerative and divisive hierarchical clustering. Hierarchical methods suffer from the fact that once we have accomplished either merge or split step, it can never be undone. This inflexibility is useful in that it leads to smaller computation costs by not having to worry about a combinatorial number of dissimilar selections[2,8]. However, such methods cannot correct incorrect decisions that once have taken. There are two methodologies that can help in improving the feature of hierarchical clustering:

1. Initially to accomplish careful analysis of object associations at every hierarchical partitioning or
2. By integrating hierarchical agglomeration and further methods by first using a hierarchical agglomerative algorithm to group objects into micro clusters[6,9]. Then implement macro clustering on the micro clusters using alternative clustering method such as iterative relocation.

II. LITERATURE REVIEW ON HIERARCHICAL CLUSTERING

Hierarchical clustering is a technique of cluster analysis which seeks to construct a hierarchy of clusters. The worth of a pure hierarchical clustering technique suffers from its incompetence to achieve adjustment, once a merge or split decision has been executed. Then it resolves neither undo what was prepared previously, nor perform object swapping between clusters. Thus merge or split decision, if not well selected at some step, may lead to somewhat low-quality clusters. One favorable direction for refining the clustering quality of hierarchical methods is to integrate hierarchical clustering with new methods for multiple phase clustering. So in this research, we designate a rare improved hierarchical clustering algorithm that overcomes the boundaries that occur in pure hierarchical clustering algorithms. Hierarchical clustering algorithms are either top-down or bottom-up. Bottom-up algorithms requisite every file as a singleton cluster at the outset and then sequentially combine (or agglomerate) pairs of clusters until all clusters have been merged into a single cluster that contains all files. Bottom-up hierarchical clustering is therefore called hierarchical agglomerative clustering or HAC. Top-down clustering needs a process for splitting a cluster[5,7,10]. It proceeds by splitting clusters recursively until distinct archives are reached.

A. Hierarchical clustering

The Hierarchical Clustering algorithm is shown below. Here we first compute the $N \times N$ similarity matrix C. The algorithm then accomplishes N - 1 steps of merging the maximum similar clusters. In every reiteration, the two most similar clusters are merged and the rows and columns of the merged cluster i in C are restructured. The clustering is stored as a list of merges in A. I indicate which clusters are still available to be merged. The function SIM(i, m, j) computes the similarity of cluster j with the merge of clusters I and m. For some Hierarchical Clustering algorithm, SIM(i, m, j) is simply a function of $C[j][i]$ and $C[j][m]$. We will now refine this algorithm for the different similarity measures of single-link and complete-link clustering and group average and centroid clustering. The merge criteria of these four variants of Hierarchical Clustering are presented.

Algorithm:

HC($d_1, \ldots, d_N$)

Step-1: for n ← 1 to N

Step-2: do for i ← 1 to N

Step-3: do C[n][i] ← SIM(d_n, d_i)

Step-4: I[n] ← 1 (keeps track of active clusters)

Step-5: A ← [ ] (assembles clustering as a sequence of merges)

Step-6: for k ← 1 to N – 1

Step-7: do (i, m) ← arg $\max_{\{i,m:I \neq m \wedge I[i]=1 \wedge I[m]=1\}}$ C[i][m]

Step-8: A.APPEND (i, m) (store merge)

Step-9: for j ← 1 to N

Step-10: do C[i][j] ← SIM(i, m, j)

Step-11: C[j][i] ← SIM(i, m, j)

Step-12: I[m] ← 0 (deactivate cluster)

Step-13: return A

III. ENHANCED AGGLOMERATIVE HIERARCHICAL CLUSTERING ALGORITHM

A. Cluster Divergence

In order to choose which clusters should be combined (for agglomerative), or where a cluster should be split (for divisive), a measure of dissimilarity between sets of clarifications are required. In furthermost methods of hierarchical clustering, this is achieved by use of an appropriate metric (a measure of distance between pairs of clarifications), and a linkage criterion which identifies the dissimilarity of sets as a function of the pairwise distances of observations in the sets. Here we applied a metric function to integrate the performance of the proposed algorithm using squared Euclidean distance formula[13,14]. The Enhanced Agglomerative Hierarchical Clustering comprises of the following three parameters need to be considered.

1. Single Linkage
2. Complete Linkage
3. Group Average

Algorithm

Step-1: Scan the Entire Database

Step-2: Collects the reduced data set by using agglomerative technique.

Step-3: Partition the reduced dataset.

Step-4: Eliminate Outliers.

Step-5: Cluster Labeled data as Partial Cluster using Squared Euclidean distance equation.

$$||a-b||_2^2 = \sum_i (a_i - b_i)^2$$

Setp-6: Initially each item $x_1, \ldots, x_n$ is in its own cluster $C_1, \ldots, C_n$.

Step-7: Repeat until there is only one cluster left:

Step-8: Merge the nearest clusters, say C_i and C_j. The result is a cluster. One can cut the tree at any level to produce different clustering. A little thought reveals that the nearest clusters are not well-defined, since we only have a distance measure $d(x, x')$ between items. This is where the variations come in:

$$d(C_i, C_j) = \min x \in C_i, x' \in C_j \; d(x, x')$$

This is known as *single-linkage*. It is identical to the minimum spanning tree algorithm. Single can set a threshold and stop clustering once the distance between clusters is above the threshold. Single-linkage tends to produce long and skinny clusters.

$$d(C_i, C_j) = \max x \in C_i, x' \in C_j \; d(x, x').$$

This is known as *complete-linkage*. Clusters tend to be compact and roughly equal in diameter.

$$d(C_i, C_j) = \frac{\sum x \epsilon C_i, x' \in C_j d(x, x')}{|C_i| \times |C_j|}$$

Step-9: The result is the *average* Euclidean distance between items. Somewhere in between single-linkage and complete-linkage and a million other ways you can think of.

Step-10: Cluster formation followed after the incorporation of Squared Euclidean distance

IV. EXECUTION WITH RESULTS

In this experimental research we used WEKA as a data mining tool to estimate the cluster performances[15]. The accuracy of different algorithms has been analyzed then we designed a new algorithm called Enhanced Agglomerative Hierarchical Clustering, it is the superlative proper algorithm having better clustering performance. In this research work we applied proposed algorithm on Sensor Discrimination dataset to estimate the performance. The experiment has been conducted with Sensor Discrimination dataset available on UC Irvine Machine Learning Repository. Authors have implemented various clustering algorithms on this dataset using WEKA tool which is developed by Machine Learning Group at the University of Waikato. The dataset used for the analysis and implementation purpose is having 2212 instances and 12 attributes and one class attribute. The experimental result with the dataset gets the following results.

Table 1. PERFORMANCE OF CLUSTERER USING TRAINING SET AS CLUSTER MODE

Clustering Techniques	Total Number of instances in dataset	Number of Clustered Instances	Number of Clusters formed	Percentage of Clustering
Hierarchical Clustering	2212	1746 466	Cluster 0 Cluster 1	79% 21%
Enhanced Agglomerative Hierarchical Clustering	2212	2151 61	Cluster 0 Cluster 1	97% 3%

Table 2. PERFORMANCE OF CLUSTERER USING PERCENTAGE SPLIT AS CLUSTER MODE

Clustering Techniques	Percentage Split 33%	Percentage Split 66%	Percentage Split 99%
Hierarchical Clustering	56% 44%	36% 64%	87% 13%
Enhanced Agglomerative Hierarchical Clustering	97% 3%	97% 3%	100%

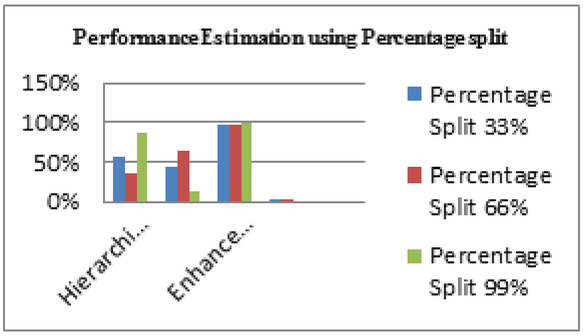

Figure 2. Performance Estimation using Percentage Split

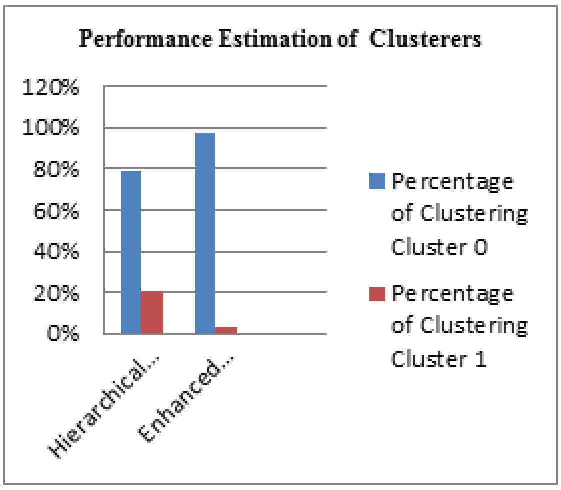

Figure 3. Performance Estimation of Clusterer

V. CONCLUTION

Data mining is an application oriented technology and is having wide applications in many fields. It also estimates, integrates and motives to guide the solution of practical problems and discover the connections between events. In my research work we proposed and developed an innovative algorithm called Enhanced Agglomerative Hierarchical Clustering (EAHC) is a bottom-up clustering method where clusters have sub-clusters, which in turn have sub-clusters, etc. There are many clustering methods, such as hierarchical clustering method, can further classify into agglomerative hierarchical methods and divisive hierarchical methods. The process of Enhanced Agglomerative Hierarchical Clustering starts with these single observation clusters and gradually combines pairs of clusters, forming smaller numbers of clusters that contain more observations. Then clusters successively merged until the desired cluster structure is obtained. In this research we estimated the performance of Hierarchical Clustering and our proposed Enhanced Agglomerative Hierarchical Clustering with results using Sensor Discrimination dataset from the UCI Machine Learning Repository. The proposed method gives admirable performance when compared the results with other algorithms. In future work if we develop, integrate and embed the EAHC algorithm with another utmost clustering algorithm called Farthest First on sensor discrimination dataset. It gives tremendous performance when compared to other clustering algorithms in the data mining system.

VI. REFERENCES

[1]. Mukhopadhyay, U. Maulik, S. Bandyopadhyay, and C. A. C. Coello, "A survey of multi objective evolutionary algorithms for data mining: part I," IEEE Transactions on Evolutionary Computation, vol. 18, no. 1, pp. 4-19, 2014.

[2]. Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques", the Morgan Kaufmann/Elsevier India, 2006.

[3]. Marjan Kuchaki Rafsanjani, Zahra Asghari Varzaneh, Nasibeh Emami Chukanlo, "A survey of hierarchical clustering algorithms", The Journal of Mathematics and Computer Science,5,3, pp.229- 240, 2012.

[4]. L. Feng, M-H Qiu, Y-X. Wang, Q-L. Xiang, Y-F. Yang and K. Liu, "A fast divisive clustering algorithm using an improved discrete particle swarm optimizer", Pattern Recognition Letters, 31, pp. 1216-1225, 2010.

[5]. Pal N.R, Pal K, Keller J.M. and Bezdek J.C, "A Possibilistic Clustering Algorithm", IEEE Transactions on Fuzzy Systems, Vol. 13, No. 4, Pp. 517- 530, 2005.

[6]. Chim H and X. Deng, "Efficient Phrase-Based Document Similarity for Clustering," IEEE Trans. Knowledge and Data Eng., vol. 20, no. 9, pp. 1217-1229 Sept. 2008.

[7]. Chee Keong Chan, Duc Thang Nguyen, Lihui Chen and Senior Member, IEEE, "Clustering with Multi viewpoint-Based Similarity Measure", IEEE transactions on knowledge and data engineering, Vol. 24, No. 6, 2012.

[8]. K. Huang, C. Chang, and K. Lin, "Prowl: an efficient frequent continuity mining algorithm on event sequences", in Data Warehousing and Knowledge Discovery,vol.3181of Lecture Notes in Computer Science, pp. 351-360, Springer, Berlin, Germany,2004.

[9]. Modha D and I. Dhillon, "Concept Decompositions for Large Sparse Text Data Using Clustering," Machine Learning, Vol. 42, nos. 1/2, pp. 143-175, Jan. 2001.

[10]. S. Huang, "A study of the application of data mining on the spatial landscape allocation of crime hot spots", in Geo-Informatics in Resource Management and Sustainable Ecosystem, vol. 398 of Communications in Computer and Information Science, pp. 1274-286, Springer, Berlin, Germany, 2013.

[11]. X. Wu, X. Zhu, G.Q. Wu, and W. Ding, "Data mining with big data," IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 1, pp. 97-107, 2014

[12]. N.Elgendy and A.Elragal, "Big data analytics: a literature review paper," in Advances in Data Mining. Applications and Theoretical Aspects, vol. 8557 of Lecture Notes in Computer

Science, pp. 214- 227, Springer, Cham, Switzerland, 2014.

[13]. Dr.M.Jayakameswariah, Mr.M.Veeresh Babu, Dr.S.Ramakrishna,Mrs.P.Yamuna, "Computation Accuracy of Hierarchical and Expectation Maximization Clustering Algorithms for the Improvement of Data Mining System", International Research Journal of Engineering and Technology (IRJET), Volume: 03 Issue: 12, ISO 9001:2008, Page 1580-1585, e-ISSN: 2395-0056, p-ISSN: 2395-0072, Dec-2016.

[14]. M.Jayakameswaraiah and S.Ramakrishna, "A Study on Prediction Performance of Some Data Mining Algorithms", International Journal of Advanced Research in Computer Science and Management Studies, Volume 2, Issue 10, ISSN: 2321-7782, October 2014.

[15]. Mahendra Tiwari, Yashpal Singh, Performance Evaluation of Data Mining Clustering Algorithm in WEKA, informaticsjournals, Volume 4, Issue 1, January-June 2012.