ISSN XXXX XXXX © 2018 IJESC

Research Article Volume 8 Issue No.5

K-Nearest Neighbor Generation in External Memory Model

Laxmi Sharma[1], Arko Bagchi[2]
M.Tech Student[1], Professor[2]
Department of Computer Science
Delhi College of Technology and Management, India

Abstract:
Clustering is the process of grouping datasets into subsets of data based on some similarity criteria. The process of clustering comes in the category of unsupervised learning. In unsupervised learning method no information about object for right answer is provided. Main objective of clustering is to divide groups having similar features and assign them as clusters. Most of the algorithms of clustering considers that the main memory of a computer system is of finite size and can contain some sets of patterns which are not able to fit in main memory. Most of the part of data set is to be stored in the portion of secondary memory when the size of data set is quite vast. The disk input output operations act as major barrier in developing efficient algorithms of clustering specially for large data sets. Wide variety of different techniques with different designs are used to design clustering algorithms for large data sets. In this paper, the algorithm is designed to maintain the performance in case of large data sets where some values or locations in memory for corresponding values are intermittently accessed. External memory algorithms are suitable to use in large data sets. The algorithm shown in this paper is designed in external memory model and also the input output operational complexity is reduced.

Keywords: Nearest neighbor clustering, Shared nearest neighbor clustering, external memory clustering algorithm, clustering of large datasets

I. INTRODUCTION

Clustering is a process in which data entities are dispensed into several different groups on the basis of similar traits shared by the entities [1]. It works on the method of searching composition in a group of untagged data items [2]. In unsupervised learning method no information about object for right answer is provided [3]. Clustering does not depend on predefined classes. It can uncover undetected relationships in a complex datasets. Main objective of clustering is to divide groups having similar features and assign them as clusters. External memory algorithms are suitable for using in large data sets. In this algorithm the pattern is followed in which same value or related memory locations are frequently accessed. Data objects in each collection are comparatively more homogenous to the object of that collection than those of the other data object collection [4]. There are certain important requirements of clustering algorithm which are: adaptability, should be able to identify clusters having different shapes, should have very less input parameters, and should be stable with respect to noise, should have no effect to the input order of data, should be adaptable according to large data sets. The algorithm should be ideal i.e. if there is increment in data size then the performance of the algorithm should not be decreased. All the clustering algorithms designed does not meet the requirements of scalability of large data sets. Here word large will reform with the transformation in the technology, mainly related to the computing speed of machines and their main memory. According to the growth in the technology the data set in today's era is large may not be large after few years because technology will change and will be able to handle that much of data [5]. But the data size as compared to the technology is increasing at much faster rate. The traditional algorithm is designed with the assumption that main memory size is finite. But in modern era's computers memory have multiple level. From each level of memory there is a different cost and performance characteristics. The data has to be stored in disk of computer if it does not fit in main memory. The main memory access time is quite faster than the disk access time. Approximately all of the clustering algorithm presumes that the size of the main memory is sufficient for the set of data. But in case of large data set, this assumption cannot be considered as realistic. So for large data sets the number of input/output is the main performance measure rather than the computational cost. Different techniques have evolved to design algorithms of large data sets. The input/output model considers the computer as it contains internal memory, processor and external memory (disk). External memory algorithms are certain algorithms which completely uses the locality of reference in the algorithms directly. The size of external memory is considered to be unlimited and external memory is partitioned into blocks of data items which are consecutive. I/O operation is the operation when block of data is transferred between RAM and disk.

II. LITERATURE SURVEY

There are several techniques designed for implementing efficient clustering algorithms and lots of work have been done for developing clustering algorithms which are as follows:
Aggarwal and Vitter in 1988 introduced an external memory model. Willet in 1988 made a survey on the basis of hierarchical clustering model. Hierarchical model of clustering is applied into document clustering. He introduces the two techniques which are Buckshot and Fractionation. Buckshot uses a standard clustering algorithm by pre-clustering the selected small sample of documents and assign the remaining to the cluster formed by pre-clustering the document. Fractionation partitioned the N documents and converted them into m buckets where each of the bucket contains N/m documents. The input parameter ρ is given in Fractionation, which symbolizes that how much each bucket is reduced i.e.

the factor for reduction of each bucket. The standard clustering document is applied and each bucket is checked if each bucket contains n documents then these documents will be clustered into n/ρ clusters. Now every cluster is considered as if it is a separate document and the entire process is repeated till only k clusters are left. A new clustering technique was developed by Zamir and Etzioni in 1998 [6]. They developed a terms known as phrase based document clustering. They just used a general suffix tree and get the data about the phrases and this information is used for clustering the document. Cluster analysis is a term which can be used for both document clustering and clustering. Huang in 1997 gives the terms k-modes which is said to be an extension of the K-means clustering algorithm [7]. This algorithm retains the scaling property of k-means by describing the term of categorical clusters and introduced a rule of incremental update for clusters. K-mode naturally get the disadvantages of k-means like seed cluster dependency and it cannot identify the number of clusters automatically. Gibson et al. in 1998 introduced a method which converts the encoded form of a dataset in a weighted graph structure where each attribute values is associated with the weighted vertices [8]. In this the iteration of various instances of weighted graph takes place by using a combination operator which is user defined, which converges to a fixed point. On reaching the fixed point, the data points are partitioned by using the weight of the basin which give rise to final clusters. This was the argument of the author. In this method the approach of dynamical system was problematic with respect to the type of clusters detected. The approach is non-intuitive if the attribute value is separated by their weight. To maintain the convergence high probability the number of basins is required which is quite significant. The clustering algorithm which was based on the number of links between various tuples. The similar records are captured by the number of links. This method give rise to satisfactory result with regard to the value of attribute which can never occur again in a single tuple. It optimizes the quality function of cluster with regard to the link number in hierarchical manner. The base of the algorithm results in cube complexity of the number of records. This is mainly the reason for unsuitability of large datasets. For using summary information of dataset, a combinatorial search based algorithm was introduced by Ganti et al in 1999. It differentiate the detected categorical clusters which is different from other algorithm.

III. THE TRADITIONAL SNN ALGORITHM

The Shared nearest neighbor is a method in which similarity of two data points is defined. Those two data points which share the common number of neighbor are said to be similar. The number of clusters in shared nearest neighbor algorithm is not specified. The number of clusters is identified automatically by the algorithm. There are several examples in which technique of SNN clustering is used which are temporal clustering and document clustering. The two approaches are followed for SNN clustering which are: figure approach and core approach.

The handling of the SNN algorithm can be done in two ways:

The neighboring data points can be clustered by using the distance which the two data points share or links weights. By specifying the distance in between two data points, a graph can be made.

The traditional shared nearest neighbor algorithm is processed in two steps which are as follows [9]:

- In first step, neighbors of all the data points are determined. These neighbors are k-nearest neighbors. These k-nearest neighbors of a particular point are ordered by ascending. The first point of first row of neighbors is its own data point because each point is neighbor of its own.
- In the second step, the calculation of nearest neighbor of each and every data point is made. Consider two points 'a' and 'b' where a < b. These are the points in which Θ neighbors are matched. Where Θ is a similarity threshold. These 'a' and 'b' are from neighborhood list of each other. Then 'b' which is the bigger index will be replaced by 'a', the smaller index. This means 'a' and 'b' can be considered as similar and 'b' can be represented as 'a'.

IV. PROPOSED ALGORITHM FOR GENERATING K-NEAREST NEIGHBOR

In this paper the k-nearest neighbor algorithm is designed in external memory model which makes it input/output efficient, hence it is suitable for large datasets. The computational complexity is same as that of traditional algorithm computational complexity because the computational steps are same as traditional algorithm but the pattern to access the data is altered for enhancing the scalability of the algorithm.

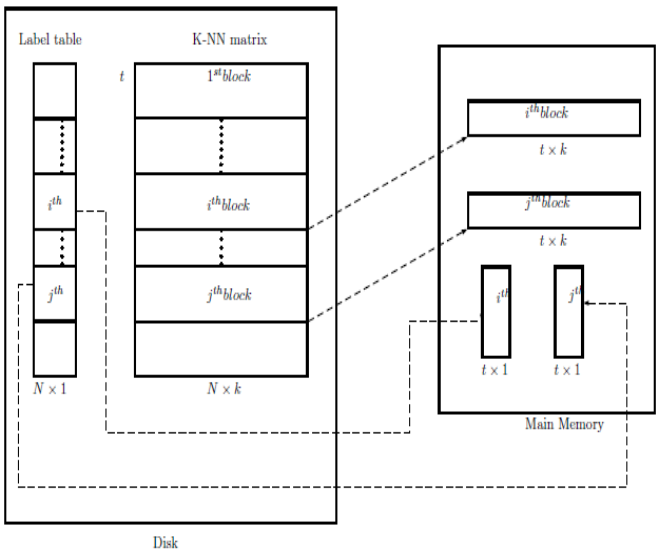

Figure.1.Transfer of blocks between Disk and Main Memory

The proposed algorithm is performed in the following way:

1. In first step of the algorithm, the k-nearest neighbor matrix is generated for efficient Input / Output. Consider the partitioning of N x D datasets into N / t blocks. These each block is of size D x t. Where t is a fixed parameter which depends on the availability of main memory, N are the number of data points and D is the dimension of data points.
2. Read two blocks S_i and S_j into main memory.
3. The corresponding points are indexed in *knn* block of matrix of size k x t. The *knn* matrix block is written in the external memory and k nearest neighbor of block S_i.
4. Repeat the process for N /t times. In this process the *knn* matrix is generated.

Traditional algorithm of shared nearest neighbor algorithm for the number of inputs is O (N^2k^2). But in proposed algorithm,

the main memory consists of two blocks each of size D x t and two blocks of size k x t.
Therefore, M = 2Dt + 2kt.
The input/output complexity of traditional algorithm is $O(N^2k^2)$. The input/output complexity of proposed algorithm is $O(N^2k^2)$ / BM which is an improvement of a BM factor improvement over the traditional algorithm. The performance of the algorithm can be analyzed by the effect of main memory by changing the size of the main memory. The size of the memory is inversely proportional to the input/output, if main memory increases input/output decreases. Thus the dependency of input/output is on the availability of main memory. In main memory the distance between each point pair is calculated. The distance is stored in temporary vector of size k x t.

V. CONCLUSION

In this paper the shared nearest neighbors are generated by following the algorithm which is designed in external memory model. The technique designed here can be used in existing techniques for large data sets. It is shown that the input/output complexity of proposed algorithm is $O(N^2k^2)$ / BM which is an improvement of a BM factor improvement over the traditional algorithm. Now the future work of this paper is to implement the clustering algorithm and analyze the exact performance of the algorithm.

VI. REFERENCES

[1] Sawsan Kanj, Thomas Bruls, and Stephane Gazut – "Shared Nearest Neighbor clustering in a Locality Sensitive Hashing framework".

[2] Data Clustering Algorithms https:/ /sites. google. com/site /dataclusteringalgorithms/home

[3] Unsupervised learning and data clustering towards data sciencehttps://towardsdatascience .com/unsupervised-learning-and-data-clustering-eeecb78b422a

[4] An introduction to clustering and different methods of clustering https://www.analyticsvidhya.com/blog/ 2016/11/an-introduction-to-clustering-and-different-methods-of-clustering

[5] Sonal Kumari, Saurabh Maurya, Poonam Goyal, Navneet Goyal – "Scalable Parallel Algorithms for Shared Nearest Neighbor Clusteringhttps://ieeexplore.ieee. org/document/ 783 9671/

[6] Oren Zamir and Oren Etzioni – "Web Document Clustering: A Feasibility Demonstration"http://citeseerx .ist.psu. edu/ viewdoc/ download?doi=10.1.1.31.75 85&rep= rep 1&type=pdf

[7] Zhexue Huang – "Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values"
http:// arbo r.ee. ntu.edu.tw/~chyun/dmpaper/huanet98.pdf

[8] K. Premalatha and A.M. Natarajan – "A Literature Review on Document Clustering"
https://scialert.net/fulltextmobile/?doi=itj.2010.993.1002

[9] SNN Clustering
http://mlwiki.org/index.php/SNN_Clustering