ISSN XXXX XXXX © 2018 IJESC

Research Article Volume 8 Issue No.9

OSC: Optimizing Storage Using Clustering Technique for Tick data

Sabha[1], V. P. Singh[2] , Vinay Gautam[3]
Research Scholar[1], Associate Professor[2], Lecturer[3]
Department of Computer Science & Engineering
Thapar Institute of Engineering & Technology, Patiala, Punjab, India

Abstract:
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. A new algorithm based on clustering technique to optimizing the storage space of the tick data applications has been proposed in this paper. The column wise partitioned of the tick dataset is considered to optimize the storage space. Then we construct binary indicator vector that contains binary information generated after matching two concurrent columns/rows and eliminate all the duplicate values. In the proposed optimizing storage using clustering Technique, compares the compression ratios of the different datasets. The results of the proposed algorithm for stock market dataset are 67% and for weather forecast dataset 31%. Extensive analysis shows that the proposed technique outperforms the existing techniques.

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

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 paper describes different clustering techniques.
This research work is based on specific approach, namely OSC: Optimizing Storage Using Clustering Technique. OSC is a hierarchical clustering approach falls under probability based clustering and centre-based clustering approach [1]. 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 paper.

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 as in figure 1. Also, a proximity measure is defined that is used in the next step. It may occur that values are missing from the data set. 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.

If 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.

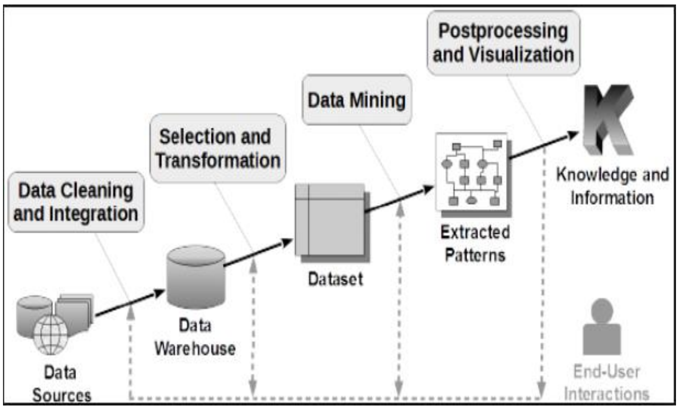

Postprocessing and Visualization
Data Mining
Selection and Transformation
Data Cleaning and Integration
Knowledge and Information
Extracted Patterns
Dataset
Data Warehouse
Data Sources
End-User Interactions

Figure 1: Clustering Process

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.

CLUSTERING TECHNIQUES

Clustering is a crucial part of internet data mining techniques which happens to be widely include with diverse areas. The classification of the clustering techniques are shown in Figure 2. Clustering Techniques 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.

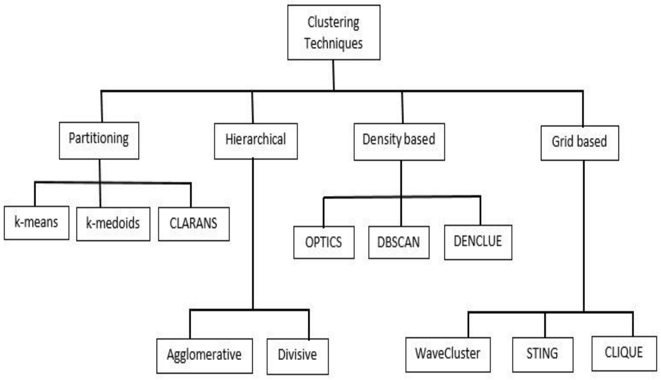


Figure 2: Clustering Techniques

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. There is no straightforward or canonical way to do this. Such groups can also overlap.

Common subdivisions include [6]:

- **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..
- **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].

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. The graphical representation of hierarchical clustering is a tree structured graph named dendrogram is shown as Figure 3.

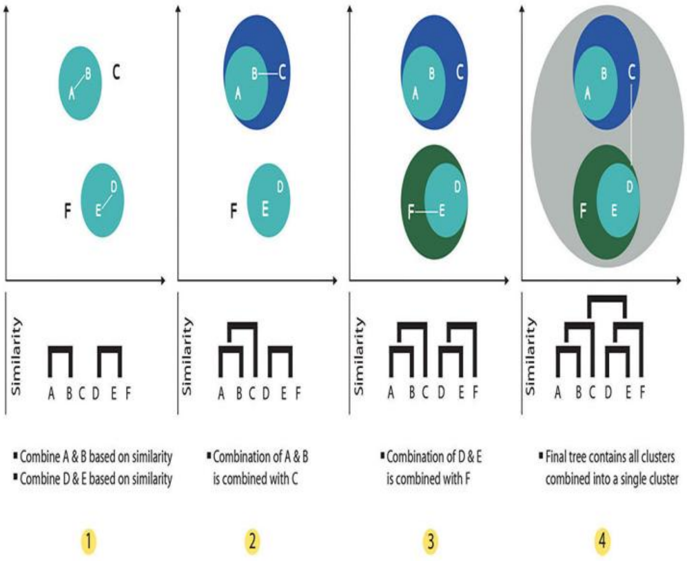

A
B
C
F
E
D
Similarity
A B C D E F
▪Combine A & B based on similarity
▪Combine D & E based on similarity
▪Combination of A & B is combined with C
▪Combination of D & E is combined with F
▪Final tree contains all clusters combined into a single cluster
1
2
3
4

Figure 3: The basic illustration of hierarchical clustering

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].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].

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.

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

- CLARA (Clustering Large Applications)[14] 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. 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.

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 [15].

Search-based Clustering

A well-known search technique is simulated annealing. This technique has been used to solve clustering problems[16].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.

Evolution-based Clustering

Evolution-based clustering approaches are inspired by natural evolution [17]. GAs has been applied for clustering the most out of the evolution-based clustering approaches. Generally, GAs consists of the following: problem encoding, initialization. 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 initialization 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.

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 [18]. 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. It can be used information for maximum likelihood estimation.

Density-based Clustering

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

General Density-based Clustering

Clusters grow in any direction, based on density alone [20]. Outliers also do not disturb density-based algorithms. 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.

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 attribute space.

- **DBSCAN** finds clusters of arbitrary shape and able to find clusters in the high dimensional spatial database [21]. 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).
- **OPTICS** [22] 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 $o(n\log n)$ in the case of indexing structure
- **DENCLUE(DENsity**
- **Based CLUestEring)**[23]: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 partitioning, 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.

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. 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 parameters. 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 [24]. They stores statistical information about data in nodes of trees where nodes represent grid cells. 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.
- **WaveCluster**-WaveCluster is clustering approach which uses different strategy and uses wavelets transforms and multi-resolution method [25]. It uses multi-resolution approach of wavelets to find erratic shaped clusters at different levels of resolution. A wavelet transforms is a signal processing method that uses various frequency bands. This method helps to find clusters of data points at different level of detail.
- CLIQUE [26, 27] is a grid-based and density-based algorithm developed to cluster high-dimensional data.

LITERATURE REVIEW

This paper covers study of different clustering techniques which are used to formulate problem. Literature review is explained below:

Storage of tick data is a very important task. Now a days huge amount of data is sending resulting to big data so there is a need to optimize the storage. Akram et al. factual studied the law of one price on different stock market [25].while Ahmad et al.focussed on summarizing the tick data time series [29]. In the past decades S Guha et al. developed a hierarchal algorithm that is ROCK which employs a links and distances when the clusters are merged [11]. ZHuang also proposed the k means algorithm. This algorithm uses clustering dataset for the categorical values. It proposed two algorithms that extend the k means algorithm for categorical domain and other one with mixed and numeric domain[4]. R.Xu et al. proposed the best clustering algorithm [19].

Now move on towards data compression technique, B Han et al. used conventional data compression techniques. As they perform data mining based compression techniques by reordering or grouping the data matrix and by post processing the redundancy in data matrix[8] .But this technique is not applicable for the execution of large datasets.so we propose the technique for optimizing the storage using clustering technique for tick data.

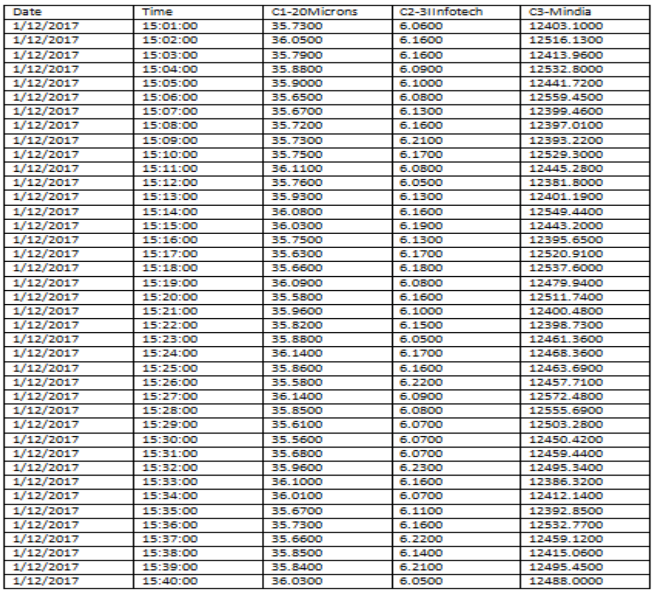
The previous work focused on the storage of tick data only by applying binary matrix [1]. But In our research work, we design and implement optimizing storage using clustering technique for tick data .we also compute the execution time of the processor. Propose technique starts with k partitions of tick 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 analyzed. Extensive analysis shows that the proposed technique outperforms existing techniques.

Proposed Approach

The propose storage optimizing clustering technique which is regarded as a clustering line of a tick data vector. This 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 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 [28]. The variation of tick data in the storage size has also been analyzed.

Table 1: Clustering after Partition

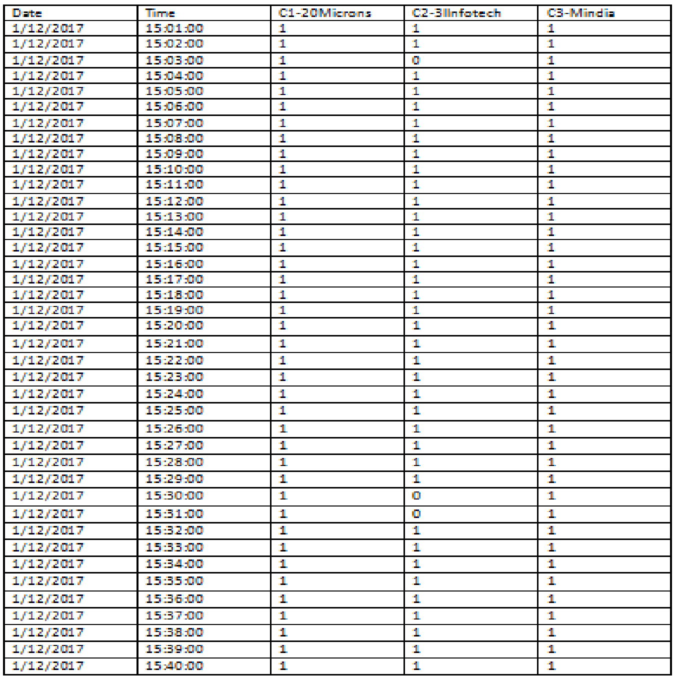


Date	Time	C1-20Microns	C2-3IInfotech	C3-Mindia
1/12/2017	15:01:00	35.7300	6.0600	12403.1000
1/12/2017	15:02:00	36.0500	6.1600	12516.1300
1/12/2017	15:03:00	35.7900	6.1600	12413.9600
1/12/2017	15:04:00	35.8800	6.0900	12532.8000
1/12/2017	15:05:00	35.9000	6.1000	12441.7200
1/12/2017	15:06:00	35.6500	6.0800	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 the 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 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 case.

Table 2: Binary indicator vector



Date	Time	C1-20Microns	C2-3IInfotech	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, 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*.

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

- Load the data parameters from object and event.
- Get Weather Data from Repository
- Load the Clusters from the data parameters.
- Divide the data into number of partitions say k.
- Cluster the columns c1,c2….cn
- **while**$|c| > k$**do**
- $s \leftarrow \infty$ (Compute the storage size)
- **for** all pairs of clusters $(C_i, C_j), with\ C_i \in P;\ C_j \in P$**do**
- Merge clusters $C_i and\ C_j$ into new cluster C_i (that is a unique data of date and time)
- $c' \leftarrow c \setminus \{C_i\} \setminus \{C_j\} \cup \{C_i\}$(select the clusters one by one or select all clusters)
- Create the binary indicator vector *I* from *M*
- S1= storage size required to store the decomposition
- **if**$s1 < s2$ **then**
- $c^* \leftarrow c'$ (The best clustering found so far)
- $s \leftarrow s1$
- **end if**
- Compute Storage Size of Data Per Column.
- Reduce Repeating Values to Optimize Storage Space

Experimentation 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 is one millionth (10^{-6}) of a second. Two datasets are considered: one is stock market and other is weather forecasting. The appraisal of propose technique is done on the following parameters such as Total data volume, required space, optimized data, compression ratio based on different parameters.

Results Analysis

The comparison of compression ratio for existing and propose approach for different datasets by using two partitions as shown in Table 3 and Figure 4.

Table 3: Compression Ratio of existing and propose results at k=2

Datasets / No of Partitions	Weather Forecast		Stock Market	
	SOHACC	OSC	SOHACC	OSC
K=2 (Compression Ratio)	1.31787	1.93199	1.08388	1.4484

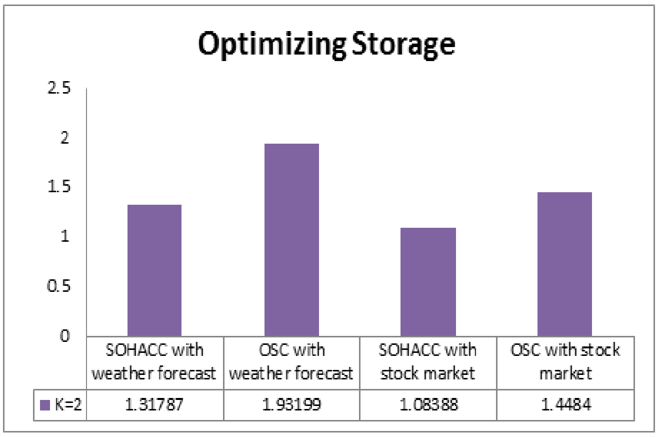


Figure 4: Storage optimization for tick data at k=2

The comparison of compression ratio for existing and propose approach for different datasets by using three partitions as shown in Table 4 and Figure 5.

Table 4: Compression Ratio of existing and propose results at k=3

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

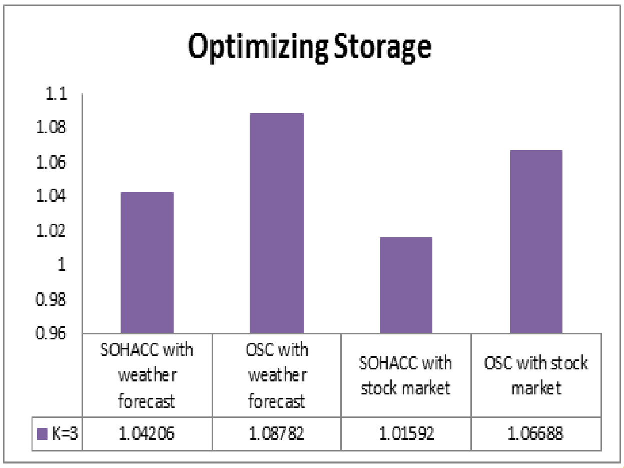


Figure 5: Storage optimization for tick data at k=3

The comparison of compression ratio for existing and propose approach for different datasets by using four partitions as shown in Table 5 and Figure 6.

Table 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 (Compression Ratio)	1.04135	1.08627	1.0007	1.0314

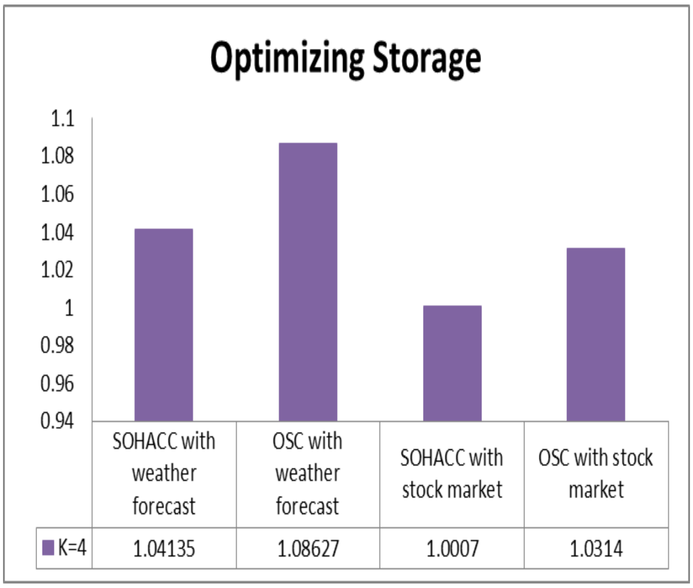


Figure 6: Storage optimization for tick data at k=4

Conclusion

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 paper, aim is to optimize the storage space using clustering technique.

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 compute the compression ratio. The variation of tick data in the storage size has also been analyzed. The percentage of OSC for stock market is 67% as compared to SOHAC which is 59%. Similarly, the percentage of OSC for weather forecast is 31% as compared to SOHAC which is 21% . So from the results we conclude that Extensive analysis shows that the proposed technique outperforms existing technique.

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

[15]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.

[16] 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.

[17]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.

[18]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..

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

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

[21]J.SwarndeepSaket and S.Pandya."An Overview of Partitioning Algorithms in Clustering Techniques."

[22] 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.

[23]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.

[24]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.

[25]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.

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

[27] 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.

[28]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).

[29]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.