A New Approach in Big Data with Fast Clustering Algorithm

Devisetty Bhargava Sai Kumar[1] , O.Srinivas[2]
[1]M.Tech (CSE), G.V.R&S college of Engineering and Technology., A.P., India.
[2]Asst. Professor, Dept. of Computer Science & Engineering, G.V.R&S college of Engineering and Technology.,, A.P., India.

Abstract: Big data is fast growing technology in IT world. It is the part of data mining. For huge data sets or databases big data is most popular. Clustering is the one of the most important technique in the data mining to find out the similar patterns and similar features as one cluster. Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, and information privacy. In this paper, with the hace theorem an adopted FAST algorithm works in two steps. In the first step, features are divided into clusters by using graph-theoretic clustering methods. In the second step, the most representative feature that is strongly related to target classes is selected from each cluster to form a subset of features. Features in several clusters square measure comparatively freelance; the clustering-based strategy of quick includes a high likelihood of manufacturing a set of helpful and independent options. To make sure the potency of quick, we tend to adopt the economical minimum-spanning tree (MST) agglomeration methodology.

Keywords: Clustering, ACO, MST, Fast Algorithm.

1. Introduction:

Every day 3 billion kilobytes of data are produced and today 90 percent of the data in the web were created within the last two years. Our ability for data making has never been so dominant and massive since the creation of the information technology in the early 19th century [1]. One example like Prime Minister Narendra Modi has discussed with the Pakistan's last Prime Minister Nawaz Sharif about two nation development and interrelated cooperation against terrorism such online debate offer a new resources to logic the public happiness and make feedback in real-time, and are mostly engaging compared to media, such as radio as well as TV broadcasting. In another instance, a public picture distribution site, flickr, which achieved 2.5 million photos per day.Each photo is assumed the size of 2 megabyte, this needs 5 terabytes storage every single day and as an old axiom elaborates that a single picture has value of lacks of words. The pictures on Flicker are a huge tank for us to search the human civilization, social proceedings, public relationships, disasters, and so on, only if we have the power to attach the massive amount of data.

The performance, robustness, and utility of classification algorithms square measure improved once relatively few choices square measure involved inside the classification. Thus, selecting relevant choices for the event of classifiers

has received a wonderful deal of attention. With the aim of choosing a collection of fine choices with regard to the target concepts, feature set alternative is associate degree economical suggests that for reducing property, removing orthogonal data, increasing learning accuracy, and up result quality [2]. Many feature set alternative methods area unit projected and studied for machine learning applications. They will be divided into four broad categories: the Embedded, Wrapper, Filter, and Hybrid approaches. The embedded ways in which incorporate feature selection as space unit of the employment methodology and area unit typically specific to given learning algorithms, and so may be lots of economical than the other three categories. Ancient machine learning algorithms like decision trees or artificial neural networks square measure samples of embedded approaches. The wrapper ways in which use the predictive accuracy of a planned learning algorithmic program to examine the goodness of the chosen sub-sets, the accuracy of the tutorial algorithms is usually high. However, the generality of the chosen choices is limited and additionally the method quality is very large.

Most feature choice criteria in pattern recognition are defined with relation to a specific classifier or cluster of classifiers. For instance, [Kittler, 1980] show strategies for selecting a tiny low set of options that optimizes the expected error of the closest neighbor classier. Similar work has self-addressed feature choice for the Box classifier [Ichino and Sklansky, 1984a], the linear classifier [Ichino and Sklansky, 1984b] and therefore the Bayes classifier [Queiros and Gelsma, 1984]. Other work (aimed at removing feature redundancy once options are extremely correlated) is predicated on activity a principal parts analysis to find a reduced set of new unrelated options defined by combining the original options victimization the eigenvectors [Morgera, 1986; Mucciardi and Gose, 1971] [3]. To our information, the matter of finding the littlest set of mathematician options that is sufficient to construct a uniform hypothesis (regardless of the shape of the hypothesis) which is that the topic of this paper has not been self-addressed.

2. Related Work:

Demirkan as well as Delen (2013) have described some research guidelines such as involved with practical statistics for big data [4]. This indicates utilizing open-source, free-of-charge data/text mining methods also connected business tools (e.g. R, RapidMiner, Weka, Gate, etc.). New techniques {need to|should incorporate solutions for relocating these resources to the cloud as well as produce effective and economical solutions for discovering information and patterns from quite large/big data sets

Constrained Ant Colony improvement Ant Colony improvement with completely different Favor (ACODF) applies ACO to be used in information bunch. The distinction between the ACODF and ACO is every hymenopter in ACODF solely visits a fraction of the full bunch objects and therefore the range of visited objects decreases with each cycle. ACODF conjointly incorporates the methods of simulated hardening associate degreed tournament choice and ends up in an rule that is effective for clusters with clearly outlined boundaries. However, ACODF doesn't handle clusters with discretional shapes, clusters with outliers and bridges between clusters well.

3. Big Data Characteristic (HACE Theorem)

HACE theorem is theorem to model the BIG DATA characteristics. Big Data starts with large-volume, Heterogeneous, Autonomous sources with distributed and decentralized control, and seeks to explore Complex and Evolving relationships among data [1] . These characteristics make it an intense challenge for discovering useful knowledge from the Big Data. In a native sense, we can imagine that a number of blind men are trying to size up a giant elephant (see Fig. 2), which will be the Big Data in this context. The goal of each blind man is to extract conclusion of the elephant according to the part of information he collects during the procedure. Because each individual's opinion is restricted to his native area, it is expected that the blind men will each conclude independently that the elephant "feels" like a rope, a wall, a tree, a mat, or a snake depending on the part each of them is limited to.

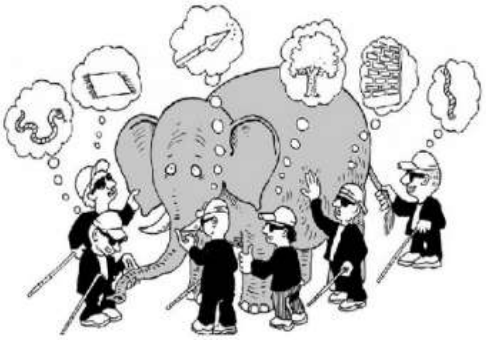

Figure 2. The blind men and the enormous elephant: the restricted view of each blind man leads to a biased conclusion.

To make the problem even more complex, let us accept that 1) the elephant is increasing quickly and its posture varies continually, and 2) each blind man may have his own information sources that tell him about subjective knowledge about the elephant (e.g., one blind man may exchange his feeling about the elephant with another blind man, where the exchanged knowledge is intrinsically subjected). Exploring the Big Data in this scenario is equivalent to form various information from different sources (blind men) to help to draw a best possible illustration to uncover the actual sign of the elephant in a actual way. Certainly, this job is not as simple as enquiring each blind man to designate his spirits about the camel and then getting an skilled to draw one single picture with a joint opinion, regarding that each separate may express a different language (varied and diverse information sources) and they may even have confidentiality concerns about the messages they measured in the information exchange procedure.

4. Data Set:

Our dataset is works on accident dataset for the classification and clustering of the data according to the proposed algorithm ACO. This is the dataset taken from the well known company to make results efficient. The dataset present all the required information for the actions performed by the proposed algorithm. Each and every record contains the attributes. The continuous attributes of all the data sets are discredited before applying it to the Ant Colony Optimization.

5. ACO Algorithm:

ACO [5] is associate rule supported the behavior of the important ants in finding a shortest path from a supply to the food. It utilizes the behavior of the important ants whereas looking for the food. It has been discovered that the ants deposit an explicit quantity of pheromone in its path whereas traveling from its nest to the food. Again whereas returning, the ants area unit subjected to follow the same path marked by the secretion deposit and once more deposit the secretion in its path. During this method the ants following the shorter path area unit expected to come earlier and hence increase the quantity of secretion deposit in its path at a quicker rate than the ants following a extended path.

Ant Colony optimization (ACO) [6] could be a branch of freshly developed swarm intelligence has been used for classification. Swarm intelligence could be a field that studies “the emerging collective intelligence of teams of straightforward agents”. In teams of insects, that sleep in colonies, like ants and bees, a private will solely do straightforward tasks on its own, whereas the colony's cooperative work is that the main reason deciding the intelligent behavior it shows.

Algorithm:

Step: 1 -Check the dataset.
Step: 2- Arrange all the attributes.
Step: 3- Processing all the data
Step: 4-Divide the data into Clusters.
Step: 5-Classification occurs
Step: 6- Classification Done according to the Dataset.
Step: 7-Results

6. Experimental Results:

In this section, the proposed system is implemented by using ACO to get the relation between centroids. Our dataset is population survey dataset. Based on the attributes given in the dataset and no of data in the dataset our results will show the efficiency and performance of the ACO. Our implementation process as follows.

1. Firstly select the CSV file with total number of clusters 10, 100, 1000.
2. Upload the CSV file for processing (dividing) the CSV file into no of attributes.
3. Now select the any of the attribute, form the no of clusters and divide into centroids.
4. If k=no of centroids, given 4, each centroid will form the relevant data in the form of clusters.
5. Csv file data n=1741, k=4, the formation of each centroid is given below.
6. Centroid : 1 @ (514,14) clusters count are : 47
7. Centroid : 2 @ (55,35) clusters count are : 1329
8. C entroid : 3 @ (220,22) clusters count are : 362
9. C entroid : 4 @ (1193,3) clusters count are : 3
10. Cluster Ensemble runs in 0.053328201 seconds.
11. Number of Centroids : 4

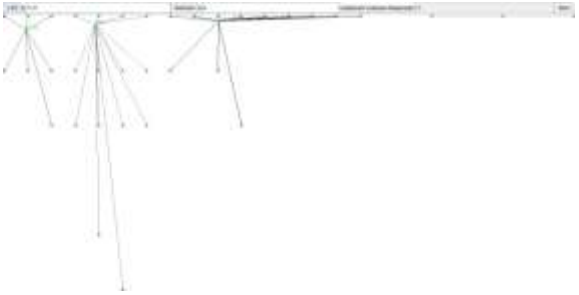

Fig-3, Formation of centriods for the relevant clusters

7. Conclusion:

Clustering in big data becomes the most popular in the current world. In this paper, with the hace theorem an adopted FAST algorithm works in two steps. In the first step, features are divided into clusters by using graph-theoretic clustering methods. In the second step, the most representative feature that is strongly related to target classes is selected from each cluster to form a subset of features. Features in several clusters square measure comparatively freelance; the clustering-based strategy of quick includes a high likelihood of manufacturing a set of helpful and independent options. To make sure the potency of quick, we tend to adopt the economical minimum-spanning tree (MST) agglomeration methodology.

8. References:

[1] Big data analysis using hace theorem, deepak s. tamhane, sultana n. sayyad, y (IJARCET) Volume 4 Issue 1, January 2015.

[2] Ant Colony Optimization in Diverse Engineering Applications: an Overview,R. Geetha, G. Umarani Srikanth.

[3]Ant Colony Optimization and Data Mining,Ioannis Michelakos, Nikolaos Mallios, Elpiniki Papageorgiou, Michael Vassilakopoulos.

[4] HACE-CSA: Heterogeneous, Autonomous, Complex and Evolving based Contextual Structure Analysis of the big data for Data Mining G Manoj Kumar, G.Vara Lakshmi, (IJSETR), Volume 4, Issue 1, January 2015.

[5] Efficient Algorithms for Identifying Relevant Features Hussein Almuallim and Thomas G. Dietterich.

[6]. Ganti, V., Gehrke, J., Ramakrishnan, R.: CACTUS – clustering categorical data using summaries. In Chaudhuri, S., Madigan, D., eds.: Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, ACM Press (1999) 73–83

About Authors:

Devisetty Bhargava Sai Kumar is a student of Computer Science and Engineering from G.V.R&S college of Engineering and Technology. Presently Pursuing M.Tech(CSE) from this college. He received B.Tech from Nalanda Institute of Engineering College of Engineering and technology, budampadu, guntur.

O.SRINIVAS is A Assistant Professor Department of CSE at G.V.R&SCollege of Engineering and Technology in Guntur. He received M.Tech in Computer Science and Engineering from JNTU K. He Gained 2 years Experience in Teaching. He is a Good Researcher in Programming.