Optimizing Storage Using Clustering Technique for Tick data

Thesis submitted in partial fulfillment of the requirements for the award of degree of

Master of Technology

in

Computer Science and Engineering

Submitted by

Sabha

Roll No. 801632043

Under the supervision of:

Dr. V.P. Singh
Associate Professor, CSED

Dr. Vinay Gautam
Lecturer, CSED

COMPUTER SCIENCE AND ENGINEERING
DEPARTMENT THAPAR INSTITUTE OF
ENGINEERING AND TECHNOLOGY PATIALA-147004

June 2018

CERTIFICATE

I hereby certify that the work which is being presented in the thesis entitled, "*Optimizing Storage Using Clustering Technique for Tick Data*", in partial fulfillment of the requirements for the award of degree of Master of Engineering in *Computer Science and Engineering* submitted in Computer Science and Engineering Department of Thapar Institute of Engineering and Technology, Patiala, is an authentic record of my own work carried out under the supervision of *Dr. V.P. Singh* and *Dr. Vinay Gautam* refers other researcher's work which are duly listed in the reference section.

The matter presented in the thesis has not been submitted for award of any other degree of this or any other University.

Signature:

(Sabha)

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

9/7/16

(Dr.Vinay Gautam)

Lecturer,CSED

(Dr.V.P. Singh)

AssociateProfessor,CSED

ACKNOWLEDGEMENT

The successful completion of any task would be incomplete without acknowledging the people who made it possible and whose constant guidance and encouragement secured the success.

First of all I wish to acknowledge the benevolence of omnipotent God who gave me strength and courage to overcome all obstacles and showed me the silver lining in the dark clouds with the profound sense of gratitude and heartiest regard. I express my sincere feelings of indebtedness to **Dr. V.P. Singh and Dr. Vinay Gautam** for their positive attitude, excellent guidance, constant encouragement, keen interest, invaluable co-operation, generous attitude and above all their blessings. She has been a source of inspiration for me.

I am grateful to **Dr. Maninder Singh**, Head of Department and **Dr. Ashutosh Mishra**, P.G. Coordinator, Computer Science and Engineering Department, Thapar Institute of Engineering and Technology for the motivation and inspiration for the completion of this thesis. I will be failing in my duty if I don't express my gratitude to **Dr. S.S. Bhatia**, Senior Professor and Dean of Academics Affairs in the institute for making provisions of infrastructure such as Library facilities, Computer Lab equipped with internet facility immensely useful for the learners to equip themselves with latest in the field.

Sabha

(801632043)

ABSTRACT

The clustering problem widespread in day-to-day life where data can be found, mined, or generated for most situations imaginable. To manage huge amount of data, it is require to group the similar type data. As the number of possible data sets grows and the data sets become larger in both number of data points and variables, the automation of this process through clustering algorithms is increasingly important. Tick data is data generated by various applications periodically that is why it is require keeping track the values changing over time and also requiring optimizing redundant data to reduce storage space. Here in this thesis, our aim is to optimize the storage space using clustering technique and to compute time complexity of propose method.

The approach starts with k partitions of tick dataset. The partitions are based on the columns of tick data. After the partition the number of clusters is obtained and then merges the clusters and finally the clusters are obtained in the normalized form. The next step is to construct binary indicator vector that contains binary information generated after matching two concurrent columns and rows. The algorithm also counts the zeroes and ones that occur in the tick data. The next step is to eliminate all the rows which are having duplicate values. The propose approach also compute the compression ratio and execution time that varies as per the number of clusters selected and system configuration. Performance analysis in terms of execution time in seconds varies as per the number of clusters selected and system configuration.

Keywords: Tick data, Clustering Technique, Structure Query Language.

.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

The clustering techniques are used to found, mined, or generate data for most of real time applications. It is a useful practice to group or cluster data that is of similar type. The data is growing rapidly therefore the datasets become larger in both number of data points and variables. The automation of this process through clustering algorithms is increasingly important. Different approaches have been proposed in the past decades, indicating that this problem is neither new nor solved. This chapter describes different clustering techniques. But this research work is based on specific approach, namely Storage Optimization Hierarchal Agglomerative Clustering (SOHAC). SOHAC is a hierarchical clustering approach falls under probability based clustering and centre-based clustering approach[1]. In connection different probability-based approaches have been discussed in the chapter followed by density-based clustering approaches. Clustering is a process of dividing files into collections of comparable objects. Both such collection contains objects that are comparable to each other and different to objects in other collections. Now the clustering approaches are described in this chapter.

1.1 DEFINITION

It is important just before note the difference between supervised learning and unsupervised learning. Supervised learning uses a priori labelled data to classify newly encountered data, whereas in unsupervised learning, no a priori labelled data is available and the cluster structure must be inferred from the data alone as shown in Figure 1.1. Looking at this from a machine learning perspective, resulting clustering structure denotes a data concept. Thus, clustering can be seen as the unsupervised learning of a hidden data concept.

As clustering analysis is an unsupervised learning technique that is applied if no knowledge about the dataset available. Clustering helps to determine the intrinsic grouping from unlabelled training data to predict the unlabelled data from the labelled set of training points.

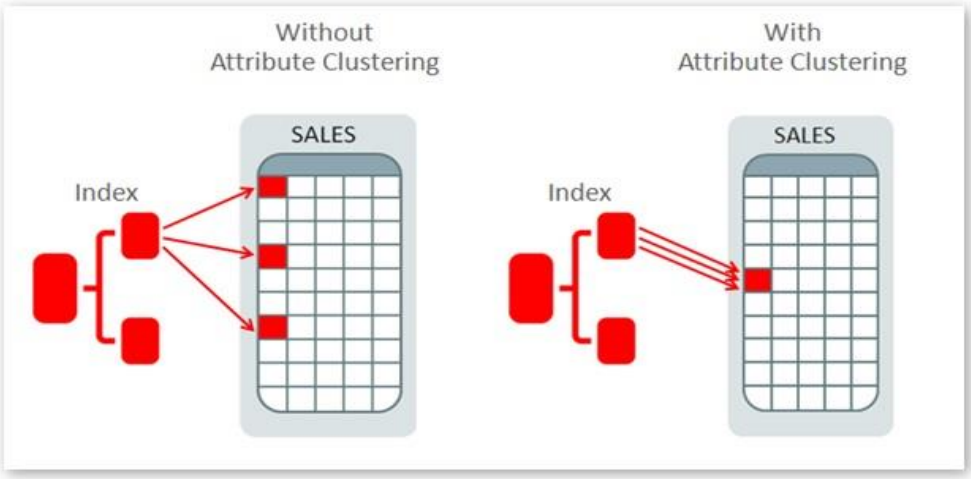


Figure 1.1: The basic illustration of clustering

It is important just before note the difference between supervised learning and unsupervised learning. Supervised learning uses a priori labelled data to classify newly encountered data, whereas in unsupervised learning, no a priori labelled data is available and the cluster structure must be inferred from the data alone as shown in Figure 1.1. Looking at this from a machine learning perspective, resulting clustering structure denotes a data concept. Thus, clustering can be seen as the unsupervised learning of a hidden data concept.

1.2 CLUSTERING PROCESS

Clustering algorithms typically include the following three steps[2]:

1. Definition of model and proximity measure
2. Clustering
3. Validation of the result

In the demonstration step, the structure of clusters is determined. This includes, for example, the number of clusters to be found and details on the features such as type and scale. In the definition step, cluster structure and criteria that separate clusters are defined. Also, a proximity measure is defined that is used in the next step. It may occur that values are missing from the data set.

Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their understanding of what constitutes a cluster and how to efficiently find them. Popular notions of clusters

include groups with small distances between cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem.

Data missing can be separated into three groups[3]: (1) in some attributes, (2) in a number of patterns, and (3) randomly. If one attribute or pattern misses all values, that attribute or pattern should be removed from the data set as shown in Figure 1.2.

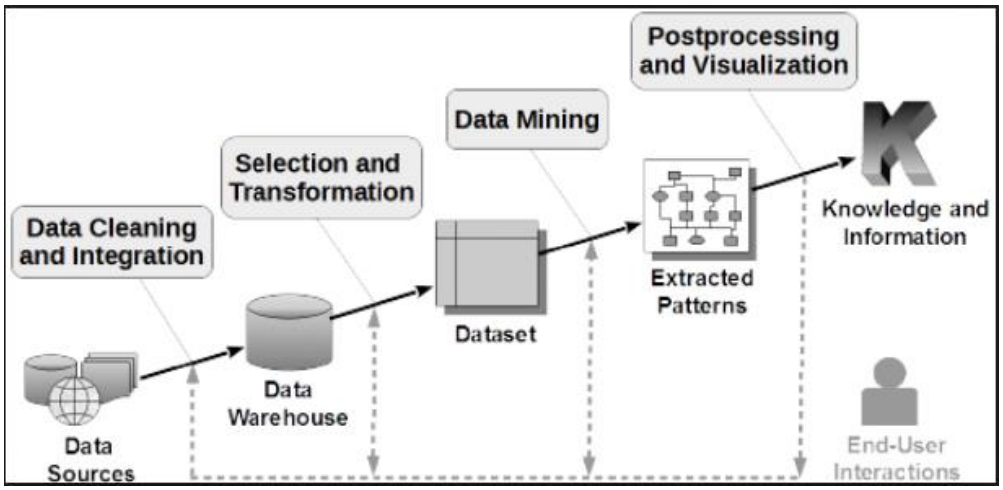


Figure 1.2: Clustering Process

It the no. of missing values is limited; there are two ways to deal with missing values: (1) Replace the missing values before the clustering starts, or (2) Deal with missing values during clustering. Thus, there may be a pre-processing step before the aforementioned steps if many values are missing in the data set.

1.3 DATA TYPES

Data-clustering algorithms depend on the data types that need to be handled by them [4]. A data type can be defined as the degree of quantization in the data. Attributes can be categorized as being discrete or continuous. Discrete attributes have a finite number of possible values. Attributes can be defined as either quantitative or qualitative. Quantitative attributes are associated with numerical data, while qualitative attributes are associated with categorical data.

A special categorical type of attributes is the binary attribute. Binary attributes have exactly two values. Examples include true or false, male or female, and inclusive or exclusive. In real life applications, various more complex data types exist, for example image data or spatial data. In addition, attributes of a single data point may be of

different data types. For such data sets, the chosen similarity or dissimilarity measures need special thought.

1.4 CLUSTERING TECHNIQUES

Clustering is a crucial part of internet data mining techniques which happens to be widely include with diverse areas. Clustering Analysis or clustering algorithms is one of the main analytical associated with data mining[5]. These algorithms are useful to group the user generated data in such a way that number of similar data points generally known as one cluster or (similar cluster) and number of dissimilar data points generally known as second cluster or (dissimilar cluster) in ways that clusters within a group having similar data points is different from other clusters.

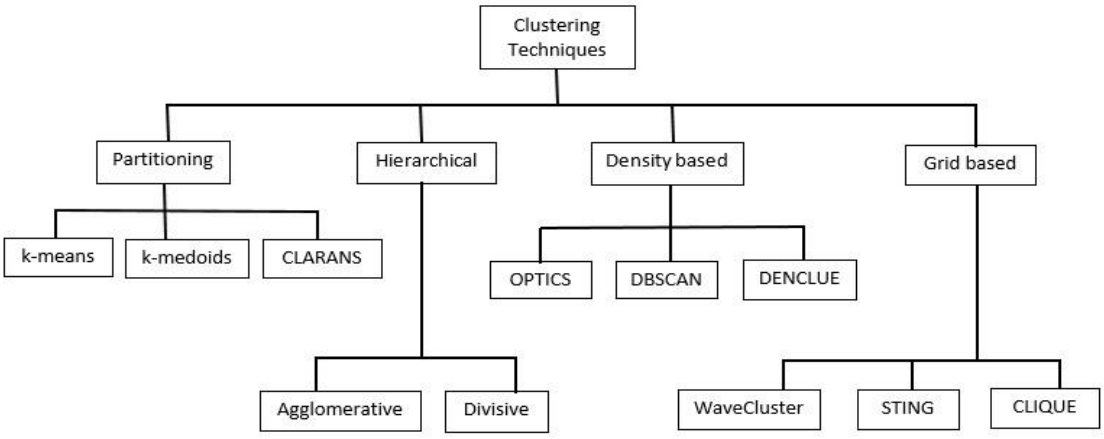


Figure 1.3: Clustering Techniques

Clustering analysis is a major tool used in lot of research areas covering image analysis, data compression, pattern recognition, computer graphics, bioinformatics and information retrieval. As clustering analysis is definitely unsupervised learning technique this really is applied wounded passengers no knowledge for the dataset. Clustering algorithms can be categorized in many ways as shown in Figure 1.3.

Clustering algorithms helps to determine the intrinsic grouping from unlabelled training data to calculate the unlabelled data from the labelled pair training points. A most popular distance measure between the details points is useful for clustering i.e. Euclidean distance.

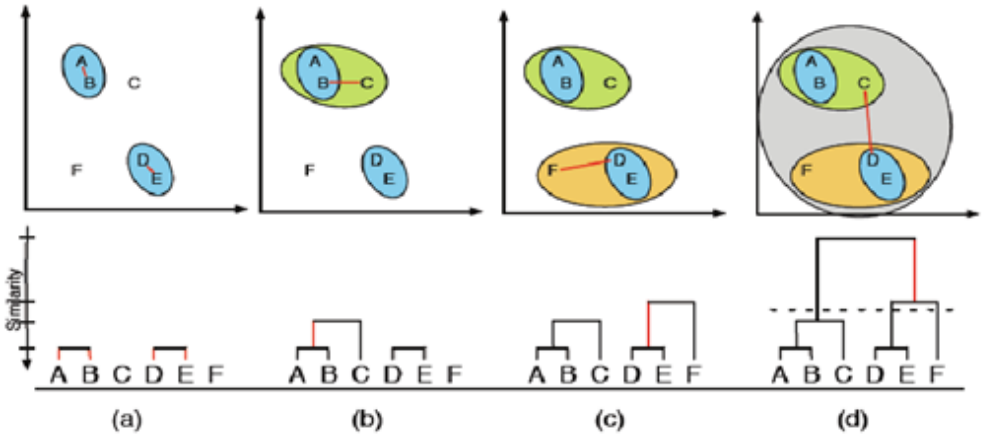


Figure 1.4: Example of hierarchical clustering

There is no straightforward or canonical way to do this. Such groups can also overlap. Common subdivisions include[6]:

- **Hierarchical and partitioning:** Hierarchical clustering algorithms build clusters gradually over multiple levels, while partitioning clustering algorithms create a one-level clustering.
- **Agglomerative and divisive:** Agglomerative methods work bottom-up, starting with one cluster for each object and merging those until a halting criterion is met[7]. Divisive methods work top-down, starting with one cluster of all data points and splitting until a halting criterion is met as shown in Figure 1.4.
- **Monothetic and Polythetic:** Polythetic methods consider all vector features at once; most clustering algorithms take this approach. Monothetic methods look at the vector features sequentially.
- **Hard and fuzzy:** In hard clustering, all objects are assigned to exactly one cluster. Such approaches find strict partitions and thus result in disjoint clusters. In fuzzy clustering, all objects are assigned degrees of membership in several clusters. A function is used to assign this probability. The clusters fuzzy clustering algorithms have as output are not partitions[8].
- **Deterministic and stochastic:** Traditional techniques use a deterministic method to cluster the data and labeling may be used as a stochastic method.
- **Incremental and non-incremental:** If incremental methods may be employed to solve this problem. If not, non-incremental methods suffice.

1.4.1 Hierarchical Clustering

Hierarchical clustering provides many hierarchical decomposition with the given objects[9]. Hierarchical algorithms follow recursive process which is often broken into two approaches: top-down (or divisive approach) and bottom-up (or agglomerative approach). In Agglomerative, it commences with the as anyone cluster and merges the group of objects which can be close together and keeps on merging prior to the termination condition holds. In Divisive, it commences with group of objects in the exact cluster together with a cluster is parse out into several clusters. Hierarchical algorithms are generally known as "nested number of partitions" which is often represented by means of tree structure called dendrogram. Kind's hierarchical algorithm includes agglomerative clustering and divisive clustering.

Finding appropriate clusters to merge or split is done depending on similarity or dissimilarity of objects in the clusters. So, as a certain likeness amongst objects in a cluster is assumed, the (dis) likeness amongst clusters can be used rather than (dis)similarity between individual objects[10]. This generalization is known as linkage metric. The way in which this linkage metric is derived affects nearness. Such methods are called graph methods. All association parameters for hierarchical clustering have, under reasonable assumptions, a time complexity of $o(n)^2$

The graphical representation of hierarchical clustering is a tree structured graph named dendrogram is shown as Figure 1.5.

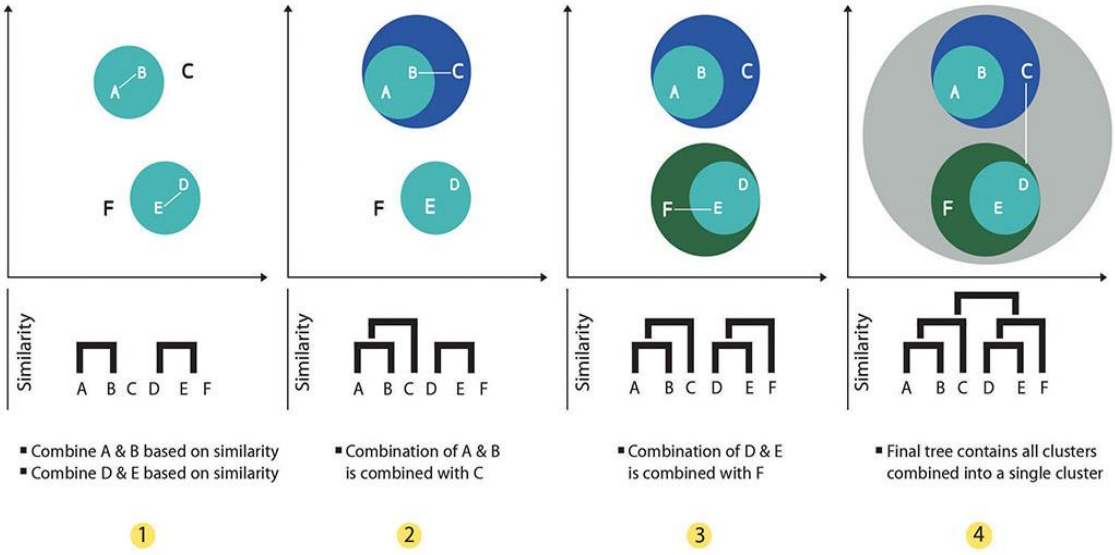


Figure 1.5: The basic illustration of hierarchical clustering

Now, below section give a non-exhaustive list of hierarchical clustering algorithms:

- BIRCH is an agglomerative technique. It was considered to cluster very large mathematical information sets in Euclidean space.
- CHAMELEON is an agglomerative algorithm. It utilizes dynamic modeling in cluster aggregation. It merges two clusters only if interconnectivity and closeness between them are high enough, relative to internal interconnectivity and closeness.
- COBWEB is an algorithm for categorical data. It utilizes incremental learning and also belongs to conceptual or model-based learning. The dendrogram created is called a classification tree. COBWEB does reconsider clustering decisions.
- CURE (Clustering Using Representatives) is an agglomerative technique that is accomplished of identifying non-spherical clusters in large databases and with wide variances in size. It is robust to outliers.
- ECS Method is a divisive algorithm, based on splitting clusters into two new clusters to maximize inter-cluster sum of squares.
- ROCK is an agglomerative algorithm for categorical attributes. It is based on number of links between records, not on any distance function[10].

1.4.2 Centre-based Clustering

Centre-based clustering technique can be divided into 2 approach categories: centroid as well as medoids. In centroid approaches, clusters are represented by gravity centre of data points in that cluster. In medoids approaches, clusters are represented by means of the data points closest to the gravity centre. Centre-based clustering techniques are efficient for huge and great dimensional databases[11].

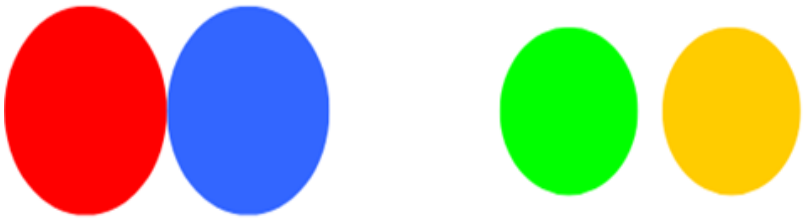


Figure 1.6: Centre based Clustering

The centre based clustering with 4 clusters is shown in Figure 1.6. These techniques find convex shaped clusters. Thus, when clusters of arbitrary shapes have to be found, centre-based approach is likely not the best choice. Even though, this technique is still popular clustering algorithms[12]. It's simple and straightforward. Beside efficiency of centre based algorithms in general, k-means often terminates at a local optimum and is sensitive to outliers. It is not as effective for high-dimensional data as some other algorithms. As k-means algorithm takes the centroid approach to represent clusters, it doesn't work properly with its attributes, while it does work well with mathematical attributes. k-modes is derived from k-means algorithm and was proposed to cluster categorical data. First, k initial modes are selected, to their respective nearest mode. A parametric version and a non-parametric version were proposed.

- K-prototypes originate from k-modes and k-means, designed for clustering of data sets with mixed attribute types.
- K-probabilities extend k-modes algorithm and were, like k-prototypes, designed for the clustering of data sets with mixed attribute types.
- FCM (Fuzzy C-Means) is a fuzzy centroid based clustering algorithm.

Some centre-based medoids clustering algorithms include:

- K is a fixed priori number in which clusters are obtained according to fixed points called cluster centroids[13]. K-Means follows iterative approach in which each object intends to partition into n clusters with nearest mean is shown as Figure 1.7

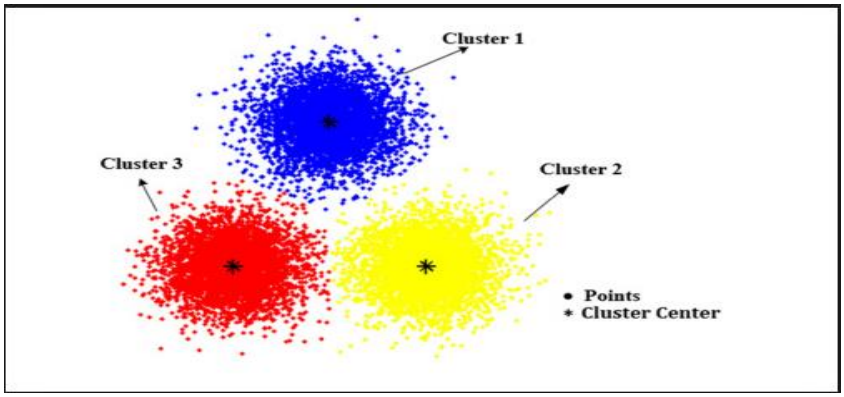


Figure 1.7: K-means Clustering

Steps of Algorithm: Let data points be $x_1,x_2,x_3,\ldots\ldots,x_n$ and centeroid clusters be $c_1,c_2,\ldots\ldots,c_c$.

- Select randomly k cluster centers.
- Compute Euclidean distance function and assign data points and centroids to their closest center of cluster.
- Compute mean or centroid in each value of cluster.
- Repeat all steps until condition is met and points are assigned as cluster.
- K-Medoids technique is a partitioning clustering technique that is extension of K-Means algorithm[14]. K-Medoids makes use of medoids rather than mean to find cluster. K-Medoids clustering attempts to minimalize the average difference amongst points. K-Medoids technique is an object representative technique which overcomes drawbacks of handling noise and outliers in K-Means. It starts with randomly selected n objects as medoids and represents the clusters which have its medoids place near to them-medoids methods represent clusters by one data point. This enables the algorithm to deal with all attribute types.
- K-Medoids is similar to k-means algorithm and is just an extension of it. It overcomes drawback of k-means by handling outliers efficiently. Rather than finding centroid of clusters used for representing clusters. The algorithm starts with finding initially k Medoids and assigning data points to these Medoids depending upon distance between points and Medoids. After, it swaps Medoids with non-Medoids points and finds new rearrangement of clusters until desired membership is obtained. It uses Medoids as they do not depend upon extreme values whereas while calculating mean of clusters, it needs to consider that values also.
- PAM (Partitioning around Medoids) is an iterative algorithm[15], improving found clusters with each step. Medoids are first created by determining a representative data point per cluster. Dissimilarity to all non-center data points is then calculated upon which clustering is then based as shown in Figure 1.8.

Medoids can be re-assigned when an improvement is found.

- CLARA (Clustering Large Applications)[16] uses multiple sample subsets and presents the best clusters found from the sample sets. CLARA overcomes the weakness of k-Medoids algorithm by the method of sampling.

Instead of finding Medoids for whole dataset, it finds samples of datasets of size and applies PAM on each of the obtained sample by finding medoids of each of the sample. If random sampling has been performed in sufficient way, then medoids of sample can approximate the Medoids of whole dataset. So, CLARA creates multiple sample and produce best result out of this. For example, if medoids of whole dataset is not contained in medoids of samples, then it does not produce best results. So, efficiency of CLARA depends on the sample size s. The experiment shows that CLARA produces good results with 5 samples of 40+k size. The complexity of each iteration of algorithms is where s is sample size and n is size of dataset. As compared to complexity of PAM, (n-k) does have any power so producing linear time complexity.

- This algorithm works on the same principle as of CLARA and is based upon randomized search. This technique proposed CLARANS to find clusters in dataset as a search problem in graphs where nodes of graphs represent set of k object indicating selected. Similarly, PAM can also be viewed as a problem of graph to find the minimum in it. Neighbors are examined to find the node with which it can be replaced. This process continues until minimum is obtained. CLARA also uses the same approach as of PAM but it only searches in subgraph examining fewer neighbors than PAM.
- CLARA search for the minimum in many subgraphs which is a time-consuming process. On the other hand, CLARANS also search in subgraphs like CLARA do but, it differs from CLARA in a way that CLARANS assumes subgraph as a sample of neighbors in previous search whereas CLARA assumes subgraphs as a sample of nodes at the start of search process.

Figure 1.8: Partitioning based Clustering

1.4.3 Probability-based Clustering

Most clustering techniques described in the previous sections are deterministic. Algorithms based on those techniques guarantee a local optimal solution. In contrast, stochastic techniques cannot guarantee an optimal solution. However, they generate near-optimal solutions quickly. Also, convergence to optimal solutions is guaranteed asymptotically as shown in Figure 1.9. Stochastic approaches allow for perturbations in directions that are non-optimal (locally) with non-zero probabilities[17].

We mention three different (possibly overlapping) points of view for probability based clustering in the following sections. Section 1.4.3.1 describes search-based clustering, Section 1.4.3.2 pertains evolution-based clustering and finally, Section 1.4.3.3 details model-based clustering.

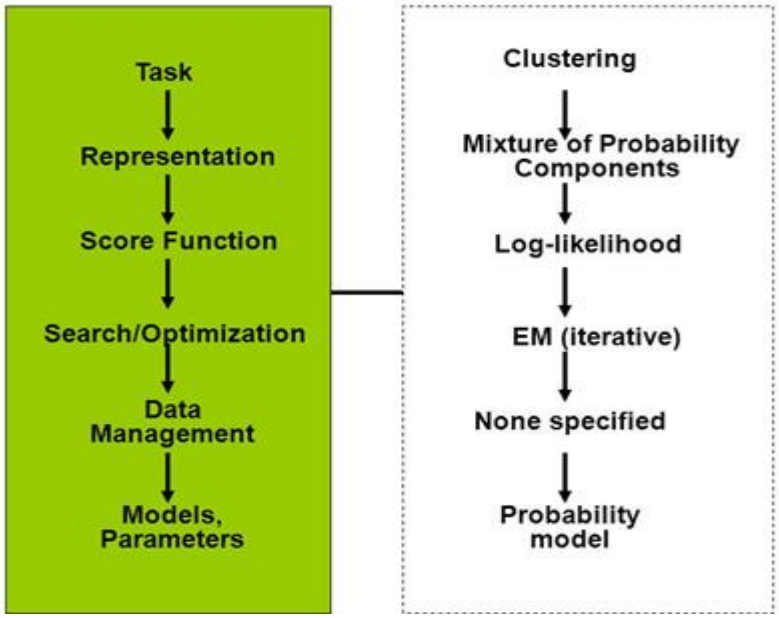


Figure 1.9: Probability-based Clustering

1.4.3.1 Search-based Clustering

A well-known search technique is simulated annealing. This technique has been used to solve clustering problems[18]. The perturbation operator in simulated annealing is similar to the k-means scheme: For example, SARS (Simulated Annealing using Random Sampling) takes the simulated annealing approach, based on decomposition.

For this algorithm to work, the clustering problem is transformed into a graph partitioning problem. SARS explicitly addresses excessive disc access problems during annealing.

Tabu search is similar to simulated annealing. Tabu search is a common heuristic algorithm used to solve combinatorial optimization problems. Tabu search uses steepest descent to improve solutions. After finding a locally optimal solution, some perturbations are done. A number of recent solutions are recorded in the tabu list. These solutions are not allowed to be visited for a no. of repetitions. This allows tabu search to escape local optima. Adaptation of tabu search is a clustering problematic. Some adaptations were also proposed for solving the fuzzy clustering issue. Also, a tabu search technique was combined with the fuzzy c-means technique. Many search-based clustering algorithms are stochastic algorithms, but some deterministic algorithms exist:

1.4.3.2 Evolution-based Clustering

Evolution-based clustering approaches are inspired by natural evolution[19]. GAs has been applied for clustering the most out of the evolution-based clustering approaches. Generally, GAs consists of the following: problem encoding, initialisation. Problem encoding is problem dependent. In addition, the evaluation function used to determine the fitness of a particular solution is also problem dependent. However, even for the same problem, different encodings or evaluation functions may be suitable.

The initialisation phase takes care of the (random) construction of the first population. Typically, GAs then iteratively creates new populations using the genetic operators. Crossover operator takes parent solutions and combines these into new child solutions. The mutation operator takes a solution and modifies it slightly with a certain probability as shown in Figure 1.10.

Data streams have recently attracted attention for their applicability to numerous domains including credit fraud detection, network intrusion detection, and click streams. Stream clustering is a technique that performs cluster analysis of data streams that is able to monitor the results in real time. A data stream is continuously generated sequences of data for which the characteristics of the data evolve over time.

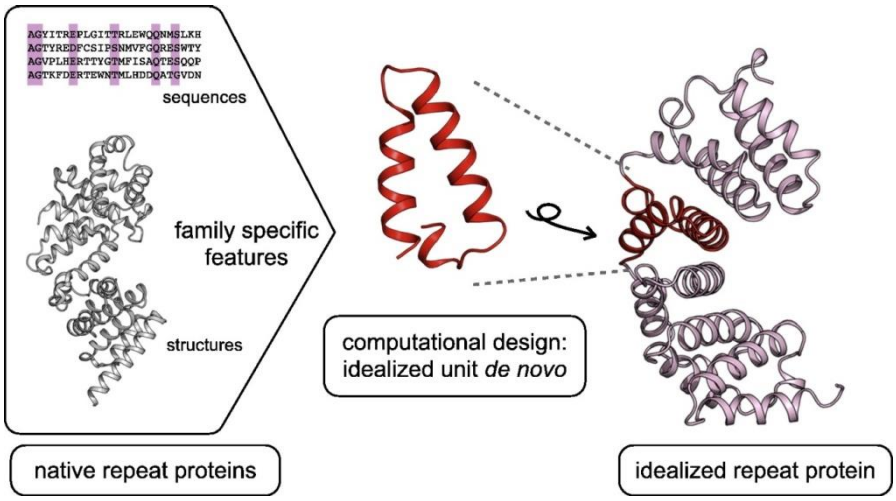


Figure 1.10: Evolution-based Clustering

The selection operator selects a number of new solutions based on survival of the fittest. The best found solution is stored separately from the population and will be the output at the end of all iterations. The sensitivity to the selection of parameters is a major problem in the use of GAs.

- A GA was applied to locate great preliminary bunch centers, following that the k-means Technique was applied for the final clustering. That hybrid approach outperformed the pure GA.
- GKA sees the globally maximum partition of knowledge collection in to confirmed quantity of clusters.
- GKMODE (Genetic k-MODEs) is like to GKA, but uses the k-modes driver and permits illegal strings. It works on the one-step fuzzy k-modes algorithm instead of the crossover driver to increase convergence.

1.4.3.3 Model-based Clustering

Model-based clustering techniques assume which info factors may be classified using a combination of possibility disseminations, wherever every such circulation fits another cluster[20]. The design is frequently applied to signify the kind of limitations and geometric homes of the covariance matrices. Model-based clustering calculations try to improve the match between types and data. This means that the more the info shapes to the design, the better model-based clustering calculations performed. EM (Expectation Maximization) is a general statistical algorithm as shown in Figure 1.11.

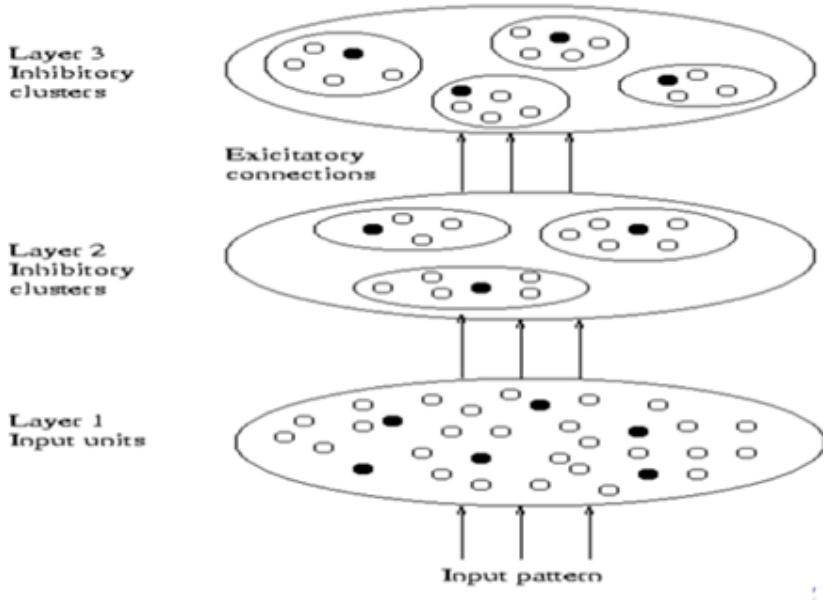


Figure 1.11: Model based Clustering

It can be used information for maximum likelihood estimation. It can be seen as an extension to k-means. EM is simple, stable, and robust to noise.

- COBWEB creates a hierarchical clustering through simple incremental conceptual learning. The number of classes is automatically adjusted.
- CLASSIT extends COBWEB for continuous data clustering. Auto Class estimates the optimal number of clusters through Bayesian statistical analysis.
- COOLCAT clusters categorical attributes using entropy. In an initialization step, a sample from the data set is used to find a set of cluster, after which all other data points are assigned to these clusters in an incremental step.
- STUCCO uses newly defined contrast-sets to find meaningfully different groups. A subset is selected from significant contrast-sets, selected from all possible combinations of attribute values using a tree searching method.
- Artificial Neural Networks (ANNs) are based on their biological counterparts[21]. ANNs only process numerical vectors, are inherently parallel, and may learn interconnection weights adaptively. Learning systems, suffer from issues in balancing plasticity and stability.
- As the EM algorithm is so popular, a framework for model-based clustering was based on it. This algorithm comes under the category of Model-based clustering and is the extension of K-means algorithm. EM decides the

membership of clusters using probabilistic distribution function and finds the best suitable parameter for this function in an iterative manner. It starts with randomly labeling data point's clusters and then estimates initial parameters from this labeled data.

- It finds local maximum of the estimated parameters. For the correct estimate of parameters, number of iterations need to be done which is time-consuming making this algorithm expensive[22]. Also, can produce less realistic results in finite steps.

1.4.4 Density-based Clustering

In this section, we define density-based clustering techniques. We first discuss general density-based clustering in Section 1.4.4.1. Then, we look at grid-based clustering in Section 1.4.4.2.

1.4.4.1 General Density-based Clustering

Clusters grow in any direction, based on density alone[23]. Outliers also do not disturb density-based algorithms. Knowing the number of clusters beforehand is not necessary, since density based clustering algorithms can find the natural number of clusters automatically. In general, scalability is very good, but interpretability is worse than for other clustering approaches. Choosing the density threshold well is of high importance, and a difficult task. Also, a metric space is required, so spatial data clustering is the main application as shown in Figure 1.12.

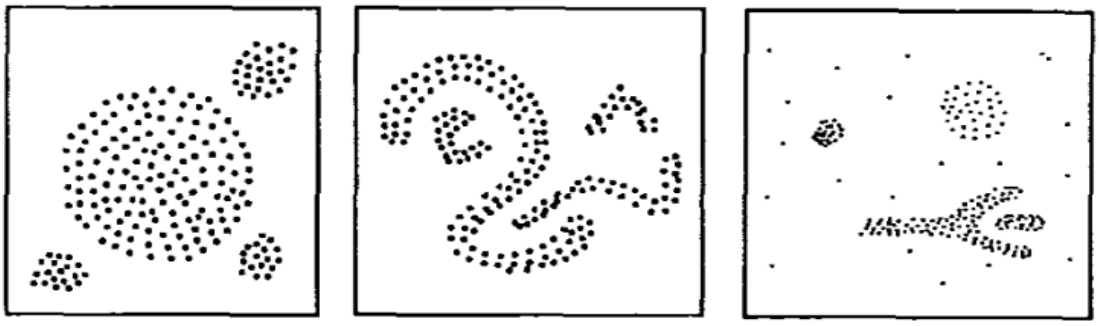


Figure 1.12: Density based Clustering

Two major approaches for density-based clustering algorithms can be identified. In the first approach, density is pinned to training data points. In the second approach,density is pinned to a point in the attributespace. Algorithms for the first approach include:

- **DBSCAN** finds clusters of arbitrary shape and able to find clusters in the high dimensional spatial database[24]. It requires user to specify input parameters which can be a tedious task and may affect the clustering. It uses spatial index for finding neighbors in the effective manner It requires two input parameters from a user which is: size of neighborhood (Eps) and minimal number of points in neighborhood (N). If a neighborhood of each point contains N points, then it is labeled as core point and other points are found connected by above points to form a cluster. Objects are labeled as border points if they have fewer points than N in their neighborhood. Points left are known as outliers or noise. DBSCAN finds clusters of arbitrary shape and able to find clusters in the high dimensional spatial database. It requires user to specify input parameters which can be a tedious task and may affect the clustering. It uses spatial index for finding neighbors in the effective manner which improves its time complexity from $o(n^2)$ to $o(\text{nlog}\,n)$
- **OPTICS**[25] is an extension of DBSCAN and works on the same approach as of DBSCAN. DBSCAN has weakness that clustering output is sensitive to input parameters so that different input parameters provide the different number and different arrangement of clusters. OPTICS overcomes this weakness by creating an ordering of points which can automatically extract clusters in data. The time complexity of OPTICS is similar to DBSCAN $o(n^2)$ to $o(\text{nlog}\,n)$ in the case of indexing structure.
- **DENCLUE(DENsity Based CLUestEring)**[26]: DENCLUE works on a different approach from DBSCAN and OPTICS and uses density function for finding clusters in data. It assumes that objects are influenced by other objects and uses influence function to find it. DENCLUE uses different influence functions, so able to generalize partitional, hierarchical and density-based clustering algorithms depending upon the choice of this function. It can handle outliers very well and can work efficiently on high-dimensional datasets.

Algorithms for the second approach include DENCLUE (DENsity-based CLUstEring). DENCLUE focuses on its density function's local maxima. DENCLUE is superposition of multiple influence functions. The algorithm is very robust regarding noise.

1.4.4.2 Grid-based Clustering

Grid based method are depend on space partitioning rather than data partitioning. Space partitioning is depending on the grid characteristics of the input data whereas data partitioning is about data membership in regions resulted from space partitioning.

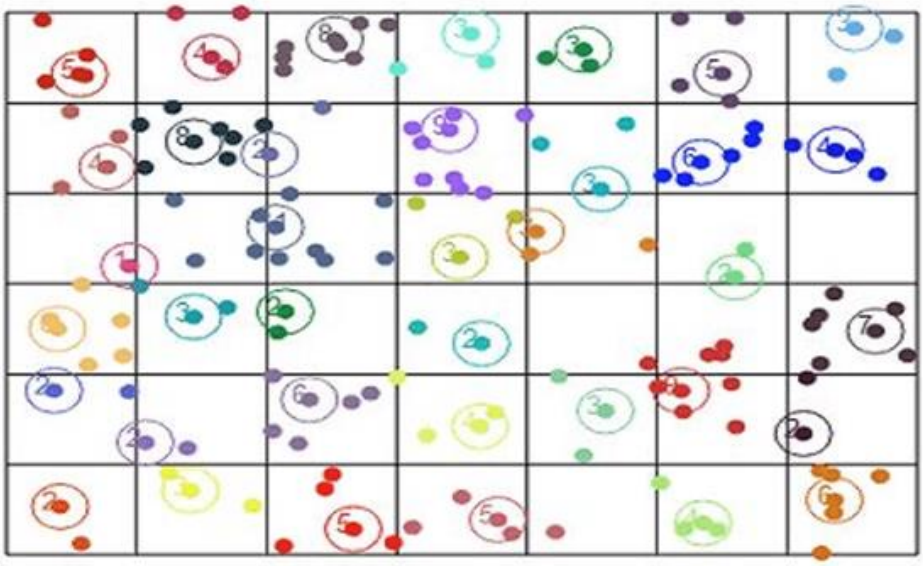


Figure 1.13: Grid-based Clustering

By this way, they become independent from data ordering and can work with data of different data types. Merging of cells in grid and cluster membership is decided by predefined parameter as shown in Figure 1.13. Traditional grid based algorithms are WaveCluster and STING.

- **STING** - The algorithm STING (Statistical Information Grid-based methods) uses hierarchical structure to break the spatial data space into number of cells[27]. They stores statistical information about data in nodes of trees where nodes represent grid cells.

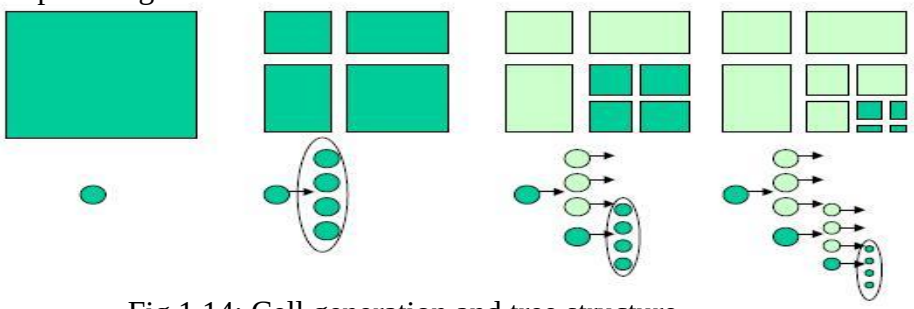


Fig 1.14: Cell generation and tree structure

- For each node in tree, it computes point and attribute-dependent measures: mean variance, minimum, maximum and type of distribution. These parts are summed up as we go higher in the hierarchy as minimum of certain node is equal to minimum of its children. Figure 1.14 represents cell generation and formation of tree structure. STING is highly scalable as new cells can be embedded in the grid easily and region queries can be resolved by scanning only appropriate cells at each level of hierarchy. During cluster formation, grid cells need to be merged depending upon some parameter where only parents are merged but not children and result in formation of clusters having vertical and horizontal boundaries. When Grid is formulated in $o(n)^2$ time, some cells are analyzed and connected to form clusters. It uses multi-resolution method that depends upon the number of leaves in tree.
- **WaveCluster-** WaveCluster is clustering approach which uses different strategy and uses wavelets transforms and multi-resolution method[28]. It uses multi-resolution approach of wavelets to find erratic shaped clusters at different levels of resolution. A wavelet transforms is signal processing method that uses various frequency bands. This method helps to find clusters of data points at different level of detail. WaveCluster can form clusters of high quality and have time complexity of $O(n)$ which makes it efficient than other algorithms. It can handle outliers effectively and able to perform clustering in high-dimensional data. Figure 1.15 shows the application of WaveCluster where leftmost image shows clustering at high resolution and rightmost image shows clustering at lowest resolution.

Fig 1.15: Application of WaveCluster

- CLIQUE[29,30] is a grid-based and density-based algorithm developed to cluster high-dimensional data.
- BANG-clustering is a hierarchical clustering algorithm. The algorithm works with grid-based segments, stored in a so-called special BANG-structure. From this structure, a dendrogram can be calculated.
- FC (Fractal Clustering) is an algorithm for numeric attributes. It uses an incremental structure, which is beneficial for memory, but is data order dependent.
- GRIDCLUS is a hierarchical algorithm developed for large data set clustering.

1.5 Formation of thesis

Chapter 2 Literature Review - It describe review of the main algorithm suggested under the category of storage optimization hierarchical agglomerative clustering algorithm. It also discusses other variants of the main algorithm and provides the comparison based on different datasets.

Chapter 3 Research Problem –It describes the research gaps to formulate problem solution. This discusses the circumstances which led to formulate the problem and set the objectives to achieve problem stated. It also specifies the methodology need to solve the research problem**.**

Chapter 4 Design of Storage Optimization– In this chapter, first existing implementation of algorithm is discussed and then, the design of propose work is explained in detail.

Chapter 5 Experiments and Result – In this chapter the experiments are performed on different tools /technologies using visual studio and c# and then comparison the results.

Chapter 6 Conclusion and Future Scope – This section gives conclusion of the thesis and, summary of the contributions and the future scope.

CHAPTER 2

LITEARTURE SURVEY

This chapter cover study of different clustering techniques which are used to formulate problem. Literature review is as followed:

Mago, Nikhit, et al. (2018) [9] have proposed improvement of internet data. Clustering is typically a Unit Understanding method that allows branding data in to lightweight and dissimilar clusters will help provide some crucial insight. Clustering discovers hidden knowledge that'll neonatologists in identifying neonates who're in peril and also assists in neonatal diagnosis. Alongside, that paper also evaluates the amount of clusters to possess designed for that practices applying Figure Coefficient.

Krishna et al. (2018) [13] have discussed three techniques for choosing the right volume of clusters which is same as knee, outline and gap statistic methods. The three techniques are elbow, silhouette and gap statistic methods. The total sum of 52 drugs (known to complete something against 5-HT receptor) employing their properties the same as Molecular Fat, logP, Big Atoms, H-bond Donors (HBD), H-bond Acceptors (HBA), polar region (PSA), volume of simply rotatable securities (RB) and half-life time within the medication are created by way of a table which was utilized for analysis.

Pizzuti et al. (2018) [18] described genetic algorithm for sensing a residential position style in tracked graphs is proposed. The technique optimizes a training purpose that contains the mixtures of node similarity and structural connectivity. The parts acquired via the method are created by nodes having both similar features and large url density. Checks on artificial techniques along with comparison with five state-of-the-art techniques reveal that genetic method might be really competitive and obtains program divisions more accurate than others acquired via the regarded methods.

Piñeros-Niñe et al. (2017) [30] Hydroxycinnamic acids are phenolic compounds which are having health promotion qualities die to their antioxidant activity. Potato tubers of 113 genotypes of Solanum tuberosum Phureja of the Colombian Key Collection, landraces of potatoes, and professional cultivars were evaluated the content of

hydroxycinnamic acids. This experiment shows an extensive difference in hydroxycinnamic acids articles and germplasm which explodes the breeding programs that was contributed to human health.

Gilpin et al. (2015) [10] Hierarchical clustering is a popular method in several parts with several common algorithms. But, all active purpose to information tools a selfish heuristic algorithm without having any specific target function. In this purpose it can formalize hierarchical clustering being an integer linear development (ILP) disadvantage to a nutritious target purpose and the dendrogram qualities enforced as linear constraints. This technique introduce the fresh purpose of implying the small data units locating the world wide perfect creates more accurate results. Formalizing hierarchical clustering being an ILP with limitations has a few advantages beyond locating the world wide optima. Satisfying the dendrogram limitations as example transitivity may make book matter alterations as an example obtaining hierarchies with overlapping clusterings. It is potential to offer limitations to scribe guidance such as must–url, cannot–url, must–link–before etc. Ultimately, nevertheless correct solvers exists for ILP we show that the simple randomized algorithm along with linear development (LP) peace could be utilized to offer projected solutions faster.

Patra et al. (2015) [7] Cluster evaluation in a large dataset is often a good use of problem generally in most areas of Technology and Engineering. One crucial clustering approach is hierarchical clustering, which parts hierarchical (nested) structures of settled dataset. The sort are not to be found in clustering big dataset as this sort of equally retains entire dataset in major storage or operates dataset repeatedly from extra storage of the machine. Similarly these are possibly significant issues for group evaluation in big datasets.

Krisztian Buza et al. (2014) [1] propose Storage optimization of agglomerative hierarchical clustering algorithm that reduces the storage space of data. The algorithm uses algorithm for decomposing the different attributes of data into matrix. Finally, it computes the compression ratio of the data.

Buza et al. (2011) [19] Clustering could possibly be the simple several distinguished data evaluation exactly how structure big datasets and create a human-understandable overview. This paper concentrate on the case after the information has several

categorical qualities, thus cannot be displayed at a loyal way in the Euclidean space. It is abide by the graph-based paradigm and propose a graph-based genetic algorithm for clustering, the flexibleness that may mainly be as a result of chance of applying different kernels. As the method may naturally be parallelized, while using and testing, it provides the computations about several CPUs. The studies reveal that just in case there are efficiently clusterable information and algorithm products well. It also completes the studies on the basis of correct medical data.

Han et al. (2011) [8] A few dimensional data matrix continues to be respected in several applications. The lossless retention of data matrix alongside produces benefits for storage but moreover process transmission. It proposes a story of data-mining-based retention method comprising of three methods: reordering and class data matrix tips and lines by co-clustering; post-processing to greatly help promote disclose redundancy in data matrix; data retention by the normal compressor. This technique attempted the method for the manufactured dataset and five UCI real-life datasets. The new effects suggest our method may improve retention expenses about 24% or lengthier to 68%. Final effects are linearly proportional to data matrix size, which is obviously faster than other opposition methods.

Tomasev et al.(2011) [2] High-dimensional knowledge happen obviously in lots of domains, and offer frequently a great concern for conventional knowledge mining techniques, equally with regards to efficiency. Clustering becomes hard with the increasing sparsity of such knowledge, in addition to increasing problem in unique ranges between knowledge points. This paper creates a story perception on the issue of clustering high-dimensional data. It validate the idea by showing that hubness is a wonderful method of calculating position centrality still another high-dimensional knowledge chaos, and by proposing many hubness-based clustering remedies, presenting that significant modems could be utilised effortlessly as chaos prototypes or as classes throughout scouting about for centroid-based chaos configurations. New effects demonstrate great effectiveness inside our remedies in numerous settings, especially in the current existence of great levels of noise.

Akram, Q. et al. (2009) [26] This paper investigates the validity of what the law claims of only one price (LOP) in global economic areas by considering the amount, rating and period of inter-market price differentials for credit and lending solutions (‘one-way

arbitrage'). Employing a unique information selection for three substantial money and international deal areas that handle an level of much a lot more than seven weeks at beat size, look for which LOP holds might, but numerous cheaply substantial violations with the LOP arise. The period price mentioning violations is sufficient making this helpful attempting to have one-way arbitrage options as an easy way to minimize credit expenses and/or maximize earnings on provided funds. The technique also describes that such options decrease with all the current pace of the marketplace market and improve with market volatility.

Dionne et al. (2009) [15] compared with conventional techniques keen on intraday data, the strategy has two significant advantages. First, our risk assess comes with a larger informational material since it views all observations. To the full overall risk assess, the method includes special the consequence of arbitrary business durations within the effectuation of arbitrary effects, and for studying the connection between these factors. Ergo, we learn that this information in the full time passed between transactions is firmly related risk evaluation, which is obviously based on predictions from asymmetric-information models available on the market microstructure literature. Next, after the design continues to be predicted, the ivar is generally computed by any trader for time skyline in line with exactly the same information predicated on number prerequisite of choosing data and calculating the design again following the skyline changes. Backtesting effects display our method constitutes respected strategy for testing intraday risk for traders who are able to be effective throughout the market.

Nanopoulos et al. (2009) [13] Cultural tagging is definitely an increasingly common phenomenon with substantial affect for the duration of study It comprehend and realize the Web. For the product range of Internet places aren't self-descriptive, such as for instance photographs, tagging is the only real design of associating these people ideas obviously indicated in text. Consequently, people should assign brands to Internet places, and draw recommenders are usually now being made to promote the re-use of current brands within the consistent way. Nevertheless, a label however and undoubtedly conveys the non-public perception of the customer upon the printed resource. That particular perception has to be taken into consideration when assessing the similarity of places with support of tags. With this report, it concentrates on similarity-based clustering of printed things, which support a few programs in social-

tagging techniques, like information series, offering guidelines, or perhaps the establishment of consumer profiles and the discovery of topics. It shows that the certainly be suggested to recapture and use the multiple prices of similarity reflected with the labels given to same item by different users. The technique proposes the model of items, the labels integral who has given the labels within the multigraph structure. To find clusters of connected things, it raise spectral clustering, a technique effectively employed by the clustering of complicated knowledge, right into a technique that catches multiple prices of similarity between any two items. The tests with two true social-tagging knowledge units show which new technique is more advanced than main-stream spectral clustering that ignores the current existence of multiple prices of similarity considered the items.

Kurucz et al. (2007) [18] assess different heuristics for hierarchical spectral clustering in big call graphs. Spectral clustering without extra heuristics frequently generates really unpredictable turmoil styles or poor clusters which may contain numerous disconnected elements, possibly true that is definitely popular for a number of information places but, to understanding, perhaps not described with the literature. Divide-and-Merge, a recently described postfiltering technique is helpful to remove low quality offices within the binary wood hierarchy. This technique propose a change option that enables elizabeth-way pieces in each point by immediately selection unbalanced or poor clusters before splitting them further.

Ben-David et al. (2006) [12] Stability is often a very popular application to ensure the validity of sample-based algorithms. In clustering its might generally use the variables with the algorithm, like the number *e* of clusters. Whatever the upsurge in acceptance of security in practical applications, there has been small theoretical evaluation applying this notion. This paper presents their state idea of security and analyzes almost all their simple properties. Really remarkably, the one's evaluation is generally that for important taste rating, security is completely dependent upon the behaviour of the target purpose that the clustering algorithm is attempting to minimize. If the target purpose has a unique worldwide minimizer, the algorithm is secure, usually, it's unstable. At the end security is not really a well-suited application to confirm, it depends on the symmetries of the data and that is often unrelated to clustering parameters. This technique show that the latest effects for center-based clustering along with

spectral clustering, and help the ideas by many cases when the behavior of security is counter-intuitive.

Rakhlin et al. (2005) [21] propose the expression of K-means clustering for clinical risk minimization method on a form HK and obviously estimate the protecting quantity in that class. Next, it demonstrates that security of K-means clustering is famous through the geometry of HK in relation to the underlying distribution. It shows that for a unique worldwide minimizer, the clustering choice is secure regarding overall changes of the data, while for the facts of numerous minimizers, the progress samples defines the modify between security and instability. While for finite availability of minimisers the outcome employs from multinomial circulation estimates, the facts of infinite minimizers requires more refined tools. At the end it shows that security with the functions in HK implies security of the centers with the clusters. Because security is generally collection to use within choosing might be clusters in practice, ideally like its evaluation can a start for finding theoretically seated dishes for selecting K.

Ahmad et al. (2004) [30] Many economic schedule techniques are nonstationary and their size faculties are time-dependent. This paper offers an instant point summarization and outlook software to analyse nonstationary, risky and high-frequency schedule data. Multiscale wavelet evaluation is used to separate out the news headlines, cyclical fluctuations and autocorrelation effects. The software may make verbal signs to cause it out each effect. The overview outcome is used to reason regarding the future behaviour daily point and to offer a prediction. Studies round the intra-day National currency spot forex expenses are described. The e-mail handle details are in contrast to a neural process outlook framework.

ZHuang (1998)[4] proposes the k means algorithm. This algorithm uses clustering dataset for the categorical values. It proposes two algorithms that extend the k means algorithm for categorical domain and other one with mixed and numeric domain.

S Guha et al.(1999)[8] develops a hierarchal algorithm that is ROCK which employs a links and distances when the clusters are merged. It extends non metric measures with a relevant situations. Doman Experts are also a main part of Knowledge.

CHAPTER 3

PROBLEM FORMULATION

3.1 Research Gaps

As brought out in chapter 2 that the clustering techniques has been used to optimize the storage space for tick data but still there are few gaps left as given below:

- As discussed above that tick data is data which generates per movement of time so it may happen that some data is redundant which is not removed by using clustering technique to optimize storage.
- To improve the execution speed of search and analytic queries.

3.2 Objectives

The objective of the thesis is:

- To implement the Optimizing Storage Using Clustering Technique algorithm and remove the redundant data.
- To use binary indicator vector approach for row and column reduction.
- To compare the compression ratio of existing and proposed algorithm.

CHAPTER 4

PROPOSED METHODOLOGY

4.1 Proposed Approach

Propose Storage Optimizing Clustering technique which is regarded as a clustering line of a tick data vector.

Our approach starts with k partitions of tick dataset. The partitions are based on the columns of tick data. After the partition we get the number of clusters and finally get the clusters in the normalized form as shown in Table 4.1. The next step is to construct binary indicator vector that contains binary information generated after matching two concurrent columns. The algorithm also counts the zeroes and ones that occur in the tick data. The next step is to eliminate all the rows which are having duplicate values .The propose approach also compute the compression ratio and execution time that varies as per the number of clusters selected and system configuration.

It has been observed that k is usually relatively small: for instance, of the storing the information of financial transaction, the consumer is more interested in the decomposition into K=2 or k=3 partitions. The proposed approach is used to optimize the tick data. Subsequently, most of the object fit to separate clusters. Then, move towards the rows which have only zeros in the standard columns. The cells of such rows may be eliminated in the analyzed decomposition without loss in information. Thus, to be able to determine the number of cells necessary for the storage of the analyzed decomposition, the rows which have only zeros within their typical columns are need to reply upon the cells. Performance analysis in terms of execution time in seconds varies as per the number of clusters selected and system configuration[31]. The variation of tick data in the storage size has also been analysed. Extensive analysis shows that the proposed technique outperforms existing techniques.

Table 4.1: Clustering after Partition

Date	Time	C1-20Microns	C2-3IInfotech	C3-Mindia
1/12/2017	15:01:00	35.7300	6.0600	12403.1000
1/12/2017	15:02:00	36.0500	6.1600	12516.1300
1/12/2017	15:03:00	35.7900	6.1600	12413.9600
1/12/2017	15:04:00	35.8800	6.0900	12532.8000
1/12/2017	15:05:00	35.9000	6.1000	12441.7200
1/12/2017	15:06:00	35.6500	6.0800	12559.4500
1/12/2017	15:07:00	35.6700	6.1300	12399.4600
1/12/2017	15:08:00	35.7200	6.1600	12397.0100
1/12/2017	15:09:00	35.7300	6.2100	12393.2200
1/12/2017	15:10:00	35.7500	6.1700	12529.3000
1/12/2017	15:11:00	36.1100	6.0800	12445.2800
1/12/2017	15:12:00	35.7600	6.0500	12381.8000
1/12/2017	15:13:00	35.9300	6.1300	12401.1900
1/12/2017	15:14:00	36.0800	6.1600	12549.4400
1/12/2017	15:15:00	36.0300	6.1900	12443.2000
1/12/2017	15:16:00	35.7500	6.1300	12395.6500
1/12/2017	15:17:00	35.6300	6.1700	12520.9100
1/12/2017	15:18:00	35.6600	6.1800	12537.6000
1/12/2017	15:19:00	36.0900	6.0800	12479.9400
1/12/2017	15:20:00	35.5800	6.1600	12511.7400
1/12/2017	15:21:00	35.9600	6.1000	12400.4800
1/12/2017	15:22:00	35.8200	6.1500	12398.7300
1/12/2017	15:23:00	35.8800	6.0500	12461.3600
1/12/2017	15:24:00	36.1400	6.1700	12468.3600
1/12/2017	15:25:00	35.8600	6.1600	12463.6900
1/12/2017	15:26:00	35.5800	6.2200	12457.7100
1/12/2017	15:27:00	36.1400	6.0900	12572.4800
1/12/2017	15:28:00	35.8500	6.0800	12555.6900
1/12/2017	15:29:00	35.6100	6.0700	12503.2800
1/12/2017	15:30:00	35.5600	6.0700	12450.4200
1/12/2017	15:31:00	35.6800	6.0700	12459.4400
1/12/2017	15:32:00	35.9600	6.2300	12495.3400
1/12/2017	15:33:00	36.1000	6.1600	12386.3200
1/12/2017	15:34:00	36.0100	6.0700	12412.1400
1/12/2017	15:35:00	35.6700	6.1100	12392.8500
1/12/2017	15:36:00	35.7300	6.1600	12532.7700
1/12/2017	15:37:00	35.6600	6.2200	12459.1200
1/12/2017	15:38:00	35.8500	6.1400	12415.0600
1/12/2017	15:39:00	35.8400	6.2100	12495.4500
1/12/2017	15:40:00	36.0300	6.0500	12488.0000

Binary indicator vector from a tick data

In works of literature, there are numerous clustering algorithms that can make non-overlapping surfaces in a way these surfaces together cover all instances. Thus, one answer for issue explained would be to group columns of a data matrix using one of the main-stream clustering algorithms. As the Table 4.2 shows how binary change sign matrix is derived from a data matrix. Catalogue columns are Date and Time column. Tick information matrix is shown in the top of the figure, as the equivalent indicator matrix is shown in the bottom. List column is the Time column in this example.

Table 4.2: Binary indicator vector

Date	Time	C1-20Microns	C2-3Infotech	C3-Mindia
1/12/2017	15:01:00	1	1	1
1/12/2017	15:02:00	1	1	1
1/12/2017	15:03:00	1	0	1
1/12/2017	15:04:00	1	1	1
1/12/2017	15:05:00	1	1	1
1/12/2017	15:06:00	1	1	1
1/12/2017	15:07:00	1	1	1
1/12/2017	15:08:00	1	1	1
1/12/2017	15:09:00	1	1	1
1/12/2017	15:10:00	1	1	1
1/12/2017	15:11:00	1	1	1
1/12/2017	15:12:00	1	1	1
1/12/2017	15:13:00	1	1	1
1/12/2017	15:14:00	1	1	1
1/12/2017	15:15:00	1	1	1
1/12/2017	15:16:00	1	1	1
1/12/2017	15:17:00	1	1	1
1/12/2017	15:18:00	1	1	1
1/12/2017	15:19:00	1	1	1
1/12/2017	15:20:00	1	1	1
1/12/2017	15:21:00	1	1	1
1/12/2017	15:22:00	1	1	1
1/12/2017	15:23:00	1	1	1
1/12/2017	15:24:00	1	1	1
1/12/2017	15:25:00	1	1	1
1/12/2017	15:26:00	1	1	1
1/12/2017	15:27:00	1	1	1
1/12/2017	15:28:00	1	1	1
1/12/2017	15:29:00	1	1	1
1/12/2017	15:30:00	1	0	1
1/12/2017	15:31:00	1	0	1
1/12/2017	15:32:00	1	1	1
1/12/2017	15:33:00	1	1	1
1/12/2017	15:34:00	1	1	1
1/12/2017	15:35:00	1	1	1
1/12/2017	15:36:00	1	1	1
1/12/2017	15:37:00	1	1	1
1/12/2017	15:38:00	1	1	1
1/12/2017	15:39:00	1	1	1
1/12/2017	15:40:00	1	1	1

In the context of our issue, two typical columns are regarded as similar, should they usually change in same row. To be able to sell vicinity steps from literature, we define a binary change indicator matrix *I* around a tick information matrix *M*. Except entries of the index column, all of the entries of the binary change indicator matrix *I* are often 0 or 1 based on whether or not the worth of a cell in the tick information matrix *M* is equal to value of cell in the exact same column and the prior rows of *M*:

$$\mathrm{I}(i,j) = \begin{cases} M(i,j) \text{if the jth column is the index column in M} \\ 0 \text{ if i } > 1 \; and \; M(\mathrm{i},\mathrm{j}) = \mathrm{M}(\mathrm{i}-1,\mathrm{j}) \\ 1 \text{ otherwise} \end{cases}$$

wherever $M(i.j)\ and\ I(I,j)$ denote the items in the ith line and line of the tick information matrix *M* and binary modify sign matrix *I* respectively.

Following constructing the binary modify sign matrix I, we are able to use its typical columns (i.e., all the columns take index column) as situations in conventional clustering algorithms. Despite proven fact that conventional clustering formulas aren't designed to make maximum partitions with regards to our problem, as we will show in

the tests, if we utilize the partitioning of the columns produced by conventional clustering formulas we are able to obtain considerable changes with respect to the necessary space for storing compared to the event of keeping the initial tick information matrix. In the next part, we produce a clustering algorithm that straight optimizes the space for storing necessary to keep the decomposed tick information matrix.

Algorithm: Optimizing Storage using Clustering Technique for Tick Data

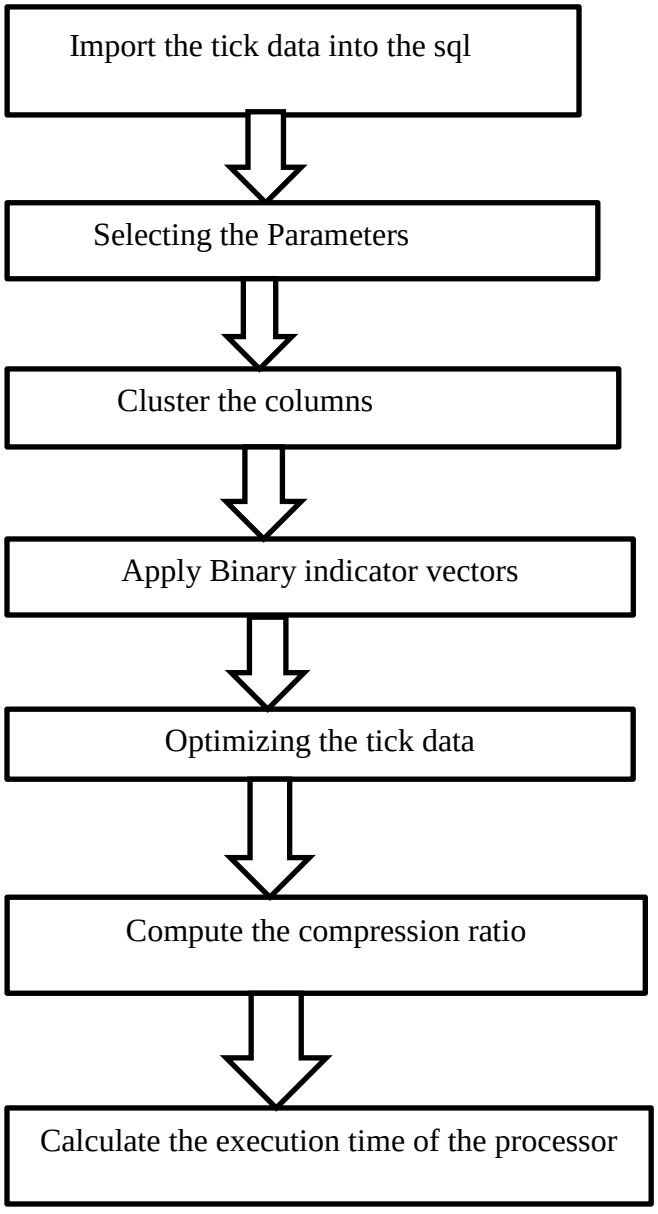
---

Require: object as sender, Event e, Tick information vector M, No. of partition K, Clustering column c.

Ensure: storing the data, clustering of rows and columns, compute the storage, optimize the clusters

1. Load the data parameters from object and event.
2. Get Weather Data from Repository
3. Load the Clusters from the data parameters.
4. Divide the data into number of partitions say k.
5. Cluster the columns c1,c2….cn
6. **while** $|c| > k$ **do**
7. $s \leftarrow \infty$ (Compute the storage size)
8. **for** all pairs of clusters $(C_i, C_j), with\ C_i \in P;\ C_j \in P$ **do**
9. Merge clusters $C_i and\ C_j$ into new cluster $C_i^{'}$ (that is a unique data of date and time)
10. $c' \leftarrow c \setminus \{C_i\} \setminus \{C_j\} \cup \{C_i^{'}\}$ (select the clusters one by one or select all clusters)
11. Create the binary indicator vector *I* from *M*
12. S1= storage size required to store the decomposition
13. **if** $s1 < s2$ **then**
14. $c^* \leftarrow c'$ (The best clustering found so far)
15. $s \leftarrow s1$
16. **end if**
17. Compute Storage Size of Data Per Column.
18. Reduce Repeating Values to Optimize Storage Space
19. Performance Computation in Terms of Seconds varies as per the number of clusters selected and system configuration.

4.2 Flowchart of propose Technique



4.3 Pseudocode

Step 1 – Firstly Import the tick data into the sql. Data records are stored in a fixed var format. There are two basic categories of format i.e. static length and variable length. Static length data types have a fixed length from which they never deviate.

Step 2 – Selecting the Parameters of the Tick data such as symbol, series, ISIN, Date, time.

Step 3 – Cluster the columns of tick data. Merge the unique column into one cluster.

Step 4 – Apply the Binary Indicator vector that contains binary information generated after matching two concurrent columns.

Step 5 – After applying the binary indicator vector, the data which is obtained are optimized.

Step 6 - Then compute the total data volume, optimized data volume, optimized required space, and compression ratio of the tick data.

Step 7 – To calculate the performance of execution time of the processor.

CHAPTER 5

EXPERIMENTS AND RESULTS ANALYSIS

5.1 Experiments analysis

The propose technique is evaluated using Visual Studio tool. The appraisal of propose technique is done on the following parameters such as Total data volume, required space, optimized data, compression ratio, Time to process(in Msec) based on different parameters. For comparison, a microsecond (us or Greek letter mu plus s) is one millionth (10^{-6}) of a second. I had taken two dataset one is stock market and other is weather forecasting. In the stock market Symbol, Series, ISIN, Date, Time parameters are used. On the other hand in the stock market Column type, Column head, Date, Time parameters are used.

The appraisal of propose technique is done on the following parameters such as Total data volume, required space, optimized data, compression ratio, Time to process (in M sec) based on different parameters.

1. Total Data Volume-It is the storage space required (in memory or disk) to hold the entire volume of data. For every buyer, there is a seller, and each transaction contributes to the count of total volume.

$$Total\ Data = \frac{Cell\ Storage\ Size * total\ numbers\ of\ \ Rows * (total\ numbers\ of\ columns - index\ columns)}{1024.0\ F * 1024.0\ F}$$

2. **Optimized Data Volume**-It is the volume of disk space required to hold data after reducing the repeating rows/columns.

$$Total\ Optimized\ Data = \frac{Cell\ Storage\ Size * total\ numbers\ of\ Optimized\ Rows * (total\ numbers\ of\ columns - index\ columns)}{Unit\ Conversion}$$

3. Optimized Space-It is the number of bytes that can be saved after running the optimization algorithm.

$$Required\ Space = Total\ Data\ volume - Optimized\ Data$$

4. Compression Ratio- It is illustrating as the ratio of number of bits saved after the compression to the number of bits required before compression.

$$Compression\ Ratio = \frac{Total\ Data\ Volume}{Optimized\ Space}$$

5. Time To Process - Time is taken to accomplish the method in ms. An ms is one-thousandth of another and is generally utilized in calculating the time for you to read or write from the hard drive.

5.2 Results Analysis

The comparison of compression ratio for existing and propose approach for different datasets by using two partitions as shown in Table 5.1 and Figure 5.2.

Table 5.1: Compression Ratio of existing and propose results at k=2

DataSets / No of Partitions	Weather Forecast		Stock Market	
	SOHACC	OSC	SOHACC	OSC
K=2	1.31787	1.93199	1.08388	1.4484

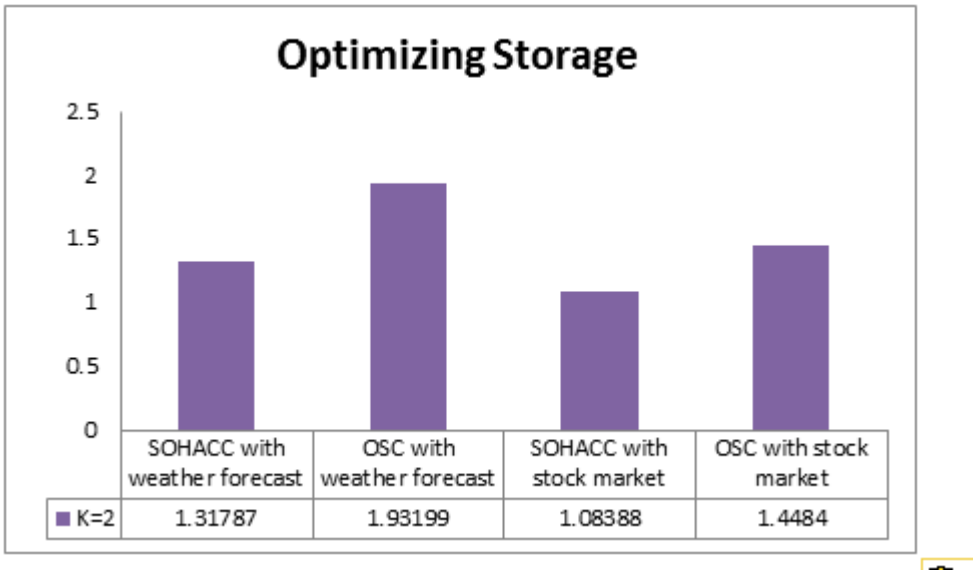


Figure 5.2: Storage optimization for tick data at k=2

The comparison of compression ratio for existing and propose approach for different datasets by using three partitions as shown in Table 5.3 and Figure 5.4.

Table 5.3: Compression Ratio of existing and propose results at k=3

DataSets / No of Partitions	Weather Forecast		Stock Market	
	SOHACC	OSC	SOHACC	OSC
K=3	1.04206	1.08782	1.01592	1.06688

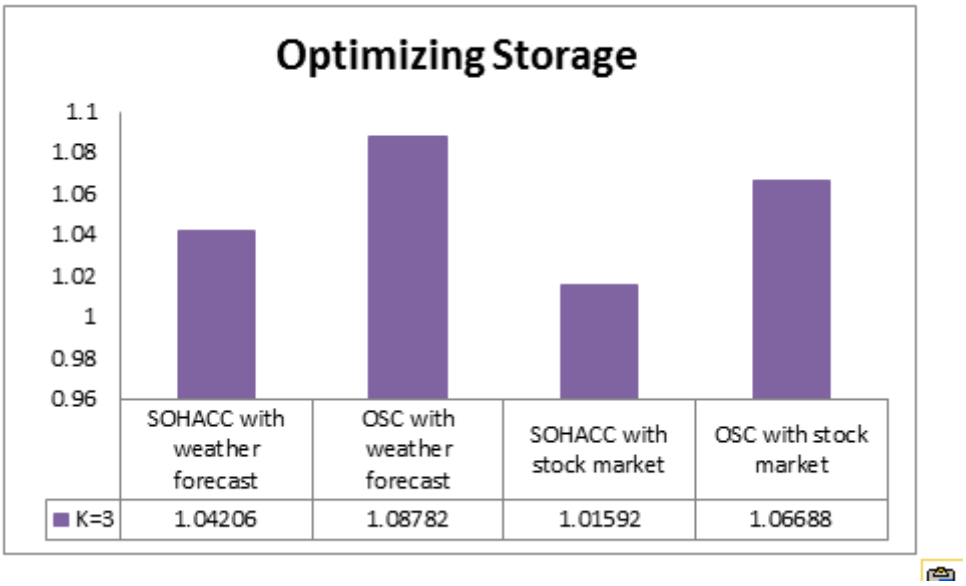


Figure 5.4: Storage optimization for tick data at k=3

The comparison of compression ratio for existing and propose approach for different datasets by using four partitions as shown in Table 5.5 and Figure 5.6.

Table 5.5: Compression Ratio of existing and propose results at k=4

DataSets / No of Partitions	Weather Forecast		Stock Market	
	SOHACC	OSC	SOHACC	OSC
K=4	1.04135	1.08627	1.0007	1.0314

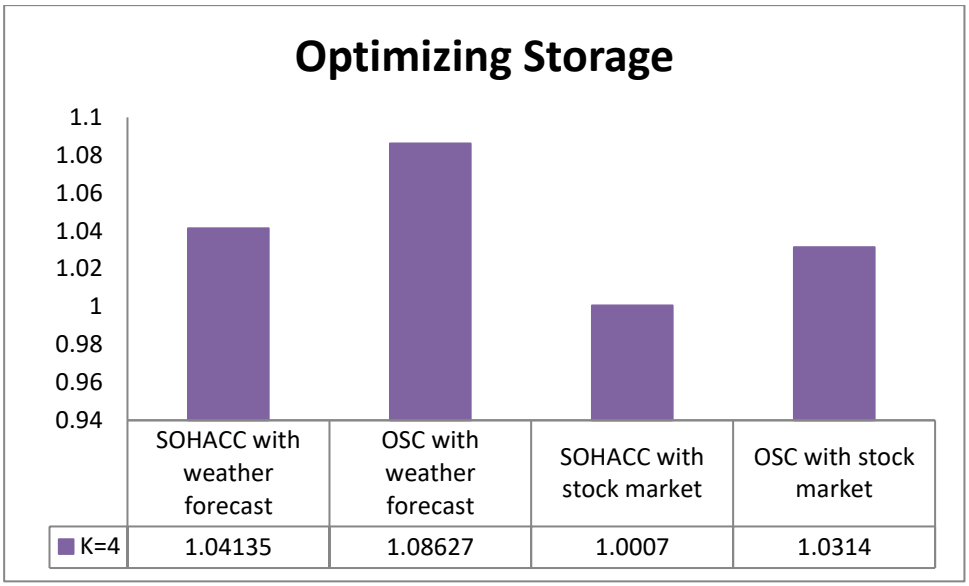


Figure 5.6: Storage optimization for tick data at k=4

First, datasets need to be uploaded on SQL. This task is accomplished in steps: firstly, directory is made on SQL and then, data is uploaded in the directory created as shown in Figure 5.7

	ShareID	DATE1	OPEN_PRICE	HIGH_PRICE	LOW_PRICE	CLOSE_PRICE	LAST_PRICE	PREV_CLOSE	TTL_TRD_QNTY	TTL_TRD_VAL	TTL_TRADES
1	1	2016-04-12	32.00	32.00	31.10	31.25	31.45	31.65	14816	465688.10	132
2	1	2016-04-13	31.80	35.90	31.20	35.00	35.00	31.25	396181	13824951.70	1967
3	1	2016-04-18	34.95	36.00	33.50	34.20	34.85	35.00	82109	2893153.95	605
4	1	2016-04-20	35.45	35.45	33.25	33.75	33.85	34.20	30807	1045957.15	206
5	1	2016-04-21	33.90	35.00	32.25	33.00	33.30	33.75	48174	1593805.60	253
6	1	2016-04-22	32.20	33.30	32.15	32.40	32.70	33.00	27303	888744.90	224
7	1	2016-04-25	32.50	32.80	32.00	32.25	32.35	32.40	18081	584973.15	115
8	1	2016-04-26	31.50	33.40	31.50	32.05	32.00	32.25	9142	294870.15	119

Figure 5.7: Loading of the dataset in the Sql

In the First experiment, k=2 means two partitions are selected from the dataset of stock market. The parameters such as Total data volume, Optimized data volume, optimized required space and compression ratio are calculated and shown in Figure5.8.

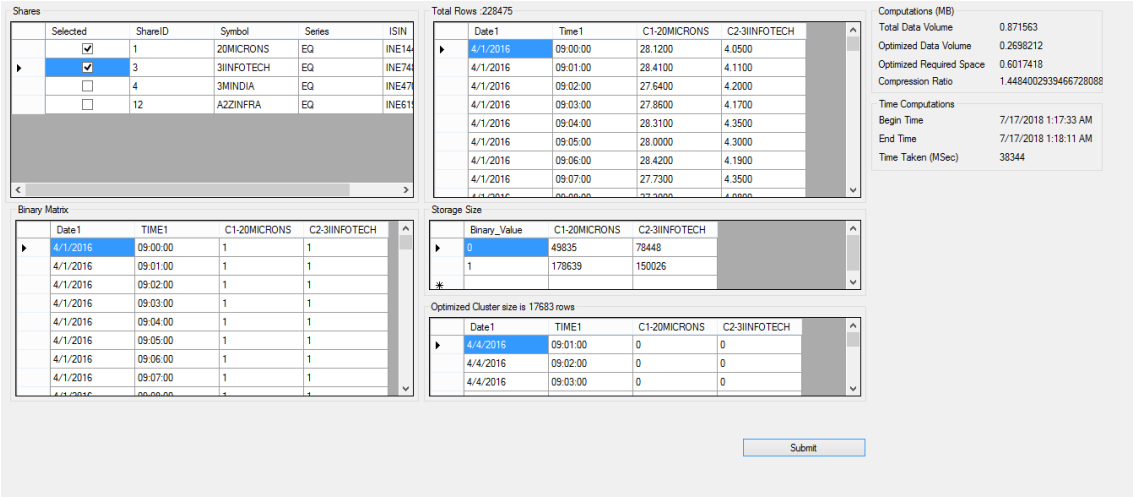


Figure 5.8: selection of two partitions from the dataset of stock market

In the second experiment, k=3 means three partitions are selected from the dataset of stock market. The parameters such as Total data volume, Optimized data volume, optimized required space and compression ratio are calculated and shown in Figure 5.9.

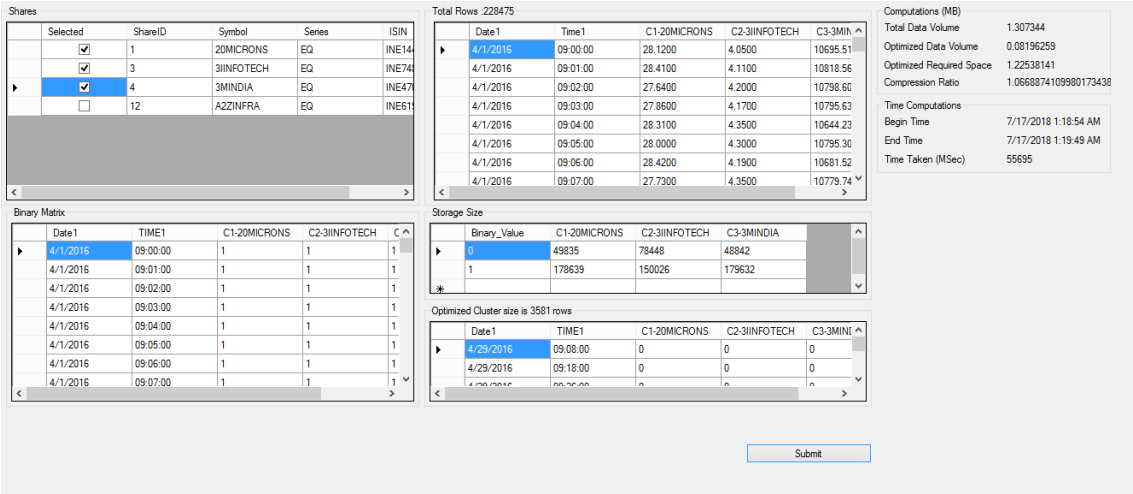


Figure 5.9: selection of three partitions from the dataset of stock market

In the Third experiment, k=4 means four partitions are selected from the dataset of stock market. The parameters such as Total data volume, Optimized data volume, optimized required space and compression ratio are calculated and shown in Figure 5.10

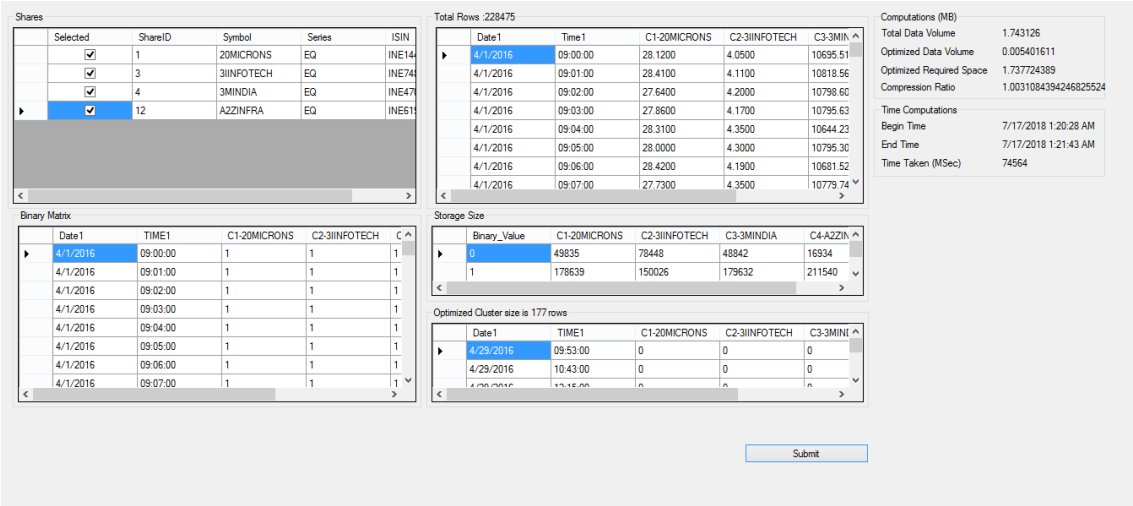


Figure 5.10: selection of four partitions from the dataset of stock market

CHAPTER 6

CONCLUSION AND FUTURE WORK

Tick data is data generated by various applications periodically that is why it is require keeping track the values changing over time and also requiring optimizing redundant data to reduce storage space. Here in this thesis, our aim is to optimize the storage space using clustering technique and to compute time complexity of propose method.

Propose technique starts with k partitions of tick dataset. The partitions are based on the columns of tick data. After the partition the number of clusters is obtained and the merge the clusters and finally the clusters are obtained in the normalized form. The next step is to construct binary indicator vector that contains binary information generated after matching two concurrent columns and rows. This algorithm also counts the zeroes and ones that occur in the tick data. The next step is to eliminate all the rows which are having duplicate values .The propose approach also compute the compression ratio and execution time that varies as per the number of clusters selected and system configuration. Performance analysis in terms of execution time in seconds varies as per the number of clusters selected and system configuration. The variation of tick data in the storage size has also been analysed. Extensive analysis shows that the proposed technique outperforms existing techniques.

Future Work

In the Future work one may implement local research, genetic algorithm for obtaining perfect partitioning. In order to reduce the computation time of an algorithm regression model will be used that is based on time series techniques. Also in near future n (number of hubs) Hub based algorithm will be used in which few columns of tick data will be treated as hubs. Factorization algorithm is also one of the techniques which could cluster the columns. Multivariate time collection will be applied for the storage of data.

PUBLICATION

Sabha, Dr.V.P. Singh, Dr.Vinay Gautam, “Optimizing Storage Using Clustering Technique for Tick data” (communicated to International Journal of Engineering Science and Computing)

REFERENCES

[1] G I. Nagy and K. Buza. "Sohac: Efficient storage of tick data that supports search and analysis." *Industrial Conference on Data Mining*. Springer, Berlin, Heidelberg, 2012.

[2] N.Tomašev,M.Radovanovic,D.Mladenic and M.Lvanovic. "The role of hubness in clustering high-dimensional data." *IEEE Transactions on Knowledge & Data Engineering* 1 (2013): 1.

[3] Tan, Pang-Ning. Introduction to data mining. Pearson Education India, 2006.

[4] Z. Huang. "Extensions to the k-means algorithm for clustering large data sets with categorical values." *Data mining and knowledge discovery* 2.3 (1998): 283-304.

[5] P. Berkhin. "A survey of clustering data mining techniques." *Grouping multidimensional data*. Springer, Berlin, Heidelberg, 2006. 25-71.

[6] T. Kanungo, DM.Mount, NS.Netanyahu,CD.Piatko,R.Silverman and AY.Wu. "An efficient k-means clustering algorithm: Analysis and implementation." *IEEE Transactions on Pattern Analysis & Machine Intelligence* 7 (2002): 881-892.

[7] BK. Patra, and S. Nandi. "Effective data summarization for hierarchical clustering in large datasets." *Knowledge and Information Systems* 42.1 (2015): 1-20.

[8] B.Han and Z.Yang. "Data matrix compression by using co-clustering." *Fuzzy Systems and Knowledge Discovery (FSKD), 2011 Eighth International Conference on*. Vol. 4. IEEE, 2011.

[9] N.Mago, RD.Shirwaikar, UD.Acharya, KG.Hegde,LES.Lewis and M.Shivkumar. "Partition and Hierarchical Based Clustering Techniques for Analysis of Neonatal Data." *Proceedings of International Conference on Cognition and Recognition*. Springer, Singapore, 2018.

[10] S.Gilpin and L.Davidson. "A flexible ILP formulation for hierarchical clustering." *Artificial Intelligence* 244 (2017): 95-109.

[11] S.Guha, R. Rastogi, and K. Shim. "ROCK: A robust clustering algorithm for categorical attributes." *Data Engineering, 1999. Proceedings., 15th International Conference on*. IEEE, 1999.

[12] S.Ben-David, U.VonLuxburg, and D.Pál. "A sober look at clustering stability." International Conference on Computational Learning Theory. Springer, Berlin, Heidelberg, 2006.

[13] A.Nanopoulos, HH.Gabriel, and M.Spiliopoulou. "Spectral clustering in social-tagging systems." *International Conference on Web Information Systems Engineering*. Springer, Berlin, Heidelberg, 2009..

[14] TVS.Krishna, AY.Babu, and RK.Kumar. "Determination of Optimal Clusters for a Non-hierarchical Clustering Paradigm K-Means Algorithm." *Proceedings of International Conference on Computational Intelligence and Data Engineering*. Springer, Singapore, 2018.

[15] T.Velmurugan,. "Efficiency of k-means and k-medoids algorithms for clustering arbitrary data points." Int. J. Computer Technology & Applications 3.5 (2012): 1758-1764.

[16] FB.AI Abid. "A Novel Approach for PAM Clustering Method." *International Journal of Computer Applications* 86.17 (2014).

[17] RT.Ng, and J.Han. "CLARANS: A method for clustering objects for spatial data mining." *IEEE transactions on knowledge and data engineering* 14.5 (2002): 1003-1016.

[18] M.Kurucz,A.Benczur,K.Csalogany and L.Lukacs. "Spectral clustering in telephone call graphs." *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*. ACM, 2007.

[19] C.Pizzuti, and A.Socievole. "A Genetic Algorithm for Community Detection in Attributed Graphs." *International Conference on the Applications of Evolutionary Computation*. Springer, Cham, 2018.

[20] K.Buza, A.Buza, and PB. Kis. "A distributed genetic algorithm for graph-based clustering." *Man-Machine Interactions 2*. Springer, Berlin, Heidelberg, 2011. 323-331..

[21] R.Xu and D. Wunsch. "Survey of clustering algorithms." *IEEE Transactions on neural networks* 16.3 (2005): 645-678.

[22] A.Rakhlin, and A.Caponnetto. "Stability of $ k $-means clustering." *Advances in neural information processing systems*. 2007.

[23] J.Swarndeep Saket and S.Pandya. "An Overview of Partitioning Algorithms in Clustering Techniques."

[24] M.Ester,HP.Kriegel,J.Sander and X.Xu. "A density-based algorithm for discovering clusters in large spatial databases with noise." *Kdd*. Vol. 96. No. 34. 1996.

[25] M.Ankerst,MM.Breunig,HP.Kriegel and J.Sander. "OPTICS: ordering points to identify the clustering structure." *ACM Sigmod record*. Vol. 28. No. 2. ACM, 1999.

[26] H.Rehioui,A.Idrissi,M.Abourezq and F.Zegrari. "DENCLUE-IM: A new approach for big data clustering." *Procedia Computer Science* 83 (2016): 560-567.

[27] QF.Akram, R.Dagfinn , and L.Sarno. "Does the law of one price hold in international financial markets? Evidence from tick data." *Journal of Banking & Finance* 33.10 (2009): 1741-1754.

[28] C.Piñeros-Niño,CE.Narvaez-Cuenca,AC,Kushalappa and T.Mosquera. "Hydroxycinnamic acids in cooked potato tubers from Solanum tuberosum group Phureja." *Food science & nutrition* 5.3 (2017): 380-389.

[29] R.Agrawal,JE.Gehrke,D.Gunopulos and P.Raghavan. "Automatic subspace clustering of high dimensional data for data mining applications." U.S. Patent No. 6,003,029. 14 Dec. 1999.

[30] S.Saini, and P. Rani. "A Survey on STING and CLIQUE Grid Based Clustering Methods." *International Journal of Advanced Research in Computer Science* 8.5 (2017).

[31] S.Ahmad, T.Taskaya-Temizel, and K. Ahmad. "Summarizing Time Series: Learning Patterns in 'Volatile'Series." *International Conference on Intelligent Data Engineering and Automated Learning*. Springer, Berlin, Heidelberg, 2004.

LIST OF ABBREVIATIONS

ABBREVIATIONS	DETAILS
SOHAC	Storage Optimization Hierarchical Agglomerative Clustering
BIRCH	Balanced iterative reducing and clustering using hierarchies
CURE	Clustering Using Representatives
ECS	Elastic Container Service
ROCK	Robust Clustering using links
PAM	Partitioning around Medoids
CLARA	Clustering Large Applications
GA	Genetic Algorithm
GKA	Genetic k-Means
GKMODE	Genetic k-MODEs
ANNs	Artificial Neural Networks
DBSCAN	Density Based Spatial Clustering of Applications with Noise
DENCLUE	Density Based Clustering
OPTICS	Ordering Points To Identify Clustering Structure
STING	Statistical Information Grid-based methods
CLIQUE	Clustering In Quest
FC	Fractal Clustering

ORIGINALITY REPORT

8%	5%	7%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	cs.bme.hu Internet Source	2%
2	Lecture Notes in Computer Science, 2012. Publication	2%
3	www.citeulike.org Internet Source	1%
4	repositorium.sdum.uminho.pt Internet Source	1%
5	Bo Han, Zhenyu Yang. "Data matrix compression by using co-clustering", 2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 2011 Publication	1%
	T. V. Sai Krishna, A. Yesu Babu, R. Kiran Kumar. "Chapter 26 Determination of Optimal Clusters for a Non-hierarchical Clustering Paradigm K-Means Algorithm", Springer Nature, 2018 Publication	<1%

7	Krisztian Buza. "A Distributed Genetic Algorithm for Graph-Based Clustering", Advances in Intelligent and Soft Computing, 2011 Publication	<1%
8	linknovate.com Internet Source	<1%
9	www.randomiq.com Internet Source	<1%
10	Piñeros-Niño, Clara, Carlos-Eduardo Narváez-Cuenca, Ajjamada C. Kushalappa, and Teresa Mosquera. "Hydroxycinnamic acids in cooked potato tubers from Solanum tuberosum group Phureja", Food Science & Nutrition, 2016. Publication	<1%
11	Krisztian Buza, Gábor I. Nagy, Alexandros Nanopoulos. "Storage-optimizing clustering algorithms for high-dimensional tick data", Expert Systems with Applications, 2014 Publication	<1%
12	Shai Ben-David. "A Sober Look at Clustering Stability", Lecture Notes in Computer Science, 2006 Publication	<1%
13	Lecture Notes in Computer Science, 2006. Publication	<1%

14	"Classical Fuzzy Cluster Analysis", Cluster Analysis for Data Mining and System Identification, 2007 Publication	<1%
15	Lecture Notes in Computer Science, 2011. Publication	<1%
16	Kanungo, Tapas Mount, David M. Netanyahu. "An efficient [kappa]-means clustering algorithm: analysis and implementation. (Abstract)", IEEE Transactions on Pattern Analysis an, July 2002 Issue Publication	<1%

Exclude quotes	On	Exclude matches	< 5 words
Exclude bibliography	On		