ISSN: 2321-7782 (Online)

Volume 3, Issue 9, September 2015

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

Dynamic Data Model Analysis Using Hierarchical Algorithm

N. Hemalatha[1]
Kongunadu college of Engineering and Tecnology,
Trichy, India

C. Radhakrishnan[2]
Assistant Professor,
Department of Computer Science,
Kongunadu college of Engineering and Tecnology, Trichy.

N. Premkumar[3]
Assistant Professor,
Department of Computer Science,
Kongunadu college of Engineering and Tecnology, Trichy.

Abstract: Clustering in data mining is a discovery process that groups a set of data such that the intra cluster similarity is maximized and the inter cluster similarity is minimized. Hierarchical clustering is of great importance in data analytics especially because of the exponential growth of real-world data. Often these data are unlabelled and there is little prior domain knowledge available. One challenge in handling these huge data collections is the computational cost. Most of the work has focused on static data sets. There has been little work on clustering of dynamic data. In proposed system is interested in exploring algorithms are capable of finding relationships amongst the elements in a dynamic data set. In this project is aim to improve the efficiency by introducing a set of methods of agglomerative hierarchical clustering. Instead of building cluster hierarchies based on raw data points, our approach builds a hierarchy based on a group of centroids. To propose hierarchical clustering with decentralized clustering method, which is suitable for distributed and dynamic dataset. The experimental results indicate that, using the centroids based approach, computational cost can be significantly reduced without compromising the clustering performance. The performance of this approach is relatively consistent regardless the variation of the settings, i.e., clustering methods, data distributions, and distance measures.

Keywords: Chameleon Algorithm, Data Mining, Distance Measures, Hierarchical Clustering.

I. Introduction

Road traffic accidents are a social and public health challenge, as they almost always result in injuries and/or fatalities (Anderson 2009). The World Health Organization estimates over 1 million people are killed each year in road collisions. This is equal to 2.1% of the annual global mortality and an estimated social cost of $518 billion.

To significantly reduce traffic fatalities and serious injuries on public roads, need to review the characteristics of traffic accidents and identify the hidden patterns behind the accidents‟ records, referring mainly to the actual knowledge contained in the collision data rather than the raw data records themselves. For example, road safety managers or residents may be interested in the accident patterns near their common unities and not the data records.

Previous traffic safety studies show that, in most cases, the occurrences of traffic accidents are seldom random in space and time, but form clusters that indicate accident concentration areas in geographic space. A concentration area is defined as an area or location where there is a higher likelihood for an accident to occur based on historical data and spatial dependency. Thus, if we can identify the locations with the high risk on the roads, road safety managers can analyze the reasons behind the fact; and, the public can be aware of the danger, so that they can drive more carefully on the dangerous road or avoid it altogether.

There are many cities in our country which has a large Population compared to their extension. Naturally these People need a good transport system to cope with their needs like going to work, go shopping, etc. It should be done through city roads. The growth of population and the need of transportation system in one amount of traffic information in roads and cities hand and the addition of transport vehicles on the other hand, we need a good city Management all over the country. The addition of transport facilities in a town involves high financial and chronological expenses. These problems show the need of correct traffic management .As the traffic had bad effect on the air it has also undesired effects on human lives in different aspects, so it is necessary to investigate the ways to control and overcome the difficulties in this field. One of the most important problems of traffic is taking a lot of time, so we can prevent ton desired traffic effects. Today is the age of information increasing in any field; there are large databases.

In this paper is to propose an efficient agglomerative hierarchical clustering algorithm method. It does not require feature selection and extraction. This method aims to reduce the attributes in traffic accident analysis. The computational time will vary depends on the attribute selection in hierarchical clustering here to use chameleon hierarchical clustering algorithm to measure the accuracy and response time. The detailed process of the proposed work as follows.

» Gather Traffic Accident Dataset, and then eliminate missing values founded records in the dataset.

» To assign rules in the training dataset it produces the optimal rules to find the accuracy

» Apply Training dataset in to algorithm to find optimum number of clusters.

» Creating 12 rules and identify the accuracy and response time.

The paper organized as follows in section 2 discuss the related work. In section 3 briefly discuss the proposed work with problem statement. In section 4 methodologies and dataset are described in this study. In section 5 produce the detailed implementation of the proposed work with simulation result. In section 6 concludes the proposed study with future work.

II. Related Work

In the agglomerative approach ,the clustering process starts with every data element in an individual cluster, The individual cluster are, then merged on the basis of proximity until all the elements are in single cluster.[1] Proposes the work in an aggoleramative manner starting from a relatively large number of particles and combining down, to only one final particles

The individual based category compact regions are classified as different entities including groups or persons. Exploiting a set of heuristics. In people that stand close for a while are joint into groups through a connection graph build by exploring heuristics on the moving regions.[2] Proposes; the hierarchical approach is suitable for decentralized data analysis and prediction.

[3] Predict the distance between any two nodes. Distance prediction proceeds in the bottom fashion, If two nodes belongs to the same cluster this implies their relatively close to each other. So we predict the distances between then otherwise two nodes are belong to two different clusters. The hierarchical approach helps to improve the accuracy of the distance prediction.

The detailed driving data [acceleration, breaking and driver response.[4] and crash data, that would better enable identification of cause and effect. Relationship with regard to crash probabilities are difficulty not available. As a result researches have framed their analytic approaches to study the factor that affect the number of crashes occurring in the specified time period.

In the decentralized dataset each node owns co-ordinate points and local error all nodes adjust their network co-ordinates and local error via measuring their latencies to some other nodes in the system.[5] Proposes each attributes in the accident datasets measuring their latencies based on prototype and distances to other attributes in the dataset.

Centralized clustering is problematical if data is widely distributed dataset ae volatile (or) data items can't be compactly represented [6].Decentralized on the other hand is a well- know problem .even in the centralized where each data item can compared to every data item ,perfect cluster can be hand to find. Decentralized creates the additional complication that even if the correct classification can be determined with the in complete inform available, the location of the item belonging to the class also need to be discovered.

III. Proposed Work

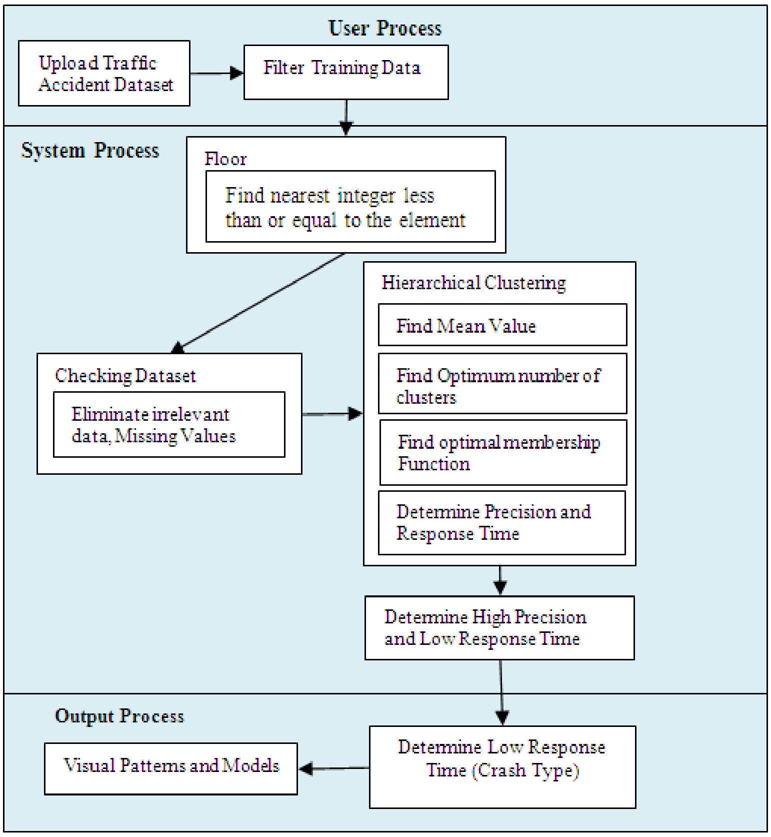
Reducing the number of traffic accidents remains one of the greatest challenges facing many societies around the world. The cost of traffic accident on society and individuals is very high. Loss of life, disability and suffering are but a few Reducing the number of traffic accidents remains one of the greatest challenges facing many societies around the world.

The cost of traffic accident on society and individuals is very high. Loss of life, disability and suffering are but a few of the impacts of traffic accidents. On average a higher proportion of Indian drivers are involved in road accidents compared to their relative population among licensed drivers.

A study of the reasons behind traffic accidents revealed four main factors: factors related to driving (the human factor); vehicle-related factors (physical environmental factors); mechanical factors, and socio-economic factors whether factor and Drinking driving. A traffic collision occurs when a road vehicle collides with another vehicle, pedestrian, animal, or geographical or architectural obstacle.

It can result in injury, property damage, and death. Road accidents have been the major cause of injuries and fatalities in worldwide for the last few decades. It is estimated that the amount of data stored in the world's database grows every twenty months at a rate of 100%. This fact shows that we are getting more and more exploded by data/ information and yet ravenous for knowledge. Data mining is a useful tool to address the need for sifting useful information such as hidden patterns from databases.

System Architecture


User Process
Upload Traffic Accident Dataset
Filter Training Data
System Process
Floor
Find nearest integer less than or equal to the element
Hierarchical Clustering
Find Mean Value
Find Optimum number of clusters
Find optimal membership Function
Determine Precision and Response Time
Checking Dataset
Eliminate irrelevant data, Missing Values
Determine High Precision and Low Response Time
Output Process
Visual Patterns and Models
Determine Low Response Time (Crash Type)

Example Traffic accident data

Accident	Gender	Age	Alcohol	Speed
1	M	Young	Yes	>100

2	M	Young	No	80-90
3	M	Middle	No	70-80
4	F	Old	No	<60
5	M	Young	Yes	70-90

Table: 1 Traffic accident data

IV. METHODOLOGY USED

4.1 Agglomerative Hierarchical Clustering

Agglomerative hierarchical clustering is a bottom-up clustering method where clusters have sub-clusters, which in turn have sub-clusters, etc. The classic example of this is species taxonomy. Gene expression data might also exhibit this hierarchical quality (e.g. neurotransmitter gene families). Agglomerative hierarchical clustering starts with every single object (gene or sample) in a single cluster. Then, in each successive iteration, it agglomerates (merges) the closest pair of clusters by satisfying some similarity criteria, until all of the data is in one cluster.

The hierarchy within the final cluster has the following properties:

- Clusters generated in early stages are nested in those generated in later stages.
- Clusters with different sizes in the tree can be valuable for discovery.

A Matrix Tree Plot visually demonstrates the hierarchy within the final cluster, where each merger is represented by a binary tree.

Process

- Assign each object to a separate cluster.
- Evaluate all pair-wise distances between clusters (distance metrics are described in Distance Metrics).
- Construct a distance matrix using the distance values.
- Look for the pair of clusters with the shortest distance.
- Remove the pair from the matrix and merge them.
- Evaluate all distances from this new cluster to all other clusters, and update the matrix.
- Repeat until the distance matrix is reduced to a single element.

Advantages

- It can produce an ordering of the objects, which may be informative for data display.
- Smaller clusters are generated, which may be helpful for discovery.

4.2 Euclidean Distance Measure:

In mathematics, the Euclidean distance or Euclidean metric is the "ordinary" distance between two points that one would measure with a ruler, and is given by the Pythagorean formula. By using this formula as distance, Euclidean space (or even any inner product space) becomes a metric space. The associated norm is called the Euclidean norm. Older literature refers to the metric as Pythagorean metric.

The Euclidean distance, data vector p and centroid q is computed as

$$d(p,q) = \sqrt{\sum_{k=1}^{n}(q_{ik} - p_{ik})^{2}}$$

4.3 Chameleon Algorithm

Clustering is a discovery process in data mining. It groups a set of data in a way that maximizes the similarity within clusters and minimizes the similarity between two different clusters. These discovered clusters can help explain the characteristics of the underlying data distribution and serve as the foundation for other data mining and analysis techniques. Clustering is useful in characterizing customer groups based on purchasing patterns, categorizing Web documents, grouping genes and proteins that have similar functionality, grouping spatial locations prone to earthquakes based on seismological data, and so on.

Most existing clustering algorithms find clusters that fit some static model. Although effective in some cases, these algorithms can break down—that is, cluster the data incorrectly—if the user doesn't select appropriate static-model parameters. Or sometimes the model cannot adequately capture the clusters characteristics. Most of these algorithms break down when the data contains clusters of diverse shapes, densities, and sizes.

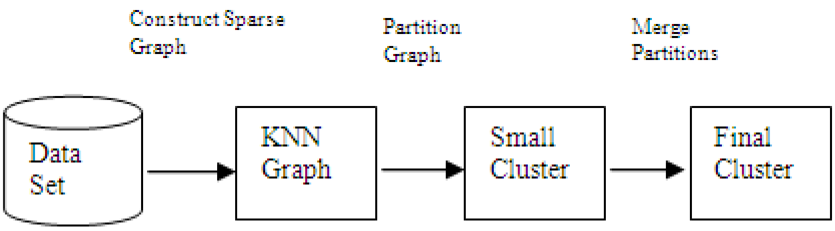
Existing algorithms use a static model of the clusters and do not use information about the nature of individual clusters as they are merged. Furthermore, one set of schemes ignores the information about the aggregate interconnectivity of items in two clusters. The other set of schemes ignores information about the closeness of two clusters as defined by the similarity of the closest items across two clusters. By only considering either interconnectivity or closeness, these algorithms can easily select and merge the wrong pair of clusters.

Chameleon is a new agglomerative hierarchical clustering algorithm that overcomes the limitations of existing clustering algorithms. The Chameleon algorithm's key feature is that it accounts for both interconnectivity and closeness in identifying the most similar pair of clusters. It thus avoids the limitations of interconnectivity or closeness. Furthermore, Chameleon uses a novel approach to model the degree of interconnectivity and closeness between each pair of clusters. This approach considers the internal characteristics of the clusters themselves. Thus, it does not depend on a static, user-supplied model and can automatically adapt to the internal characteristics of the merged clusters.

Chameleon operates on a sparse graph in which nodes represent data items, and weighted edges represent similarities among the data items. This sparse graph representation allows Chameleon to scale to large data sets and to successfully use data sets that are available only in similarity space and not in metric spaces.

Data sets in a metric space have a fixed number of attributes for each data item, whereas data sets in a similarity space only provide similarities between data items.

Chameleon finds the clusters in the data set by using a two-phase algorithm. During the first phase, Chameleon uses a graph-partitioning algorithm to cluster the data items into several relatively small sub clusters. During the second phase, it uses an algorithm to find the genuine clusters by repeatedly combining these sub clusters.



4.4 Silhouette Coefficient

The Popular method of silhouette coefficients combines both cohesion and separation .The following steps explain how compute the silhouette coefficient for an individual point , a process that consists of the following three steps .We use distances, but an analogous approach can be used for similarities.

1. For the i^{th} object, calculate its average distance to all other objects in its cluster .Call this value a_i.
2. For the i^{th} object and any cluster not containing the object ,calculate the object average distance to all the object in the object in the given cluster .Find the minimum such value with respect to all clusters ; call this values b_i
3. For the i^{th} object, the silhouette coefficient is $s_i = (b_i - a_i) / \max(a_i ,b_i)$.

The value of the silhouette coefficient can vary between -1 and 1. A negative value is undesirable because this corresponds to a case in which a_i, the average distance to points in the cluster, is greater than b_i, the minimum average distance to points in another cluster. We want the silhouette coefficient to be positive ($a_i < b_i$) and for a_i to be close to 0 as possible, since the coefficient assumes its maximum values of 1 when $a_i = 0$

4.5 Traffic Accident Dataset

This data set of traffic accidents is obtained from the National Institute of Statistics (NIS) for the region of Flanders (Belgium) for the period 1991-2000. More specifically, the data are obtained from the Belgian “Analysis Form for Traffic Accidents” that should be filled out by a police officer for each traffic accident that occurs with injured or deadly wounded casualties on a public road in Belgium. In total, 340.184 traffic accident records are included in the data set. The traffic accident data contain a rich source of information on the different circumstances in which the accidents have occurred: course of the accident (type of collision, road users, injuries …), traffic conditions (maximum speed, priority regulation …), environmental conditions (weather, light conditions, time of the accident …), road conditions (road surface, obstacles …), human conditions (fatigue, alcohol …) and geographical conditions (location, physical characteristics …). In total, 572 different attribute values are represented in the data set. On average, 45 attributes are filled out for each accident in the data set.

V. Implementations

Expected Outcomes

» Use of a clustering methodology results are in the optimum number of membership functions.

» It was found that, when the number of clusters was increased, the mean silhouette coefficient, which represents the overall quality of the clustering measurement, was decreased.

» As explained above, by increasing the number of clusters, the R-value increased and the mean silhouette coefficient was decreased. Therefore, to satisfy two different evaluations for the cluster validity, 12 clusters were selected, which more than the minimum number of 10 clusters was obtained from subtractive clustering.

» Twelve clusters were obtained from hierarchical clustering - as the optimum number of clusters, as at this value, the mean silhouette coefficient and R-value converged in the clustering algorithms. Clustering should be applied to the input and output of the training records, which comprised approximately 800 records of the overall used data. The optimum number of clusters and the number of rules should be equal; therefore, 12 rules were created. In addition, each input and output was characterized by 12 membership functions.

» Our procedure was able to identify the best model based on precision (R) and response time (t). MLP model via exhaustive search took the greatest amount of time (2.635 seconds) with the best precision (R-value of 0.89).

Results

In AHC, when data set with N points is given to be clustered, N×N distance (similarity) matrix is produced. At the beginning, every point represents one cluster. Then algorithm finds most similar cluster pairs and combines them into a single cluster. After combining most similar cluster pair, algorithm finds the next most similar cluster pair and combines them. Combining clusters continue until desired number of cluster is reached.

This method divided into two divisions:

1. Accident predicted attributes
2. Accident unpredicted attributes

Table 1 and 2 shows clustering results of Agglomerative Hierarchical clustering algorithm.

True Positive (TP), False Positive (FP), True Negative (TN), False Negative (FN) and Accuracy values have been calculated to evaluate efficiency of algorithm.

Accuracy = (TP+TN) / (TP+FP+TN+FN)

AHC algorithm has successfully distinguished anomaly attributes from normal attributes. Expected behavior from anomaly cluster is that number of members of anomaly cluster increases after accident, and anomaly cluster is supposed to contain only anomalies. Cluster 2 in AHC algorithm has shown anomaly behavior after accident.

Simulation result 1

Sim-Time(s)	TP(%)	FP(%)	TN(%)	FN(%)	Accuracy
0.47	100	0	100	0	100
0.53	100	0	100	0	100
0.58	100	0	100	0	100
0.63	100	0	100	0	100
0.74	100	0	100	0	100

Statistical Rates for Agglomerative Hierarchical Clustering Algorithm

Simulation result 2

Simulation Time	Cluster1	Cluster2
0.47	3,9	30
0.48	17	10
0.49	1,2	12
0.50	6,7	15
0.51	13,14	28
0.52	18,21	19
0.53	25	20
0.54	4	26
0.55	23.8	29
0.56	16,22	1
0.57	24,27	32
0.58	5,11	31

VI. Conclusion And Future Work

The experimental result in this paper that indicate that decentralized data systems used to find clustering of large data sets in a reliable amount of time and produce good accuracy.

The proposed work will categorized in two methods in traffic accident data that is necessary attribute for prediction and unnecessary attributes for prediction. Using chameleon hierarchical clustering algorithm based on the accuracy and response time to clustering the necessary attribute.

The simulation results will produce 100%accuracy to find cluster of predicted attributes in the traffic accident dataset.

The amount of response time will vary and comparing with other technique it will produce lesser time for clustering the large database data.

In future needs to improve the accuracy and response time to using different classification approach or using comparing technique to measure the maximum distance function and produce high accuracy with low response time.

References

1. Particle Swarm Optimization Based Hierarchical Agglomerative Clustering byShafiq Alam, Gillian Dobbie, Patricia Riddle, M. Asif Naeem
2. Decentralized Particle Filter for Joint Individual-Group Tracking by Loris Bazzani, Marco Cristani ,Vittorio Murino.
3. Pharos : A Decentralized and Hierarchical Network Coo ordinate System for Interne Distance Prediction by Yang Chen, Yongqiang Xiong , X iaohui Shi , Beixing Deng , XingLi.
4. Pharos: accurate and decentralized network coordinate system by Y. Chen Y. Xiong X. Shi1 J. Zhu B. Deng1 X. Li.
5. The Statistical Analysis of Crash-Frequency Data: A Review and Assessment of Methodological Alternatives by Dominique Lord.
6. A Method for Decentralized Clustering in Large Multi - Agent Systems by Elth Ogston , Benno Overeinder, Maarten van Steen, and Frances Brazier.
7. A Method for Decentralized Clustering in Large Multi-Agent Systems by Elth Ogston , Benno Overeinder, Maarten van Steen, and Frances Brazier.
8. Traffic Accident Segmentation by Means of Latent Class Clustering by Benoît Depaire Geert Wets, Koen Vanhoof.
9. A Generic Local Algorithm for Mining Data Streams in Large Distributed Systems by Ran Wolff, Kanishka Bhaduri,Member, IEEE, and Hillol Kargupta, Senior Member, IEEE.
10. Hierarchically Distributed Peer-to-Peer Document Clustering and Cluster Summarization by Khaled M. Hammouda and Mohamed S. Kamel,Fellow.