IDENTIFY EARTHQUAKE HOT SPOTS WITH 3-DIMENSIONAL DENSITY-BASED CLUSTERING ANALYSIS

Lei Lei

School of Information Systems and Technology, Claremont Graduate University, CA 91711
lei.lei@cgu.edu

1. INTRODUCTION

Earthquakes occur and cause loss of life and property. To help disaster preparation and prevention, it is important to predict potential earthquake risks, especially in earthquake-active areas, or earthquake hot spots. In this paper, an earthquake hot spot is defined as a region of highest earthquake risk or with significant earthquake activities within a larger area of low or normal earthquake activities. Such predictions usually explore spatial information using spatial data mining. Spatial data mining examines spatial relationships such as topology, distance, and direction [7] and aims to discover patterns of interest automatically [2]. Clustering analysis is a common spatial data mining approach that focuses on finding clusters of data objects which share similarity based on certain pre-defined criteria. Since hot spots can also be regarded as clusters of spatial data points, their identification has drawn considerable research attention in spatial data mining [2, 5, 7, 9].

An overview of various types of existing clustering algorithms, including partitioning, hierarchical clustering, and density-based clustering, can be found in [4]. Density-based clustering is usually more efficient than the other two types because density clustering of neighboring data points is usually based on local conditions and thus only requires one scan of the entire database [4]. A density-based clustering algorithm DBSCAN (Density-Based Spatial Clustering of Applications with Noise) was first introduced by Easter et al. [3]. DBSCAN grouped data points into regions based on a single value density measure that used the number of points within a given neighborhood [3]. With this measure, DBSCAN showed performance improvement over the partitioning algorithm CLARANS [6]. Hinneburg and Keim [4] introduced DENCLUE (DENsity-based CLUstEring), which used the Gaussian kernel influence function as the density measure. In DENCLUE, a hill climbing procedure moved in the direction of the gradient of the influence function [5]. Compared to DBSCAN, DENCLUE showed significant efficient improvement [4]. Based on [4], Jiang et al. [5] developed a weighted Gaussian kernel influence function as the density measure and presented SCDE (Supervised Clustering Density-based Estimation). Unlike DENCLUE, SCDE used the sign of the influence function to determine the direction in which to continue hill climbing. Compared to algorithms such as SPAM and SCMRG, SCDE generated higher maximum rewards and better fitness values [5].

One of the essential challenges for existing clustering approaches is lack of consideration for clustering features of high dimensions [4]. Many of them either emphasize on the horizontal dimension or consider the vertical dimension without using density-based clustering [1, 2, 6, 7]. Therefore, methods for using the depth data meaningfully and effectively in density based clustering algorithms deserve further research attention. This paper takes a first step towards using three dimensional variables to describe data object in supervised density-based clustering, and provides a revision to the existing SCDE algorithm by considering the impact of depth on interrelated spatial data objects to identify earthquake hot spots.

2. RESEARCH STATEMENT

The research statement is formulated as follows: given a set of spatial data points representing earthquake occurrences in an area at a given time frame, develop a density-based clustering method that is able to describe the spatial characteristics of the data objects, apply the method to extract meaningful information that will indicate the regional earthquake risk, and use the information to discover the highest risk regions (i.e. hot spots) in that area. The term "region" here refers to a clustering result and is not associated with any political or administrative boundary.

SCDE describes a data object in the form of (*<location_x, location_y>*, *<variable of interest>)* where the *variable of interest* refers to the earthquake risk. The depth variable, or the distance from the epicenter to the earthquake focus, is not considered by SCDE in the location pair, but is commonly available in many spatial earthquake data sources. Thus it is intuitive to think that depth is relevant to earthquake risk. To count its impact in earthquake hot spot identification, this paper provides a revision to SCDE. In the revision, a data object is described by a location triple of <longitude, latitude, depth> as well as the variable of interest. Accordingly, the gradient of the density function of all data objects in a dataset will be recomputed with the addition of the depth variable in the corresponding hill climbing procedure.

3. EXPERIMENTAL DESIGN AND EVALUATION

The experiment is developed as follows: The selected area is Northern California and the sample data is retrieved from querying the online USGS Northern California Earthquake Catalog database (http://www.ncedc.org/maps/). Magnitude is selected as the variable of interest in this experiment based on its geological meaning and uniqueness. In the query, the minimum magnitude is set with a system default of 3.0 and the maximum magnitude is set as unlimited. The query also includes earthquakes from all depths and excludes events with no reported magnitude. The time frame in the query is set between 2002/01/01, 00:00:00 (a system default) and 2008/10/18, 00:00:00, the later of which is a week from the day that this research was first started. After the query submission, a total of 1483 earthquake records (sample dataset *S*) are retrieved from the database.

The independent variable in this experiment is the usage of the depth of a spatial data object in SCDE. Depending on whether the depth variable will be used to describe a data object in SCDE, there are two treatments: the SCDE-with-depth treatment (T1) and the SCDE-without-depth treatment (T2). The former is the revised algorithm and the latter is the original SCDE algorithm. Three dependent variables will be measured: (1).The highest rewards *(HR)* computed from the reward function in SCDE. The reward function considers both the impact of the number of clusters and the interestingness of all resulting clusters, and *HR* finds the best tradeoff between the two. The treatment with a higher *HR* is preferred; (2). The fitness value *(FV)* calculated from the fitness function in SCDE. *FV* measures the overall reward for all resulting clusters and a higher *FV* is preferred. (3).The accuracy. Since the ultimate concern of this research is to identify hot spots of earthquakes, earthquake domain expertise is used to evaluate the clustering results. To do so, the clustering results are plotted on a three-dimensional map using a Geographic Information System (GIS) map viewer tool and overlaid with the 2008 California Seismic Hazard Map retrieved from the USGS website. This map classifies geographic locations into different zones of varying quantitative risk probabilities. Denote category *A* as the total number of data objects in the sample data set *S* that fall into the high level hazardous areas in the 2008 California Seismic Hazard Map; denote category *B* as the total number of data objects from the cluster with the highest reward and fall into the high level hazardous areas in the 2008 California Seismic Hazard Map. The accuracy (AC) of the prediction can be computed as: *Accuracy = B/A*.

Two nuisance variables are selected to control the experiment: 1. the standard deviation σ of the influence function, as [5] indicated that it is the most critical variable in SCDE that determines the size of the influence region of a data object. This research will follow the same σ value (0.75) as chosen in [5]. 2. The penalty parameter β which determines the penalty associated with the number of generated clusters. This paper will follow the same choice of β in [5]: 1.01 and 1.2. 1.01 is suggested as a good value to identify very local hot spots and 1.2 as a good value for more regional hot spots [5].

The two algorithm treatments are implemented in MATLAB R2006a and most spatial operations are performed in ArcMAP 9.3. Figure 3-1 and 3-2 show an example of the preliminary clustering results with points representing category *B* in the Accuracy formula above. More results and evaluation will be presented in the full paper.

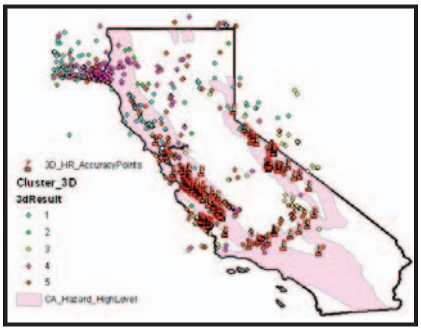

Figure 3-1 High risk points (shown as red flags) identified by the revised SCDE that fall into the high level hazardous areas (in pink) in the California Seismic Hazard Map.

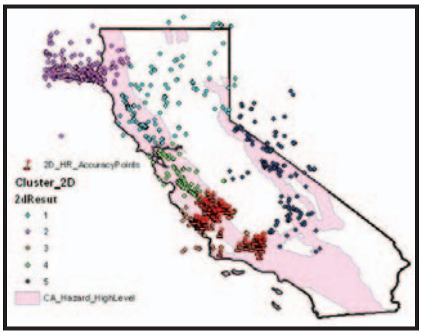

Figure 3-2 High risk points (shown as red flags) identified by SCDE that fall into the high level hazardous areas (in pink) in the California Seismic Hazard Map.

4. CONTRIBUTIONS AND CONCLUSIONS

One popular topic in spatial data mining is identifying earthquake hot spots. The vertical dimension of a data object contains relevant spatial information and has not been fully explored by existing density-based clustering algorithms. This paper provides a revision to the SCDE algorithm by introducing depth to describe data objects and develops an experimental design to examine the two treatments on an earthquake dataset. The results indicate that the treatment with the depth variable generates a higher fitness value, highest reward, and accuracy.

5. REFERENCES

[1] Dzwinel, W., Yuen, D.A., Boryczko, K., Ben-Zion, Y., Yoshioka, S. and Ito, T., "Clustering Analysis, Data-Mining, Multi-dimensional Visualization of Earthquakes over Space, Time and Feature Space," *Communications of the ACM*, vol. 26, 1998.

[2] Eick, C., Vaezian, B., Jiang, D. and Wang, J., "Discovery of Interesting Regions in Spatial Datasets Using Supervised Clustering," In *Proc. of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases*, Berlin, Germany, September 18-22, 2006, vol. 4213, pp. 127-138.

[3] Ester, M., Kriegel, H., Sander, J., and Xu, X., "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," In *Proc. of 2nd International Conference on Knowledge Discovery and Data Mining,* Portland, Oregon, Aug. 1996, pp. 226-231.

[4] Hinneburg, A. and Keim, D.A., "An Efficient Approach to Clustering in Large Multimedia Databases with Noise," In *Proc. of the 4th International Conference on Knowledge Discovery and Data Mining,* New York City, August 1998. pp. 58-65.

[5] Jiang, D., Eick, C.F. and Chen, C.S., "On Supervised Density Estimation Techniques and Their Application to Spatial Data Mining," *Proceedings of the 15th ACM International Symposium on Advances in Geographic Information Systems,* Seattle, WA, Nov. 7-9, 2007.

[6] Ng, R. T. and Han, J., "Efficient and Effective Clustering Methods for Spatial Data Mining," In *Proceedings of the 20th VLDB Conference,* Santiago, Chile, 1994, pp. 144-155.

[7] Tay, S.C., HSU, W., Lim, K.H. and Yap L.C., "Spatial Data Mining: Clustering of Hot Spots and Pattern Recognition," *The IEEE International Geoscience and Remote Sensing Symposium (IGRASS),* Volume 6, 21-25 July 2003, pp. 3685-3687.

[8] Williams, G.J., "Evolutionary hot spots data mining – an architecture for exploring for interesting discoveries," In *Proceedings of the 3rd Pacific-Asia Conference on Methodologies for Knowledge Discovery and Data Mining,* London, UK, 1999, vol. 1574, pp. 184-193.