Homework 1: ST-DBSCAN: An algorithm for clustering spatial-temporal data

Name: Kunal Parmar

UHID: 1329834

Abstract

Clustering of spatial-temporal data is one of the popular field of knowledge discovery and data mining due to the increasing amount of spatial-temporal data being produced. This paper proposes a new algorithm ST-DBSCAN, based on DBSCAN, for density-based clustering of spatial-temporal data. DBSCAN was chosen because of its' ability to discover clusters of arbitrary shapes and its' ability to process very large databases. ST-DBSCAN improves DBSCAN in three directions, namely, i) clustering based on spatial, non-spatial and temporal attributes by introducing similarity measures for spatial and non-spatial attributes, ii) detection of clusters of different densities by assigning a density factor which is the degree of density of a cluster iii) identification of adjacent clusters by setting a threshold value for the difference between the average value of cluster and the object. The proposed algorithm is applied to a spatial-temporal data warehouse to discover regions with similar seawater characteristics. Three applications of the algorithm to the data warehouse produced regions with i) similar sea surface temperatures, ii) similar sea surface height residuals and, iii) similar wave heights. ST-DBSCAN was able to cluster objects with similar characteristics from the spatial-temporal data. Application of ST-DBSCAN could be in geographical information systems, medical imaging, and weather forecasting.

Comment [F1]: No clear motivation

Comment [F2]: For what?

Comment [F3]: Good!

Comment [F4]: !?!: What is the message here?

Comment [F5]: Mentioned but not 100% convincing: How does the work increase the state of the art? What new capabilities

Comment [F6]: Vague the author should know what the applications are.

Conclusion on next page..

Nikolaos Sarafianos – 1392654

Advanced Machine Learning – Homework 1

Professor: Dr. Eick

Th. 10.29.2015

ST-DBSCAN: An algorithm for clustering spatial-temporal data

Abstract:

Spatio-temporal clustering is a process of grouping objects based on their spatial and temporal similarity. In this paper, we present a new density-based clustering algorithm ST-DBSCAN which is based on DBSCAN. We selected DBSCAN algorithm due to (i) its ability in discovering clusters with arbitrary shape; (ii) the fact that it does not require predetermination of the number of clusters; and (iii) its ability to process large databases. The proposed algorithm improves DBSCAN without changing its runtime complexity in three ways: (i) can cluster spatial-temporal data according to its non-spatial, spatial and temporal attributes; (ii) can detect some noise points when clusters of different densities exist by assigning a density factor to each cluster; and (iii) solves the problem of completely different values in the object borders of 2 sides in a cluster. We applied ST-DBSCAN to a spatial data warehouse system we designed for the purpose of this paper and present our results in 3 data mining applications.

Conclusion:

In this paper we introduced ST-DBSCAN. A new density-based clustering algorithm based on DBSCAN which exploits its key characteristics and at the same time improves its limitations. ST-DBSCAN is capable of clustering spatial-temporal data according to its non-spatial, spatial and temporal attributes and by comparing the average value of each cluster with new coming value, it tackles the problem that appears when the values in the object borders of 2 sides of a cluster are different. We also introduced a density factor which is assigned to each cluster in order to address the problem of noise points in cases where clusters have different densities. Finally, we presented 3 data mining applications of our approach by designing a spatial-temporal data warehouse which contains geographical information about different seas and described the process of KDD in each step.

Arthur Dunbar

ST-DBSCAN: An Algorithm for Clustering Spatial-Temporal Data

Abstract:

Cluster analysis is vital for discovering hidden relationships within data and it is extremely useful in engineering and scientific applications. One popular clustering algorithm is "Distance-Based Spatial Clustering of Applications with Noise" or DBSCAN.

DBSCAN clusters data based upon a distance metric and is great for detecting arbitrarily shaped clusters, but it has difficulty with certain kinds of data. DBSCAN can not handle clusters that have varying densities, it is not geared towards temporal data nor is it useful for spatial data, and it has problems with noise in the data extending the boundaries of the cluster artificially.

In this paper, we introduce an improvement upon DBSCAN called Spatial-Temporal DBSCAN (ST-DBSCAN). ST-DBSCAN maintains DBSCAN's good features and fixes several issues. First, ST-DBSCAN is able to cluster data based upon non-spatial, spatial, and temporal values. Second, it can handle clusters with varying densities much better than DBSCAN. Finally, it is less susceptible to having problems with noise points that get too close to a cluster's edges being wrongly added to that cluster. All of this is accomplished without increasing the order of the algorithm's time complexity above DBSCAN's O(n*logn).

In addition, we developed a spatial data warehouse system that is used in conjunction with ST-DBSCAN. This spatial data warehouse system is designed to efficiently store and make use of spatial-temporal data.

?!

incorrectly

kind of long!

why artificial

Good!

by doing what.

What evidence do we have that the algorithm works/ delivers what it promises?

No paragraphs in abstract!

Assigned Paper: ST-DBSCAN: An Algorithm for Clustering Spatial-temporal data

Student: Pham Dinh Nguyen

Motivation / Area of Reseal

Abstract

This work proposes an algorithm to cluster spatial temporal data. The algorithm is ST-DBSCAN, based on the well-known density-based DBSCAN algorithm. DBSCAN strength is scalability and ability to discover cluster of arbitrary shapes. We further extend DBSCAN with three important ~~contribution:~~ directions: clustering spatial temporal data according to its non-spatial, spatial and temporal attributes; efficiently detecting noise points in the presence of different densities; stabilizing border points detection for adjacent clusters. We also present a data warehouse system, which provides storage, management, clustering analysis, and visualization for a wide range of spatial temporal data. We use that system to demonstrate our algorithm and discuss the clustering results.

Good!

quite short!

Say something more positive

Conclusion

In this study, we introduce ST-DBSCAN with three major improvements to DBSCAN and mitigate its current limitation. The main extension is the ability to cluster spatial temporal data. The second extension is the introduction of density factor, allowing detection of clusters with different densities, which is useful in analyze real data sets. The third improvement is the use of the mean value of the cluster to robustly determine border points, which is critical for adjacent clusters.

We demonstrate our algorithm with real weather sensing data set, using our own data warehouse. The results are visualized and presented in a user-friendly interface, showing some interesting finding.

As an extension over DBSCAN, our algorithm requires more preset thresholds, as shown in the pseudo code. Though we only mention one heuristic method to choose these parameters, we believe there can be others heuristics, considering the spatial temporal relation.

Our implementation has the same performance with DBSCAN, which is O(n*log(n)), which is among the best runtime in clustering methods. We further improve the performance by apply R-Tree index in our database, and were able to process large real data sets. Further improvement might be in the direction of parallelizing the warehouse system.

Ha Le - 1344084

Abstract + Conclusion

Motivation / Area of research

Abstract: In this paper, we introduce a new clustering algorithm for spatial-temporal data based on DBSCAN (Density-Based Spatial Clustering Applications with Noise) method. Our algorithm, ST-DBSCAN, improves DBSCAN method in three important directions: clustering spatial-temporal data based on its non-spatial, spatial and temporal attributes, detecting noise points using different density factor for each cluster and identifying adjacent clusters by comparing the average value of a cluster with new coming value. Moreover, in this paper, we present a spatial data warehouse system, which is designed for the purpose of storing and clustering a wide range of spatial-temporal data. Experiments are conducted to demonstrate the applicability of our algorithm to real world problems.

Short!

not clear

need to say something more specific

Conclusion: We have presented an algorithm for clustering spatial-temporal data which improves DBSCAN method and shows potential applicability in solving real world problems. The proposed algorithm consists of three main contributions: the two distance metrics for spatial values and non-spatial values, the density factor for noise detection and the average comparison for identifying adjacent clusters. The experimental results suggest that ST-DBSCAN is robust for clustering spatial-temporal data. The processing time is significantly improved by our data warehouse system. We utilizes the R-Tree indexing method to handle spatial-temporal information, and we also adopts some filters to reduce the search space for spatial data mining algorithm.

What is the contribution here!