Self Organizing Map based Clustering Approach for Trajectory Data

Sanjiv Kumar Shukla[#1], Sourabh Rungta[#2], Lokesh Kumar Sharma[#3]

[#]*Department of Computer Science and Engineering, Rungta College of Engineering and Technology Bhilai (CG) - India*

***Abstract*— Clustering algorithm for the moving or trajectory data provides new and helpful information. It has wide application on various location aware services. In this study the Self Organizing Map is used to form the cluster on trajectory data. The self-organizing map (SOM) is an important tool in exploratory phase of data mining. It projects input space on prototypes of a low-dimensional regular grid that can be effectively utilized to visualize and explore properties of the data. When the number of SOM units is large, to facilitate quantitative analysis of the map and the data, similar units need to be grouped, i.e., clustered.**

***Keywords*— Trajectory Data, Self-Organizing Map, Clustering.**

I. INTRODUCTION

Clustering is a technique in descriptive modelling or unsupervised learning field of machine learning. It is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters). Clustering is useful in several exploratory pattern-analysis, grouping, decision-making, and machine-learning situations, including data mining, document retrieval, image segmentation, and pattern classification. It is the process of producing unlabeled categorized data [1]. The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. But how to decide what constitutes a good clustering? It can be shown that there is no absolute "best" criterion, which would be independent of the final aim of the clustering. Consequently, it is the user, which must supply this criterion, in such a way that the result of the clustering will suit their needs. For instance, we could be interested in finding representatives for homogeneous groups (data reduction), in finding "natural clusters" and describe their unknown properties ("natural" data types), in finding useful and suitable groupings ("useful" data classes) or in finding unusual data objects (outlier detection) [1].

The increasing use of GPS devices and other location-aware devices to capture the position of moving objects demands tools for the efficient analysis of large amounts of data referenced in space and time [3] [4] [5]. In database community, trajectories, which usually stored in moving objects databases (sometimes called trajectory databases in the literature), describe complete histories of movement. Spaccapietra et al. [14] has given the definition of trajectories semantic meaning as:

" A trajectory is the user defined record of the evolution of the position (perceived as a point) of an object that is moving in space during a given time interval in order to achieve a given goal. Trajectory: [t begin, t end] → space. "

Clustering on trajectory data is a very important data mining task for a wide variety of application fields including location aware services, geo-marketing protein analysis etc. Most of traffic planner or Geo-marketer takes interest to know the most visited place or important place with respect to product promotion. It is very useful in various applications. Trajectory data clustering provides good chance to identify the visited place and also find the similar interested place [2]. In this study, we used a self-organizing map (SOM) to identify clusters in a trajectory data. SOM is an unsupervised learning method that relates similar input vectors to the same region of a map of neurons [9]. SOMs have been used for other tasks in trajectory data such as a benchmark for model selection and to predict class. Once the SOM was used to identify the clusters, a constraint-satisfaction neural network (CSNN) was used to characterize the clusters by determining a profile for each cluster. Briefly, the CSNN is a Hopfield-type network of neurons arranged in a non-hierarchical way. There are symmetric, bi-directional weights between all pairs of neurons but there are no reflexive weights. The CSNN operates as a nonlinear, dynamic system that tries to reach a globally stable state by adjusting the activation levels of the neurons under the constraints imposed by the a priori fixed weight values. A cluster ''profile'' provides a description of a ''typical'' case in the cluster [11].

The remainder of this paper is organized as follows. In section 2, we present the related work in the area of trajectory clustering. In section 3, we will present the architecture and algorithm of self-organizing map cluster for trajectory data. In section 4 the experimental investigation is reported. Data pre-processing and result analysis is reported. Finally the work is concluded in section 5. .

II. RELATED WORKS

In cluster analysis, the goal is to partition a data set into groups of closely related data in such a way that the observations belonging to the same group, or cluster, are similar to each other, while the observations belonging to different clusters are not. Trajectory clustering, the discovery of groups of 'similar' trajectories, together with a summary of each group. Knowing which are the main routes (represented by clusters) followed by people or vehicles during the day can represent precious information for mobility analysis. For example, trajectory clusters may highlight the presence of

important routes not adequately covered by the public transportation service [2] [5].

Gaffney et al. [12] proposed trajectory clustering with mixtures of regression models. In this study they addressed the problem of clustering trajectories namely sets of short sequences of data measured as a function of a dependent variable such as time. Examples include storm path trajectories, longitudinal data such as drug therapy response, and functional expression data in computational biology, and movements of objects or individuals in video sequences. This clustering algorithm is based on a principled method for probabilistic modelling of a set of trajectories as individual sequences of points generated from a finite mixture model consisting of regression model components. Unsupervised learning is carried out using maximum likelihood principles. Specifically, the EM algorithm is used to cope with the hidden data problem (i.e., the cluster memberships). Also the generalization of the method to handle non-parametric (kernel) regression components as well as multi-dimensional outputs was developed.

Lee et al. [8] have proposed partition and group frame based trajectory-clustering technique. The advantage of this framework is to discover common sub-trajectories from a trajectory database. This algorithm consists of two phases: partitioning and grouping. The first phase presents a formal trajectory-partitioning algorithm using the Minimum Description Length (MDL) principle.

Akasapu et al. [2] augmented the relative density-based clustering algorithm for movement data or trajectory data. It utilized a k-nearest neighbours clustering algorithm based on relative density for the performing the dense cluster. This algorithm inherits the features of density-based algorithm and also it produces the noises.

III. Self Organizing Map

The SOM algorithm is based on a competitive algorithm founded on the vector quantification principle: at each cycle of life in the network, the unit from Kohonen's layer whose codebook is most similar to the input wins. This unit is given the name of Winner Unit (WU). Consequently, the WU codebook is modified to get it even closer to the input. The codebooks belonging to the units that are physically near the WU (which are part of the neighbourhood) are also put closer to the input of a given delta. The algorithm calculates a first stage during which the parameters of neighbourhood and corrections of weights are set and the codebook initialization is carried out; this stage is followed by the cyclic stage of codebook adjustment. In this stage the codebooks are modified for the network to classify the input records [6] [13].

A. Training Algorithm

Initially the weights and learning rate are initialized. The input data to be clustered are presented to the network. Once the input vectors are given, based on the initial weights, the winner unit is calculated either by Euclidean distance method or sum of products method. Based on the winner unit selection, the weights are updated for that particular winner unit using competitive learning rule. The SOM training process can be summarized in following steps [6] [10] [13]:

Step 1: Set topological neighbourhood parameters. Set learning rate initialize weights.

Step 2: While stopping condition is false, do steps 3-9.

Step 3: For each input data x, do steps 4-6

Step 4: for each j, compute squared Euclidean distance.

$$D(j) = \sum (w_{ij} - x_i)^2 ; i = 1 \ldots n, \; j = 1 \cdots n \quad (1)$$

Step 5: Find index J, when D (j) is minimum.

Step 6: For all units J, with specified neighbourhood of j and for all i update the weights.

$$w_{ij(new)} = w_{ij(old)} + \alpha [x_i - w_{ij(old)}] \quad (2)$$

Step 7: Update the learning rate.

Step 8: Reduce the radius of topological neighbourhood at specified times.

Step 9: Test the stopping condition.

IV. Experimental Investigation

The raw data of trajectory contains vehicle id, time stamp and location (longitude and latitude). Therefore raw data is required to preprocess before applying the clustering. In this section data preprocessing is reported and further the result analysis is discussed.

A. Data Pre-processing

For the task of clustering on trajectory data, we used Milan datasets. These data contain the records of moving vehicles in Milan City, Italy, which is provided by Milan Metropolitan Authority for research purpose. Data consists of positions of the vehicles, which has been GPS-tracked between April 1, 2007 and April 7, 2007, and are stored in a relational database. The data have been recorded only while the vehicles moved. Each record includes the vehicle-id, date and time, the latitude, longitude, and altitude of the position. To facilitate analysis of movement data, initial preprocessing in the database is performed, which enriches the data with additional fields: the time of the next position in the sequence, the time interval and the distance in space to the next position, speed, direction, acceleration (change of the speed), and turn (change of the direction) [1][3][4].

The most visited location is an important in term of location aware services. To identify the interest location, location aware service provider can provide the facility and plan the other marketing strategies. This regards the trajectory are split into several trajectories with respect to time. In this study we split trajectory on where vehicle is stopped between 30-60 minutes. We assume here the vehicle user stop for short period. These locations are not home and workplace. We do not consider very short time stop such as 3-15 minutes. We assume this stop may be due to traffic signals or other. Further the ending points are identified and clustering algorithms are applied on those data points to perform the group of points or clusters. The experiment is performed in MATLAB to utilize the SOM neural network tools.

B. Result Analysis

In this experiment we use the above data set. It contains 996 records. Fig. 1 shows the topology of SOM. In this study we use 5x5 size map. The Fig. 2 shows the SOM layer hits. With each neuron showing the number of input vectors that it classifies. The relative number of vectors for each neuron is shown via the size of a colours patch. Fig. 3 shows SOM neighbour weight distances. SOM layer showing neurons as gray-blue patches and their direct neighbour relations with red lines. The neighbour patches are colour from black to yellow to show how close each neuron's weight vector is to its neighbours. Fig. 4 shows the SOM weight positions. The input vectors as green dots and shows how the SOM classifies the input space by showing blue-gray dots for each neuron's weight vector and connecting neighbouring neurons with red lines. Fig. 5 represents the data on various clusters. Each cluster is shown in different colours. A SOM layer shows neurons as gray-blue patches and their direct neighbour relations with red lines in the Fig. 6.

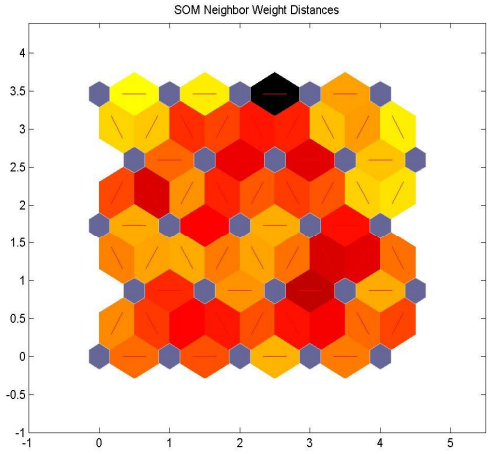

Fig. 3. SOM neighbour weight distances

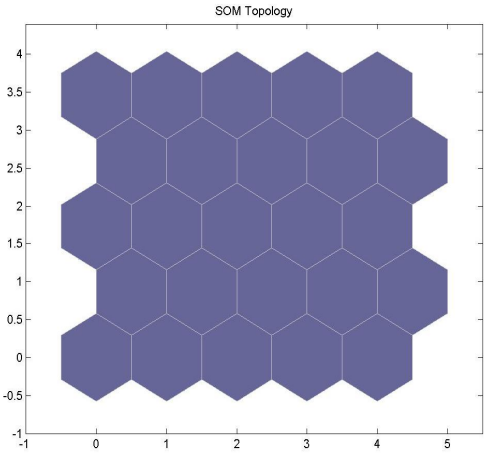

Fig. 1 Topology of SOM

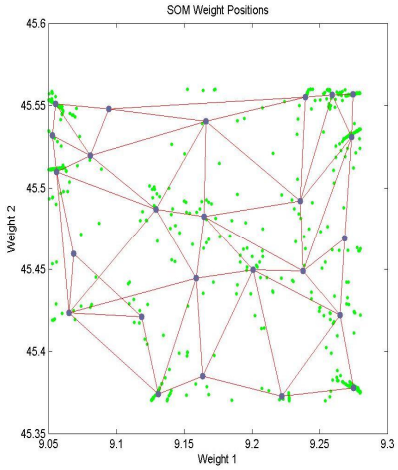

Fig. 4 Weight position of SOM

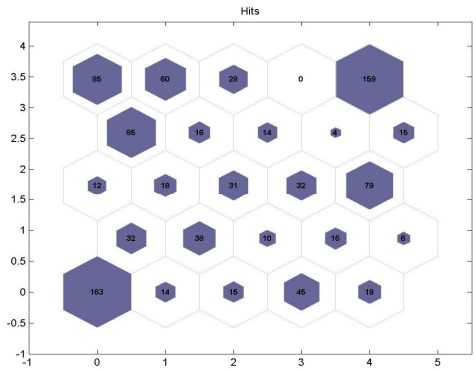

Fig. 2 SOM layer hits

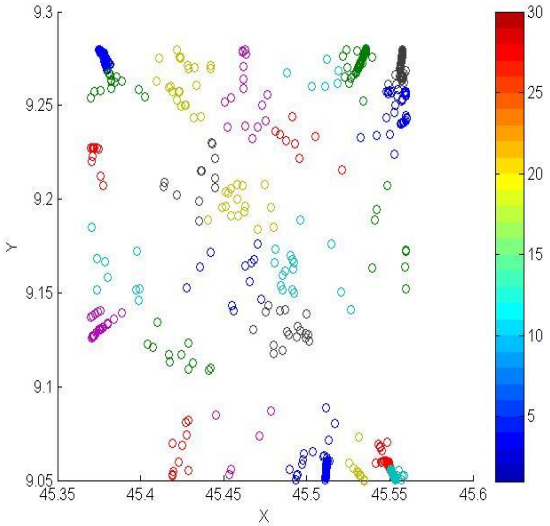

Fig. 5 Cluster data

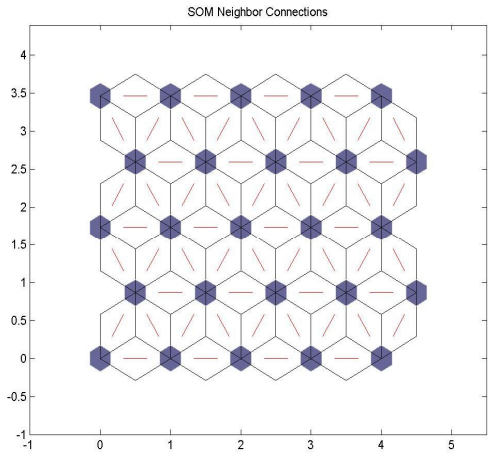

Fig. 6 SOM Neighbour Connections

V. CONCLUSION

There is clearly an increasing demand for Location Aware Service applications and while the developed retype is basic in its current stage, it is able to identify the location of an information device user, search for offers that are within a defined range and present the offers to the users, the findings of this research have provided preliminary empirical evidence about how users are willing to strike a balance between value and risk. In this study the self-organizing map based trajectory clustering framework is provided in this regards. The trajectory clustering can be used as an important tool to indentify the important location so that the services can be applied.

REFERENCES

[1] A. K. Jain, M. N. Murty, P. J .Flynn, " Data Clustering: A Review." ACM Computing Surveys, Vol. 31, No. 3, pp. 265-323, Sep. 1999.

[2] A. Akasapu et al. "Density Based k-Nearest Neighbors Clustering Algorithm for Trajectory Data", Int. J. on Advanced Science and Technology, Vol. 31, June 2011, pp. 47-57, 2011.

[3] F. Giannotti and D. Pedreschi, "Mobility, Data Mining and Privacy: Geographic Knowledge Discovery", Springer Verlag, 2008.

[4] F. Ginnotti, M. Nanni, D. Pedreschi and F. Pinelli, "Trajectory Pattern Mining", In Proceedings of the 13th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 330 – 339, 2007.

[5] G. Andrienko, N. Andrienko, and S. Wrobel, "Visual Analytics Tools for Analysis of Movement Data", ACM SIGKDD: 38-46, ISSN:1931-0145, 2007.

[6] G. Massini, "Applications of Mathematics in Models, Artificial Neural Networks and Arts", Chapter 13, Springer, 2010.

[7] J. Gudmundsson, P. Laube and T. Wolle T. "Movement Patterns in Spatio-Temporal Data", In: Shekhar, S. & Xiong, H. (eds.). Encyclopedia of GIS, Springer-Verlag, 2008.

[8] J. Lee, J. Han., and K. Whang ,"Trajectory clustering: a partition-and-group framework", In Proceedings ACM SIGMOD Int. Conf. on Management of Data: 593 – 604, 2007.

[9] J. Vesanto and E. Alhoniemi,"Clustering of the Self-Organizing Map",IEEE Tran. on Neural Networks, Vol. 11, No. 3, pp. 586-600, 2000

[10] M. H. Beale et al., "User Guide: MATLAB Neural Network Toolbox", The Math Works, Inc., 2012.

[11] M. K. Markey et al., "Self-organizing map for cluster analysis of a breast cancer database", Artificial Intelligence in Medicine 27, pp. 113–127, 2003.

[12] S. Gaffney and P. Smyth, "Trajectory Clustering with Mixtures of Regression Models",KDD 99 San Diego CA USA, pp. 63-72

[13] S. N. Sivanandam et al., "Introduction to Neural Network using MATLAB 6.0", Tata McGraw-Hill Publishing Company 2006.

[14] S. Spaccapietra, et al. "A conceptual view on trajectories", Data & Knowledge Engineering 65: 126-146, 2008..