Information Sciences 369 (2016) 388–401

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Development and validation of OPTICS based spatio-temporal clustering technique

K.P. Agrawal[a,*], Sanjay Garg[a], Shashikant Sharma[b], Pinkal Patel[a]

[a] Computer Science and Engineering Department, Institute of Technology, Nirma University, Ahmedabad 382481, Gujarat, India
[b] Space Application Center (ISRO), Ahmedabad, Gujarat, India

ARTICLE INFO

Article history:
Received 3 December 2015
Revised 21 June 2016
Accepted 26 June 2016
Available online 30 June 2016

JEL Classification:
C: Mathematical and Quantitative Methods

Keywords:
Spatio-temporal
Clustering
ST-OPTICS
ST-DBSCAN
Cluster validation indices

ABSTRACT

Spatio-temporal data mining (STDM) is a process of the extraction of implicit knowledge, spatial and temporal relationships, or other patterns not explicitly stored in spatio-temporal databases. As data are growing not only from static view point, but they also evolve spatially and temporally which is dynamic in nature that is the reason why this field is now becoming very important field of research. In addition, spatio-temporal (ST) data tend to be highly auto-correlated, which leads to failure of assumption of independence, which is there in Gaussian distribution model. Vital issues in spatio temporal clustering technique for earth observation data is to obtain clusters of, good quality, arbitrary shape, problem of nested clustering and their validation. The present paper addresses these issues and proposes their solutions. In this direction, an attempt has been made to develop a clustering algorithm named as "Spatio-Temporal - Ordering Points to Identify Clustering Structure (ST-OPTICS)" which is modified version of existing density based technique – "Ordering Points to Identify Clustering Structure (OPTICS)". Experimental work carried out is analysed and found that quality of clusters obtained and run time efficiency are much better than existing technique i.e. ST-DBSCAN. In order to improve the visualization and the interpretation of obtained micro level clusters, sincere effort has been put in to merge the obtained clusters using agglomerative approach. Performance evaluation is done in both ways i.e. qualitatively and quantitatively for cross validating the results. Results show performance improvement of proposed ST-OPTICS clustering technique compared to ST-DBSCAN algorithm.

1. Introduction

Owing to the generation of petabytes of data (may be classical, spatial, temporal or hybrid type) on daily basis from different sources, data mining tasks are required to extract hidden knowledge from it, so these voluminous data can be utilized meaningfully. Clustering is more challenging task compared to classification due to absence of class labels in dataset. In real life, there are many instances where class labels are absent especially when obtained data are from satellites. All geographic phenomena evolve over time both spatially and temporally, so there is a need to perform tasks related to clustering on spatial, temporal and spatio-temporal data, which may be useful in many ways viz. land utilization, urban and rural planning, identification of major crop areas, identification of earth quake prone areas etc.

* Corresponding author.
E-mail addresses: kpa229@gmail.com (K.P. Agrawal), sgarg@nirmauni.ac.in, gargsv@gmail.com (S. Garg), sasharma@sac.isro.gov.in (S. Sharma), pinkal08cece@gmail.com (P. Patel).

http://dx.doi.org/10.1016/j.ins.2016.06.048
0020-0255/

1.1. Clustering technique(s)

A cluster formation is nothing but partitioning dataset in some group(s) in which each object(s) in same group or can also be referred as cluster will have similar property while objects belonging to different groups/clusters will be dissimilar in nature. Clustering performed using classical datasets which are often discrete, obey Gaussian distribution model, and focus is laid on global pattern. On the other hand, in case of spatio-temporal data mining, data samples are highly auto-correlated and are embedded in continuous space instead of discrete space and at the same time focus is on the local pattern instead of global unlike in classical datasets.

Spatio-temporal data [25] can be explored in term of spatial dimension, temporal dimension and non-spatial measures. Spatial dimension relates to the location of fixed object on the earth, for example latitude and longitude are describing the location of object. Temporal dimension is described about time period attached to the data which expresses when it was valid or stored in the database. Non-spatial measures (also called attribute or characteristic data) are used to characterize non-spatial features of objects (such as name, population, and unemployment rate for a city, normalized difference vegetation index, temperature, rainfall etc).

Major requirements of spatio temporal clustering technique are as follows:

- Discovery of clusters with arbitrary shape
- Ability to handle high dimensional data
- Ability to deal with spatial, non-spatial and temporal attributes
- Independent of input data order
- Good Interpretability and usability
- Ability to deal with nested and adjacent clusters
- Should be Scalable

Proposed technique has taken care of the these requirements.

1.2. Validation indices

In real life, most of clustering techniques are very sensitive to their input parameters and results obtained from them are also different and there are no predefined structures of clusters and hence it is difficult to identify:

(i) Appropriate clustering techniques/algorithms for available dataset(s).
(ii) Appropriate clustering structure for different parameters of same clustering algorithm.
(iii) The accurate number of clusters in a dataset.

For performance evaluation, different cluster validation indices have been proposed [18,40] which belong to three categories viz. external, internal and relative indices.

Cluster validation indices have major issues in the following fields when they are used for validation purpose :

- Detection of dense and arbitrary shape clusters: As most of existing validation indices usually choose a fixed representative point from each cluster, calculate object distance and other parameters based on this point [27].
- Obtaining clusters from spatio-temporal dataset(s).

1.3. Contribution of this work

Present paper focuses on development of spatio-temporal clustering algorithm which gives dense and arbitrary shaped clusters in nature, able to overcome difficulty of presence of nested clusters and their performance evaluation using qualitative and quantitative approaches.

Remaining part of the paper is organized as follows: Section 2 deals with related work. Theoretically selected validation indices with reasoning to validate cluster's quality have been given in Section 3. In Section 4, proposed ST-OPTICS technique and its validation have been described in detail along with results, discussions, comparison with existing technique, run time performance and scalability. Paper ended with conclusion and future enhancement followed by references.

2. Related work

This section summarizes the work related to clustering carried out by various researchers in the field. Basic principle behind clustering [21] is the process of putting the objects together in groups based upon certain similarity measures with the aim of minimizing intra-cluster distance among the objects and maximizing inter-cluster distance. Clustering techniques have been categorized as Partitional, hierarchical, density based, grid based and model based clustering [25,26].

2.1. Clustering techniques

Partitioning based clustering: Technique possesses the characteristics of offering spherical shaped clusters, similar in size, it does not give natural clusters, but reallocation of an object from one cluster to another is taken care which improves