Data Mining on Moving Object Trajectories

Sonal Athavale[1] Neelabh Sao[2]

Computer Science Department
Rungta College of Engg. & Tech.,Bhilai
sonal.athavale@gmail.com
neelabhsao@gmail.com

Abstract: - Trajectory means a path followed by a moving vehicles or object. Spatio temporal clustering is a process of grouping objects based on their spatial or temporal similarity. It is also known as Trajectory or mobility data. Existing work have mainly focused on moving object is completely based on clustering algorithms which only gives us that at any particular time how many vehicles are present at any particular location, but it does not generate any useful patterns.

In this project we deal with the analysis, pre processing and modeling of traffic data in moving object database for traffic management system. And we are using clustering techniques after that we are classifying our data set to generate useful patterns using C 4.5 Algorithm.

In clustering we are using Incremental DBSCAN algorithms because as we know moving data always gets updated these data are dynamic in nature.

Keywords*:* Trajectory, Spatio temporal clustering, C 4.5 Algorithm, Incremental DBSCAN, Moving Object.

I. Introduction

The comprehension of phenomena related to movement not only of people and vehicles but also of animals and other moving objects – has always been a key issue in many areas of scientific investigation or social analysis. Many applications track the movement of mobile objects, using location- acquisition technologies such as Global Positioning System (GPS), Global System for Mobile Communications (GSM) etc., and it can be represented as sequences of time stamped locations. [1][7]

TD consists of movements of objects, which record their position as it evolves over time, the concept of uncertainty appears in various ways; data imprecision due to sampling and/or measurement errors, uncertainty in querying and answering, fuzziness by purpose during pre-processing for preserving anonymity, and so on. Although uncertainty is inherent in TD, to the best of our knowledge there is no related work in the database literature that studies its effect in the knowledge discovery process.

The distance functions and indexing methods proposed for one-dimensional time series data a not be directly applied to moving object trajectories due to their unique characteristics [2].

- Trajectories are usually two or three dimensional data sequences and a trajectory data set often contain trajectories with different lengths. Most of the earlier proposals on similarity-based time series data retrieval are focused on one-dimensional time series data [2].
- Trajectories usually have many outliers. Unlike stock, weather, or commodity price data, trajectories of moving objects are captured by recording the positions of the objects from time to time (or tracing moving objects from frame-to-frame in videos). Thus, due to sensor failures, disturbance signals or errors in detection techniques, many outliers may appear. Longest Common Subsequences (LCSS) has been applied to address this problem; however, it does not consider various gaps between similar subsequences, which lead to inaccuracy. The gap refers to a sub-trajectory in between two identified similar components of two trajectories.
- Similar movement patterns may appear in different regions of trajectories. Different sampling rates of tracking and recording devices combined with different speeds of the moving objects may introduce local shifts into trajectories (i.e., the trajectories follow similar paths, but certain sub-paths are shifted in time)[1][7].

II. Incremental Data Mining

Incremental' means changes in the existing database i.e. insertion of new data into the database or deletion of old data from the database. This is sometimes called '% of delta change in the database'. This is a very important issue now a day. Because at present most of the databases are dynamic in nature. So, we need to develop some new data mining techniques (algorithms) which can handle this dynamic feature of the database efficiently and effectively. The objective of incremental data mining algorithms is to minimize the scanning and calculation effort for newly added records. Here we improve the efficiency of newly added record updating problem. These are the few prime factors that cause to apply the incremental Mining[5][6].

- Database always gets modified.
- During each modification (insertion and deletion) the database requires scanning again. Thus it requires lot of time for rescanning.
- Hence, the Incremental version of clustering algorithm includes the logic for Insertion and Deletion as separate Dynamic operation.
- Whenever, the databases may have frequent updates and thus it may be dynamic.
- After insertions and deletions to the database, the existing clustering algorithm has to be updated.
- After that it can be expected algorithm of incremental approach performs efficiently compare to actual one [9].

III. DBSCAN (DENSITY BASED SPATIAL CLUSTERING OF APPLICATIONS WITH NOISE)

In this section, DBSCAN (density based spatial clustering of applications with noise) is a density based approach to cluster data of arbitrary shape. DBSCAN is based on two main concepts: density reachability and density connectability. These both concepts depend on two input parameters of the dbscan clustering: the size of epsilon neighborhood e and the minimum points in a cluster m. DBSCAN requires only one input parameter and supports the user in determining an appropriate value for it[5]. We performed an experimental evaluation of the effectiveness and efficiency of DBSCAN using synthetic data and real data. DBSCAN is efficient even for large spatial database[3].

DBSCAN uses the same values for all clusters. The density parameters of the “thinnest” cluster are good candidates for these global parameter values specifying the lowest density which is not considered to be noise[4].

A. DBSCAN Clustering

DBSCAN clustering One of the most common clustering algorithms and also most cited in scientific literature is Density Based Spatial Clustering of Applications with Noise (DBSCAN) which has the ability to produce arbitrary shape of clusters. Clusters are identified by looking at the density of points. Regions with a high density of points depict the existence of clusters whereas regions with a low density of points indicate clusters of noise or clusters of outliers. DBSCAN grows clusters according to a density based connectivity analysis. It defines a cluster as a maximal set of density-connected points. The key idea of density-based clustering is that for each object of a cluster the neighborhood of a given radius ( ) has to contain at least a minimum number of objects (MinPts), i.e. the cardinality of the neighborhood has to exceed some threshold[10].

DBSCAN requires two parameters: ε (eps) and the minimum number of points required to form a cluster (Minpts). It starts with an arbitrary starting point that has not been visited. This point's ε-neighborhood is retrieved, and if it contains sufficiently many points, a cluster is started. Otherwise, the point is labeled as noise [10].

IV. PROBLEM IDENTIFICATION

As the number of moving vehicles increase rapidly everyday the need for analysis, modeling and preprocessing of traffic data is vital.

When we think of moving object or trajectory data that represent traffic situated in some city or province, obvious task we would like to perform concerning everyday phenomena include detecting traffic jams, predicting traffic jams and discovering relations between traffic jams. We will also try to reduce the chances of occurring accidents [7][8].

Typical for moving object is that they have speed and clustering can be directed to detect similarly fast moving objects. Other problems are-

- For planning traffic and public mobility systems in metropolitan areas.
- For forecasting traffic related phenomena.
- For planning physical communication networks, such as new roads or railways.
- For reducing the chances of occurring accidents..

V. SOLUTION APPROACH

In this paper described, we deal with the analysis, pre processing and modeling of traffic data in moving object database for traffic management system.

- Firstly, we are clustering our Data Set by using Incremental DBscan clustering algorithm by which we will find that at any particular time how many vehicles are present at any location.
- After that we are classifying our data set to generate useful patterns using C 4.5 Algorithm.

A. C 4.5- Classification Algorithm

C4.5 is a well known Machine Learning algorithm, which automates decision tree generation. Various extensions have been made to the method to improve the efficiency, effectiveness, and generality to the domain of the trees produced. However this method has a number of disadvantages: the most significant drawback of C4.5 and similar methods is that they can not incrementally learn knowledge. The tree learned is derived from the set of data presented, and can not be easily modified. Also, the method can not incorporate previous domain knowledge easily, unless that knowledge can be represented using additional attributes for each case in the dataset.

- C 4.5 successor of ID3 is an algorithm used to generate decision tree often referred to as a statistical classifier.
- C 4.5 builds decision tree from a set of training data in the same way as ID3.
- At each node of the tree, C 4.5 chooses one attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other.

VI. CONCLUSION

Our work focuses to manage and maintain traffic systems or for diverting traffic under certain emergency circumstances or for tracking peoples movement in mall to get popularity of mall.

For this firstly we are clustering our data set and after that we are going to perform classification on that, because by doing only clustering we can not get any useful patterns only we can get is at any particular time how many vehicles are present at any location and their directions but by this it is not possible to get useful patterns only we can do is to compare the performance of various clustering algorithms.

So after clustering we are going to classify our data set, which will help us to generate useful patterns. After this we will compare the result of clustering and classification and try to generate useful patterns, and check the accuracy of both.

REFERENCES

[1] Ajaya Kumar Akasapu, Lokesh Kumar Sharma, G. Ramakrishna in Volume 9– No.5, November 2010, “Efficient Trajectory Pattern Mining for both Sparse and Dense Dataset”, International Journal of Computer Applications (0975 – 8887).

[2] R. Agrawal, C. Faloutsos, and A. N. Swami. Efficient similarity search in sequence databases. In Proc. Conf. of Foundations of Data Organization and Algorithms, 1993.

[3] Carlos Ordonez, “Clustering Binary Data Streams with DBSCAN”, San Diego, CA, USA. Copyright 2003, ACM 1- 58113-763-x, DMKD'03, June 13, 2003.

[4] Martin Ester, Hans-Peter Kriegel, Jiirg Sander, Xiaowei Xu, in 1996, “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”.

[5] Data Mining concepts and techniques by Jiawei Han and Micheline Kamber, Morgan Kaufmann (publisher) from chapter-7 ‘cluster analysis’, ISBN:978-1-55860-901-3, 2006.

[6] Dunham, M.H., Data Mining: Introductory And Advanced Topics, New Jersey: Prentice Hall, ISBN-13: 9780130888921. 2003.

[7] Applications Volume-9 –No. 5. Jae-Gil Lee, Jiawei Han & Kyu-Young Whang” Trajectory Clustering: A Partition-and-Group Framework in University of Illinois at Urbana-Champaign KAIST

[8] Naresh kumar Nagwani and Ashok Bhansali, “An Object Oriented Email Clustering Model Using Weighted Similarities between Emails Attributes”, International Journal of Research and Reviews in Computer science (IJRRCS), Vol. 1, No. 2, June 2010.

[9] I. H.Witten, Data mining: practical machine learning tools and techniques with Java implementationsSan-Francisco, California : Morgan Kaufmann, ISBN: 978-0-12-374856-0 2000.

[10] SANJAY CHAKRABORTY Prof. N.K.NAGWANI, in 2011,” Analysis and Study of Incremental DBSCAN Clustering Algorithm”, International Journal of Enterprise Computing and Business Systems, ISSN (Online) : 2230 230--8849 –8849, Vol. 1 Issue 2 July 2011.