Large Scale Data Clustering Using Various-Widths Clustering Approach

Ms. Harshal R.Agashe[1] Prof. S. S.Banait[2]
[1]PG Student [2]Assistant Professor
[1,2]Department of Computer Engineering
[1,2]K. K. W. I. E. E. R., Nashik, Maharashtra, India

Abstract— To perform a clustering widely used and most powerful technique is k-nearest neighbor. This approach required large computational cost for high dimensional datasets. The proposed work focuses on k-NN is based on various clustering widths on large scale data. We are proposing modified kNN approach with MapReduce parallel computing algorithm and clusters grouping with goal of improving the performance in terms of clustering time, pre-processing costs and querying cost while working with high dimensional data. First we are presenting the kNN method using various width clustering to efficiently extract the kNNs for input query object from the dataset. The given dataset is clustered using global width then each cluster that satisfies its predefined criteria i.e threshold value is recursively clustered using their local width. To prune unlikely clusters triangle inequality was used earlier, but we designed tree based approach in which centers of clusters grouped into the tree based index to maximize the more clusters pruning. To reduce the processing time and clustering time, we designed parallel computing algorithm based on MapReduce.

Key words: Clustering, k-Nearest Neighbor, Tree Index, large scale data, Map Reduce

I. Introduction

To perform a clustering widely used and most powerful technique is used i.e k-nearest neighbor. It plays an most used in unsupervised learning, also to measuring an quality of cluster this is mainly used. In clustering quality indexes has been proposed in many years and their different indexes also used in different area. Clustering is useful for grouping similarities also an decision making in the machine learning which including data mining, document information retrieval, image segmentation. Clustering is task of to find homogeneous groups of the studied objects. Many researchers is interested to develop a clustering using various types of algorithms. The clustering main issue is we don't have any kind of knowledge, information of data. Using hidden pattern the quality of clustering will be measure.

In many research domains K-nearest neighbor (kNN) is widely used for information retrieval and classification process. The input set of objects P and test query object Q, the kNN query extracts the k similar objects to Q from the set P. Since from last decade, the kNN problem was extensively studied by various researchers. There are number of methods introduced in order to compute extract or approximate outcomes based on requirements of applications and end users. The existing approximate based methods aiming to achieve more efficiency at the quality accuracy cost. The exact methods are very costly, but they produce the correct results. Therefore it is required to have exact technique over high dimensional data rates with minimum cost. Recently exact kNN based method proposed. This is novel kNN approach based on different widths in the cluster using that we can improve the efficiency. So that to improve this method by adding parallel computation framework and tree based cluster groupings. In single clustering[12] are used with following properties.

Large scale of dataset is task to perform an clustering in limited time and producing an better output also easy to understand it so this is focused on the preprocessing the data i.e construction of index in different distribution of data they are divide into two parts those 1.Tree based indexes.2.Flat based indexes .Tree based indexes is used binary partition method for construct a tree for given dataset. The k-NN is use flat indexes approach and the triangle inequality for efficient to prune the node of tree. Tree based indexes have technique to partition a data to recursive fashion for building the a tree of data .Using Various-width clustering approach and find k-NN search [4]using triangle inequality to efficiently find the query object and computing cost to partitioning the clusters. This process involved following operation those are cluster width learning, partitioning and merging And then construction of cluster into tree like index [8].To reduce the processing time and clustering time we designed parallel computing algorithm.

In the below sections we are going to discuss about related work done for the proposed research area. We refer some existing research paper for completing this task. It is given as follow:

II. Related work

A.Gupta [1] presented a new the method an extraction anomalies of Nuclear Power Plant time sequence created an Data with the help the Fixed Widths Clustering Algorithm. Basically the time data will be recorded successive point in the time. To find their anomalies as well as correlation and pattern of time series. Causes will be found their corrective action are taken. Using a dynamic method is to decide which cluster width will be used for clustering the data. The fixed-width clustering algorithm [1] is based on the outline Anomaly detection are done using fixed width clustering is a three stage process, (1) normalization, (2) cluster formation, and (3) cluster labeling. The anomaly detection are done using brute force method on SAX .The complexity of this algorithm is O(n2) .Here we used it for first dataset because this algorithm will gives us very accurate and it also allows an dimensionality/ numerosity reduction is done to the original dataset. Hence using this algorithm we can reduce the size of original dataset before processing it.

A Ezugwu [2] presented a new approach Performances Characterization of heterogeneous distribution of cluster resources in evaluation technique to the measurement based is characterize into specific performance and their contribution of individual resources of clusters configurations .Also identify the initial parameters are considered at the stage of selection and their allocation an computational node of an all application for execution. For selection of resources parameter (Proceesor speed, associated with the bandwidth and size of memory in each processor) and also their interactions. To map the strategies it is necessary for determining and improvement the performance of an clusters with varying their number of resource specification requirements. Demonstrated [2] it need the resource characterization and how that resources it also could be used for determine and predict to analyse the resource quality and also improve performance for user application which consist having an very high scale.

A .Almalawi [3] proposed a new method an data driven technique clustering to detect an attack on SCADA system .It is method of data acquisition in supervisory control systems having small salient part to control the critical infrastructure, for e.g power plants, various energy grids, various distribution of water systems in that it is automatically identify normal states and critical states of the given water system. Also it extract proximity which is based detection rules which will help to identifying states for monitoring purposes. a novel data-driven clustering approach that removes the need for domain experts and the purely “normal” SCADA data to build the detection models. This approach is based on only about the assumptions that “normal states”, [3] that are represented by a combination of the status and values, of multivariate process parameters in a SCADA system can be clustered into finite groups of dense clusters, and critical states in the *n*-dimensional space will take the form of noise data, also called outliers.

K.Hajebi Hong Zhang[4] was introduced an new method an fast approximate an nearest neighbor by using k nearest neighbor graph. The algorithm construct a graph is offline phase using its nearest neighbor. When query to a new point then it will perform an hill climbing to start with randomly sample to the node in given graph and lazy learner which they doesn’t learn anything from training sample data and used for just classification. We build a k-nearest neighbor (k-NN) graph and to performing an greedy search on an graph to find out the closest node to the given query. Introduce to the Graph Nearest Neighbor Search algorithm (GNNS)[4] and also to analyze their performance. Compare the GNNS algorithm with the help of KD-tree and LSH (Local Sensetive Hashing)methods apply on the real-world dataset as well as a synthetically created or generated dataset.

N.Madicar [5] proposed a new method the time series clustering which are parameter free Subsequences with various width clustering technique.Sub-sequence of Time Series (STS)an clustering of the subsequences in single time of an series which are their subparts or subsequences in that a single time series.It also the similar group of pattern in time series are combine or the cluster centroid is represent an data of every group of cluster. Algorithm will generate an results which is very very similar to the STS clustering algorithm . In addition, this algorithm will perform an better in some cases since that the widths of an various clusters will allowed to vary. This means there results of clustering can containing the clusters with they having different widths, by removing an existing requirement that every subsequence[5] will be clustered that will be must have one same length.

G.Karypis [6] was introduced a new approach CHAMELEON : Hierarchical clustering algorithm using an dynamic modeling. In this algorithm breakdown if incorrect choice of parameter of data is clustered or if they not able to find the property of clusters. In clusters data is consist off diverse shapes, density, and sizes. Using this approach we determine the smiler property of two cluster based on dynamic model. An two clusters will combine their inter-connectivity and close to each other and various type of data and their similarity matrix will be constructed. In this clustering process, first that done only two clusters are merged only if the inter-connectivity of cluster and closeness (proximity) between two clusters will check which they are comparable to the internal inter-connectivity of the clusters and closeness of items within the clusters. The merging process using the dynamic model presented in this paper facilitates discovery of natural and homogeneous clusters.

X.Wang [7] present a new method An fast extract k nearest neighbor in high dimensional data with the help of k-means and triangle inequality. This known as kMkNN (k-Means for k-Nearest Neighbor search)using we can accelerate an finding the nearest neighbors. Preprocess of k-means using an metric trees, kd-trees, or ball-tree, kMkNN [7].Using triangle inequality we reduce the computing distance calculation. An basic an main idea about this is first we can classifying an training objects into the number of different clusters which are regardless of an classes of the training objects, and then for every the given query object *q*, here we are use triangle inequality for avoid distance calculations which having an some training objects in that clusters that are far from query object *q*. The kMkNN algorithm has two stages. In the buildup stage, separatation of an dataset into the number of clusters and record to calculate the distance from each and avery training object to very very closest from the cluster center.

T.Chiang [12] presented a new method for k-NN using ranking based on multilabel classification. Ranking based model is to know to neighbor’s labels which are more comfort candidate using weighted KNN-based strategy after that assign higher weights in which candidates having more vote among all the member. The weight are form using generalized searching pattern technique. So possible to improved the accuracy of multilabel data. This ranking based approach will exploits to learning that which neighbor's labels are very trustable candidates from the weighted KNN-based strategy, and after that we assigns them into higher weights to those candidates[12] in which time we making weighted-voting decisions. The weights can also calculated with the help of generalized pattern searching method.

S.guha[13] was proposed a new method ROCK: A Robust Clustering Algorithm are used their Categorical attributes. In that clustering algorithm to study data having an Boolean and categorical attributes. Using an distances between points for clustering which are not appropriate for Boolean and categorical attributes so we use of links to find the similarities among their data points. This method is non-metric similarities measures in the relevant in situation. So that we exhibits the good scalability of clustering properties.

J.Maillo[17] was proposed a new method A MapReduce based k-Nearest Neighbor Approach for Big Data Classification. This model allows simultaneously classification of large amounts of data. In map phase to determine k-NN from different splits into the data. After that reduce the number of stage will be compute the definitive neighbors which they are getting in the map phase. This model allows the k-NN classifier to scale to datasets arbitrary size, just simply adding more computing nodes if necessary. This model allows us to simultaneously classify large amounts of unseen cases which are totally against a big (training) dataset. To do so, the map phase will determine the k-nearest neighbors in different splits of the data. Afterwards, so that reducing a stage we will compute an definitive neighbors from an list which are obtained during the map phase. The designed model allows the k-Nearest neighbor classifier to scale to datasets of arbitrary size, just by simply adding more computing nodes if necessary.

III. Problem Formulation

To design and develop a system to cluster large scale data using various-widths clustering approach.

IV. System Architecture

Fig 1 describes the overall system structure. The whole system is divided in main 2 blocks. First blocks represents methods which are used and another block will be a functionality which are used.

Figure 1 represents proposed system architecture. Processing of proposed work is takes places as following way:

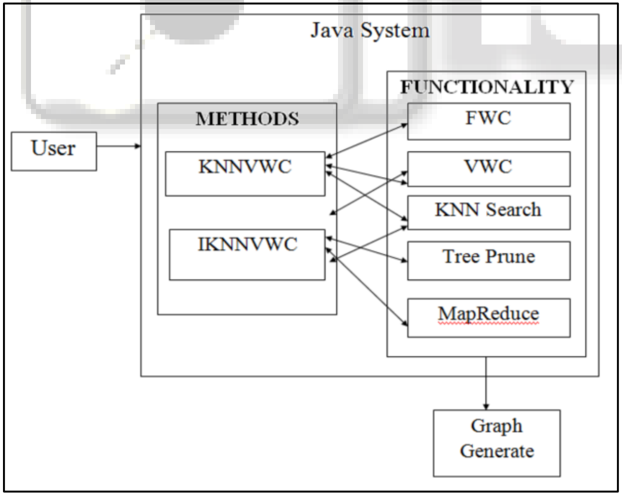

Fig. 1: System architecture

A. KNNVWC

This is an existing method of k-nearest neighbor approach which is based on various-widths clustering. In that using K-NN for searching nearest objects using triangle inequality approach we required low computing cost for partitioning the clusters of various distribution of data.

B. IkNNVWC

This is proposed method for finding best k-NN which is called as ”Improved k-NN”.In this method we can apply various k-NN method to getting an improvement the result by using parallel algorithm i.e MapReduce are used .In that we finding best k-NN methods as compare with existing k-NN.

C. Fix-Widths Clustering

Fixed width clustering will creates the number of clusters having fixed radius (width) w. Here an w is the width a parameter should be define by the user what the user want the width of cluster.

D. Various-Widths Clustering

Clustering will perform using various width approach for large scale data by performing the cluster width learning i.e creation of cluster by using threshold value which is given upto largest size of cluster will be form after they are divided to number of cluster with the help of width suits cluster.

E. k-Nearest Neighbours Search

All instances correspond to points in an n-dimensional Euclidean space. To find nearest neighbor in that cluster.

– Tree Prune: In tree pruning removing the some part of tree after tree has been build. In that various tree pruning methods are used for construction of tree.
– Map-Reduce: Map-Reduce is an programming model for implementation of system generate large dataset using parallel or distribution of cluster.

V. Conclusion

In this review paper, several existing techniques have studied and analysed in section II. Traditional methods of clustering work effectively and efficiently to identify cluster classification methods. Various clustering widths on large scale data to find k-NN by applying Map-Reduce parallel computing algorithm and grouping of clusters to improve the performance. To reduce the processing time and clustering time using parallel computing algorithm based on Map-Reduce to adequately prune additional clusters for various distribution to clusters will be form.

References

[1] Aditya Gupta, Durga Toshniwal”Extracting Anomalies from Time Sequences Derived from Nuclear Power Plant Data by Using Fixed Width Clustering Algorithm”2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI)

[2] Absalom E. Ezugwu,Marc E. Frincu,Sahalu B. Junaidu”Performance Characterization of Heterogeneous Distributed Commodity Cluster Resources”2014 IEEE.

[3] Abdulmohsen Almalawi, Adil Fahad, Zahir Tari, Abdullah Alamri, Rayed AlGhamdi, and Albert Y. Zomaya, Fellow”An Efficient Data-Driven Clustering Technique to Detect Attacks in SCADA Systems”,IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY,VOL. 11, NO. 5, MAY 2016

[4] Kiana Hajebi and Yasin Abbasi-Yadkori and Hossein Shahbazi and Hong Zhang”Fast Approximate Nearest-Neighbor Search with k-Nearest Neighbor Graph”. hajebi, abbasiya, shahbazi, hzhangg@ualberta.ca

[5] Navin Madicar Haemwaan Sivaraks Sura Rodpongpun Chotirat Ann Ratanamahatana,”Parameter-Free

Subsequences Time Series Clustering with Various-width Clusters",2013 5th International Conference on Knowledge and Smart Technology (KST)

[6] George Karypis Eui-Hong (Sam) Han Vipin Kumar"CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling",To Appear in the IEEE Computer

[7] Xueyi Wang "A Fast Exact k-Nearest Neighbors Algorithm for High Dimensional Search Using k-Means Clustering and Triangle Inequality", Proceedings of International Joint Conference on Neural Networks, San Jose, California, USA, July 31 August 5, 2011

[8] Dantong Yu and Aidong Zhang "ClusterTree: Integration of Cluster Representation and Nearest-Neighbor Search for Large Data Sets with High Dimensions",IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 15, NO. 5, SEPTEMBER/OCTOBER 2003

[9] Adrien Gaidon, Zaid Harchaoui, Cordelia Schmid" Recognizing activities with cluster-trees of tracklets" ,BMVC, Sep 2012, Guildford, United Kingdom. 2012.

[10] QING-BA0 LIU, SU DENG, CHANG-HUI LU, BO WANG, YONGFENG ZHOU"RELATIVE DENSITY BASED K-NEAREST NEIGHBORS CLUSTERING ALGORITHM"Proceedings of the Second International Conference on Machine Learning and Cybernetics, Wan, 2-5 November 2003

[11] Chanop Silpa-Anan Richard Hartley"Optimised KD-trees for fast image descriptor matching",2008 IEEE Hosein Alizadeh, Behrouz Minaei-Bidgoli and Saeed K. Amirgholipo

[12] "A New Method for Improving the Performance of K Nearest Neighbor using Clustering Technique"

[13] Sudipto Guha,Rajeev Rastogi,Kyuseok Shim"ROCK: A Robust Clustering Algorithm for Categorical Attributes"

[14] D.A. White, R. Jain, Similarity Indexing with the SS-Tree, Proc. 12th Intl. Conf. Data Eng., pp. 516-523, Feb. 1996.

[15] R. Kurniawati, J.S. Jin, and J.A. Shepherd, The SS+-Tree: An Improved Index Structure for Similarity Searches in a High-Dimensional Feature Space, Proc. SPIE Conf. Storage and Retrieval for Image and Video Databases, pp. 13-24, Feb. 1997.

[16] K. Chakrabarti and S. Mehrotra, The Hybrid Tree: An Index Structure for High Dimensional Feature Spaces, Proc. 16th Intl Conf. Data Eng.,pp. 440-447, Feb. 2000.

[17] J. Maillo , Isaac Triguero" A MapReduce-based k-Nearest Neighbor Approach for Big Data Classification",2015 IEEE,BigDataSE/ISPA.