Application Research of Clustering Algorithm in Earthquake Disaster Prediction System of Sanhe

Changxiu Hu, Zhiyuan Cui, Jinglei Lin, Meng Huang, Shuai Liu
Institute of Disaster Prevention, P.O. Box 065201, HeBei, SanHe, China
hm@cidp.edu.cn 1982312765@qq.com

Abstract. In order to better provide decision-making technical support for the government after the earthquake, this paper takes all buildings in Sanhe as the research target and simulates the earthquake damage based on the circular intensity attenuation algorithm. According to the earthquake resistance level of the building and the type of building, the damage of the building can be judged, and the rescue efficiency and effect can be improved more efficiently. This project uses artificial intelligence clustering algorithm to cluster analysis of all buildings according to the damage level, and combines the GIS platform to display the analysis results in the form of graphs. Finally, the system used PHPWord technology to provide a detailed output of a detailed earthquake damage prediction analysis document, in order to efficiently arrange and schedule emergency rescue work.

1. Introduction
Sanhe is located in the core area of the Golden Triangle, which is composed of Beijing, Tianjin and Tangshan. It is an important place in the integration of Beijing-Tianjin-Hebei and the Bohai Economic Circle. In the event of a major earthquake, the economic losses caused by the personnel were incalculable, so it was particularly important to carry out rescue quickly and accurately after the earthquake. The traditional method of post-earthquake information collection mainly used artificial site survey to obtain disaster information about the earthquake. This way required a lot of manpower and material resources, and the speed was slow and the cycle was long. There was no way to help earthquake rescue work in time. Therefore, the timing of the rescue was delayed. Therefore, grasping the disaster information and reasonable rescue suggestions in the first time after the earthquake can effectively avoid the waste of time and human resources, and make the rescue operations more efficient. The goal of this system was to quickly and accurately obtain information on the damage status of buildings in Sanhe after the earthquake, and provide detailed disaster information for earthquake emergency rescue work in a timely manner.

This paper took the building information provided by the Sanhe Earthquake Administration as the basic research unit, and combined Baidu map to establish a GIS spatial database. The spatial clustering idea and the intensity algorithm were used to deal with the problem of regional division. Based on the obtained clustering results, through accurate organization and regulation, it provided an accurate and effective earthquake damage prediction document for earthquake emergency response, which facilitated the efficient arrangement and dispatched of emergency rescue work.

2. Data analysis
This paper analyzed based on the information of more than 20,000 buildings provided by the Earthquake Administration of Sanhe and the pictures of buildings of nearly 246GB. Due to the complexity of the
original data and the large amount of data, the GIS spatial database had the characteristics of large data volume, high accessibility, joint management of attribute data and spatial data. The GIS spatial database was consistent with the original data characteristics, so GIS spatial database technology was used for data storage.

Because the original data was not conducive to the management of the later data and more specific reflection of the disaster information, the data was processed by Arcgis. This paper use ArcGIS to complete the conversion of surface data to line data and line data to point data, which facilitated more precise mining and analysis of data. The system finally created a total of 48 data tables, as shown in Figure 1.

![Figure 1. Database information](image)

Due to the large amount of data in the building picture, there were about 200,000 pictures, and each building had a number of pictures. In order to ensure the high correlation between the building dataset and the image resources, the system generated the file index information of the building dataset so that the data and the image information can be matched quickly and accurately.

These data mainly covered 11 types of buildings including civil buildings, public buildings, industrial buildings, bridges, communication systems, power systems, gas systems, gas gate stations, heating systems, secondary disaster sources, and water supply systems. Buildings of different building types were subject to different degrees of damage when they are affected by force majeure factors, and may even cause secondary damage, and the resulting impacts were greatly different.

For example, when a magnitude 7.3 earthquake occurred in Haicheng, Angang caused a series of problems, such as blast furnace shutdowns and power outages caused by freezing of molten iron. The destruction of hydropower equipment in Yingkou had caused the city to save water and the city to be paralyzed, causing huge economic losses. At the same time, secondary disasters also caused great troubles for the development of post-disaster relief work. Therefore, how to estimate the damage of the city in time based on the existing building data, and provide first-hand rescue information support for the post-earthquake rescue work was the focus of this paper.

After comprehensively exploring various building information units, this paper combined the output of 48 data tables to explore the hidden complexities between building information units based on the main attributes such as latitude and longitude information, building number, intensity information, building type and foundation type. The project team firstly divided the data into several feature sets based on the correlation between various building types and the latitude and longitude coordinate information and the characteristic attributes of the intensity information. In order to characterize the similarity of data in different regions, the ultimate goal was to use clustering algorithm to explore the disaster situation of Sanhe after the earthquake. The system used the GIS platform as the bottom layer, applied the spatial computing method, and combined the characteristic attributes of the data clusters to integrate, extracted and analyzed the spatial distribution of the affected areas according to the existing earthquake disaster impact factors, and established a mathematical model for disaster assessment. In order to obtain the distribution situation of special buildings in space, analyze the spatial correlation of continuous domains, and the evolution characteristics of disasters in space. In this way, the damage of the building was assessed, and the risk comprehensive assessment of the special buildings was carried out to avoid the interference of the secondary disasters, and the regional environment of the earthquake-stricken areas can be understood through the picture information before the earthquake. Therefore, rescue work can be carried out in a timely, accurate and efficient manner.

3. Main technical methods

The clustering algorithm was a typical algorithm for unsupervised learning and did not require tagging results. By trying to explore and discover certain patterns, to find common groups, the data was divided
into multiple categories according to the inherent similarity, so that the internal similarity was large and
the internal similarity was small. In short, the similarity of the same kind of objects was as large as
possible, and the similarity between different types of objects was as small as possible.
Commonly used clustering algorithms are as follows:

- **Clustering based on partitioning**: Given a data set with N tuples or records, the splitting method
  constructed K packets, each of which represented a cluster (condition: \( K \leq N \)). For a given \( K \),
  the grouping was changed by iterative iteration so that the records in the same group were as close as
  possible, and the records in the different groups were as far apart as possible. Create an initial
  partition by giving the number \( K \) of partitions built. Then, iterative relocation technique was used
to repeatedly change the grouping so that the objects in the same cluster were as relevant as possible,
and the objects in different clusters were as far as possible.

- **Hierarchical-based clustering**: Hierarchical clustering was divided into merged hierarchical
  clustering and split hierarchical clustering. The former was a bottom-up hierarchical clustering
  algorithm. The cluster started from the bottom layer, and each time the clusters in the previous level
  were formed by merging the most similar clusters, and the whole was completed when all the data
  points were merged into one cluster or a certain termination condition was reached. The latter
  adopted a top-down approach. Clustering started with a cluster containing all the data points, then
  split the root node into sub-clusters, and each sub-cluster continued to recursively continue to split
  until it appeared. A single-node cluster containing one data point appeared and the cluster stopped.

- **Density-based clustering**: Density-based clustering was characterized by not relying on distance,
  but on density, thus overcoming the shortcomings of distance-based algorithms that only
  "spherical" clustering can be found. The core idea was that as long as the density of points in a
  region was greater than a certain threshold, it was added to the clusters that were close to it.

- **Network-based clustering**: This kind of clustering usually divided the data space into a grid
  structure of a finite number of cells, and all processing was performed on a single unit. This was a
  very fast process because it was independent of the number of data points and only related to the
  number of cells.

- **Model-based clustering**: Model-based clustering assumed a model for each cluster and then looked
  for a data set that fitted the model well. The model may be a density distribution function of data
  points in space or something else. Such a method usually contained the underlying assumption that
  the data set was generated by a series of potential probability distributions. There were usually two
  attempts: statistical methods and neural network methods.

- **Fuzzy based clustering**: Clustering based on fuzzy set theory meant that samples belong to a certain
  class with a certain probability. More typical fuzzy clustering has fuzzy clustering based on
  objective function, similarity relationship and fuzzy relation.

In summary, because the system was mainly to quickly predict the detailed disaster area after the
earthquake, and provide a detailed earthquake damage prediction document for the emergency disaster
after the earthquake, it was more appropriate to use the network-based clustering. In this paper, the
clustering algorithm based on combination of square and distance was selected. Because the clustering
algorithm based on square and distance not only overcomes the given problem of \( K \) value, but also overcomes the
disadvantage that the data in the grid clustering algorithm must be spatially dense.

### 3.1 Clustering algorithm based on square and distance

At the beginning there were no known aggregation points on the map, so it was necessary to traverse all
the points on the map to calculate the outsourcing square of the points. If the outsourcing square that
calculated this point did not intersect the outsourcing square of the existing aggregation point, a new
aggregation point was created, and if it intersected, the point was added to the aggregation point.

Main operations:

I. Initialize the aggregation point.

II. Loop through the data points to determine if the next point intersects with aggregation point 1. If
so, the point should be aggregated into aggregation point 1; otherwise the point generates an
aggregation point 2. (If there was an intersection with multiple outsourcing squares of multiple aggregation points, it was necessary to judge that the distance of the aggregation point was relatively close, and it was classified into a relatively close aggregation point.)

III. Loop iteratively through the b process until it was completed.

However, due to the influence of seismic intensity, the clustering algorithm alone cannot achieve the actual target. Therefore, the system used a clustering algorithm based on square and distance and a circular intensity algorithm to quickly estimated the post-earthquake disaster area.

3.2 Circular intensity algorithm
The seismic intensity was the intensity of the impact of the earthquake on the surface and engineering buildings, and the seismic intensity from different regions of the epicenter was affected by the attenuation of seismic waves. In different geographical environments, the range of impacts of earthquakes varies. According to different geological environments, earthquake researchers had proposed a variety of models to match the distribution of intensity in various places. Sanhe was located in the Yanshan piedmont plain, with a uniform geological distribution, which was more suitable for the circular intensity algorithm.

\[ I_d = I_0 \times e^{-KD} \]

Among them, \( I_0 \) is the maximum epicenter intensity, \( D \) is the epicentral distance, \( e-KD \) is the attenuation factor, \( K \) is the medium absorption coefficient, and \( I_d \) is the intensity when the epicentral distance is equal to \( D \). The statistical relationship between the reference magnitude and the epicenter intensity, combined with the attenuation model, the system can dynamically generate an earthquake attenuation map. In the actual calculation process, the medium and long axis lengths were modified by judging the type of the passed parameters, and the corresponding output set was obtained.

This paper combined the results of seismic experts' research on the damage degree of various types of buildings under different intensities and the circular intensity algorithm to calculate the intensity value \( I_d \) of the building. This paper used the concept of statistics to carry out induction and classification, and conduct preliminary screening and statistics on the data. Finally, the system used the results of the screening as the basic research unit, and used the clustering algorithm based on square and distance for rapid regional division.

4. Application of Clustering Algorithm on Circular Intensity Diagram
The system used a circular intensity algorithm to draw a circular intensity map to simulate the damage of buildings after different levels of earthquakes. Then, it added the building points on the map through the API interface of Baidu map massive points, and finally used PHPWord technology to export the results. The circular intensity algorithm can be used to calculate the seismic intensity value \( I_d \) obtained from all building points. The damage degree was divided into six levels by the attributes of the building and the damage that the earthquake experts had suffered under different intensities of various types of buildings. The specific grades were: not in the main intensity circle, good, slight damage, medium damage, serious damage, and destruction.

The intensity value \( I_d \) calculated by the circular intensity algorithm can quickly label the affected building with a level of damage level. On this basis, this paper used the clustering algorithm based on square and distance to quickly divide the buildings against severely damaged and destroyed buildings. The effect is shown in Figure 2.
In the end, the system summarized the building information with more serious disasters, and combined the statistical ideas to use the Echarts technology to statistically analyze the overall damage of the earthquake damage. As shown in Figure 3, the percentage of various damage levels and the affected buildings can be clearly seen. In Figure 3, there were 18,039 buildings affected, of which 0 were destroyed, 2,569 were seriously damaged, 13,231 were moderately damaged, 2,037 were slightly damaged, and 202 were good. The system extracted the construction sites that were seriously damaged and accounts for 14.24% of all buildings. Next, the Baidu map inverse resolution address interface was used for positioning, and finally the analysis results of 46 regions were generated as Word documents for one-click output, as shown in Figure 4. The output earthquake damage prediction document can help earthquake emergency rescue, so as to efficiently arrange and dispatch emergency rescue work.
5. Conclusion
Aiming at the problem of tight rescue time and lack of information after the earthquake, this paper proposed to analyze the damage status of the building after the earthquake based on the square and distance clustering algorithm, in order to further improve the accuracy, universality, efficiency and fault tolerance of the algorithm. The system combined the circular intensity algorithm to classify the intensity level, and obtained the information of the hardest-hit area after the earthquake for rapid output, so as to provide timely information support for emergency rescue work. In the future research process, how to efficiently process large-scale building data and further improve the clustering performance of clustering algorithms in disaster analysis is the focus of the next step.

ACKNOWLEDGMENTS
This work was supported by the Special Fund of Fundamental Scientific Research Business Expense for Higher School of Central Government (No. ZY20180124 and No. ZY20160106).

Reference
[1] Aizhu Zhang. High-resolution seismic image segmentation based on intelligent optimization algorithm [D]. China University of Petroleum (East China), 2014.
[2] Qiming Tian, Wang Lizhen, and Yin Qun. Design, implementation and application of clustering algorithm based on grid distance[J]. Journal of Computer Applications, 2005(02): 294-296.
[3] Pei Du, and Cheng Xiaorong. A Comparative Density Peak Clustering Algorithm Based on K Near Neighbors[J]. Computer Engineering and Applications, 1-11.
[4] Fang Gong. Design of Intelligent Assessment System for Earthquake Damage of Group Buildings[J]. Bulletin of Science and Technology, 2018, 34(11): 164-167.
[5] Shaozhang Niu, Yiyi Ou, Jie Ling, and Guosheng Gu. Multi-density fast clustering algorithm using region partitioning[J]. Computer Engineering and Applications, 1-9.
[6] Zhuang He. Grid-based weighted average density adaptive clustering algorithm and its application [D]. Hunan University, 2012.