Refined range analysis of early warning and pre-control based on multi-element spatiotemporal clustering

Min Tu

1Jiangxi Police College, China

Abstract. There are three parts in public security big data analysis when early warning and pre-control oriented: refined analysis scope, visualization of temporal and spatial characteristics of analysis scope, and subspace emergency analysis model for early warning and controlling. These three parts are interrelated, and complete the early warning and pre-control analysis of emergency iteratively. The exploration method of refined analysis range is analysed in the paper, which is based on the calculation and comparison of multi-element spatiotemporal clustering results. This research adopts a grid division strategy to convert heterogeneous data into a normalized space-time grid, in order to improve the versatility of the framework. And the estimation of grid records without observed values using kernel density estimation method. The added clustering ensemble is to improve the stability of clustering results. It can avoid the influence of uncertain factors on the clustering results, as well as improving the real-time performance of clustering algorithm, by the optimization of clustering ensemble parameters, and the clustering ensemble deployed on Hadoop platform.

1. Introduction
Public safety incidents and cases always occur in a specific time and space range. Accurately delineating their space-time range and predicting the occurrence range of cases and events play a very important role in focusing on the scope of big data analysis and optimizing patrol law enforcement police force [1-2]. At present, there are still two problems in the delineation of the scope of analysis: First, A general high-incidence scope is usually preset by researchers based on their experience. This coarse-grained range estimation may introduce non-case high-incidence areas, thereby increasing the amount of data analysis and affecting the significance of the analysis results, or may split the complete high-incidence areas, making it difficult to find results beyond this range. Second, public safety incidents and cases are often accompanied by different symbiotic events, and there are specific temporal and spatial constraints between symbiosis and selected events. Researchers can analyze the causes of selected events based on this model and predict their further development trends and the scope of influence. In response to these two problems, this paper proposes a public safety big data analysis method for early warning and pre-control.

The occurrence of a case or event is often accompanied by multiple related events, which is called co-occurrence. Discovering the co-occurrence rules of a certain type of specific event, that is, finding the symbiotic event is very important for discovering the mechanism of the event, analyzing the scope of the event, and predicting the occurrence and scope of the event.
It is very challenging to discover the symbiosis of a particular event. Various types of symbiosis events often have specific time and space constraints [3]. Symbiotic events occur not only in the selected event location, but also in any spatial location, which is the case of teleconnection in the real world. On the other hand, the symbiotic event and the selected event do not necessarily occur at the same time, that is, there is a specific time interval. If the symbiosis event always occurs before the selected event, it may be the cause of the symbiosis event, to a certain extent, the symbiotic event can be predicted, and if the symbiotic event always occurs after the selected event, it often indicates the impact of the outcome, that is, what the selected event will cause. The symbiosis rules can be effectively used to improve the prediction effect of different events and cases in big data.

Public safety big data analysis for early warning and pre-control includes three parts: refined analysis scope, which is used to find a number of areas with high risk of emergencies from massive raw data; visualization of temporal and spatial characteristics of analysis scope [4-5], used for overview visualization. The high-level spatio-temporal characteristics of these areas and their emergencies; the sub-space emergencies analysis model for early warning and pre-control is used to select parts from the overview view to carry out visual analysis for early warning and pre-control. The three parts are interrelated and iteratively complete the early warning and pre-control analysis of emergencies. This paper analyzes in detail the exploration method of refined analysis scope based on the calculation and comparison of multi-element spatiotemporal clustering results.

2. Refined analysis range

The distribution of emergencies in a city is often unbalanced in time and space. Emergencies and cases in certain regions and time ranges often occur frequently, while in other regions and ranges often occur less frequently. By externally performing spatio-temporal clustering and aggregation of emergencies, the distribution characteristics of the events in the space-time range can be discovered. By visually browsing and comparing the spatial distribution, the regions and time intervals with higher event ranges can be accurately located. These areas are the focus of follow-up research and provide necessary analysis objects for follow-up research.

2.1. Analytical grid establishment and vacancy value estimation

The grid division strategy is adopted to divide the space covered by the data into several rectangular grids of equal size. Each grid may contain 0 to multiple data points, and its value represents the value of the grid. This research needs to use a grid with a value to estimate a grid without a value. The method of kernel density estimation is introduced, that is, use the nearest k grid values with observation sites to infer the grid values without observation records. In addition, an improved kernel density estimation method with dynamic bandwidth is proposed.

Let $A=\{a_1, a_2, \ldots, a_n\}$ be the grid set to be estimated, and $B=\{b_1, b_2, \ldots, b_n\}$ be the grid set with observation records. The estimated value of $a_i$ is:

$$\hat{f}(a_i) = \frac{f(b_j)}{N} \sum_{j=1}^{N} \frac{1}{\max(h, d_{ik})} K\left(\frac{d_{ij}}{\max(h, d_{ij})}\right)$$

(1)

Where $h$ is the minimum allowable bandwidth, $d_{ik}$ and $d_{ij}$ are the distance from $a_i$ to the k-th nearest valued grid and the distance to $b_j$, respectively, and $f(b_j)$ is the average recorded value of $b_j$. For the convenience of calculation, epanechnikov kernel is used in this project, the kernel function is $K(u) = \frac{3}{4} (1 - u^2) \text{I}(\|u\| \leq 1)$. When the constraints are met, $l=1$, otherwise $l=0$. Equation (1) uses dynamic core bandwidth to obtain a more continuous and flat effect.

In order to further reduce the randomness of the kernel width, a Gaussian kernel based on dynamic covariance is also proposed. Let $a_i$ be the coordinate of the grid to be estimated, and $\mu$ be the weighted average coordinate of the k nearest valued grids of $a_i$, that is, $x_\mu = \sum_{i=1}^{k} f(b_j) x_i / \sum_{i=1}^{k} f(b_j)$. $y_\mu$ is the same. The estimated value of $a_i$ is:

$$\hat{f}(a_i) = \frac{1}{2 \pi |V|} e^{-\frac{1}{2}(a_i - \mu)^T V^{-1}(a_i - \mu)}$$

(2)
\[
V = \begin{bmatrix}
\sigma^2_x & \text{cov}_{xy} \\
\text{cov}_{xy} & \sigma^2_y
\end{bmatrix}
\]  \hspace{1cm} (3)

\(V\) is the covariance matrix of the k nearest valued grid values to \(a_i\), where \(\sigma^2_x\) and \(\sigma^2_y\) are the variances of the \(x\) and \(y\) coordinates of the \(k\) nearest valued grids to \(a_i\), \(\sigma^2_x = \sum_{i=1}^{k} (x_i - \mu_x)^2 b_i / (\sum_{i=1}^{k} b_i - 1)\), \(\sigma^2_y\) is the same. \(\text{cov}_{xy}\) is the weighted covariance of the \(x\) and \(y\) coordinates, \(\text{cov}_{xy} = \sum_{i=1}^{k} b_i (x_i - x_m)(y_i - y_m) / (\sum_{i=1}^{k} b_i - 1)\).

In theory, the above two methods meet the requirements. Relatively speaking, the Epanechnikov kernel has simple calculation and fast convergence speed, while the Gaussian kernel function has high calculation accuracy and smooth convergence path, but the calculation speed is slightly slower.

2.2. Cluster ensemble

The purpose of spatiotemporal clustering is to reduce the time and space complexity of each grid data. The standard clustering algorithm has some shortcomings in determining the clustering parameters and the stability of the clustering results. Therefore, A clustering ensemble algorithm based on co-correlation matrix is adopted. Let \(G=\{g_1, g_2, \ldots, g_N\}\) be the standard grid set. An \(n \times n\) co-ordination matrix \(M\) (\(n\) is the number of grids) is defined to determine the association between any two grids, Then \(g_{ij}\) is defined as:

\[m_{ij} = \frac{1}{T} \sum_{t=1}^{T} \delta \left( \pi_t(g_i), \pi_t(g_j) \right)\]  \hspace{1cm} (4)

\(T\) is the number of base clustering. The purpose of dividing by \(T\) is to normalize the matrix. \(\pi_t\) is a basic cluster, \(\pi_t(g_i)\) and \(\pi_t(g_j)\) represent the class numbers to which \(g_i\) and \(g_j\) are assigned in each basic cluster. If \(\pi_t(g_i) = \pi_t(g_j)\), then \(\delta \left( \pi_t(g_i), \pi_t(g_j) \right) = |\pi_t|\), otherwise \(\delta \left( \pi_t(g_i), \pi_t(g_j) \right) = 0\). Among them \(|. . .\)| is the operation of taking the number of clusters. This project not only counts the number of times the two grids are grouped together, but also considers how many classes are assigned to each basic cluster. In the case of many divisions, if the two grids are still divided into one class, then the weights of positions corresponding to \(g_i\) and \(g_j\) in the co-correlation matrix are increased, so this project uses the number of classes generated by each basic cluster as the weight. Repeating \(T\) times to establish the basic clusters, the relationship between the grids can get stable estimation results. The base cluster generation method and the final cluster merging algorithm will be introduced in the next step.

In terms of generating basic clusters, existing literature has compared the effects of hierarchical clustering, Gaussian mixture model clustering, DBSCAN density clustering, and K-Means clustering on the spatial clustering of ocean grid data through experiments. According to its research conclusions, K-Means clustering and Euclidean distance are used to generate basic clusters by changing the number of clusters.

In order to merge multiple basic clusters, first define the evaluation index of the merge effect. Let \(\pi_1 = \{c_1, c_2, \ldots, c_m\}\) and \(\pi_1 = \{c^1, c^1, \ldots, c^n\}\) are two base clusters, namely \(\pi_1\) and \(\pi_1\) divide the data set into \(m\) and \(n\) categories, and define \(\theta(\pi_1, \pi_1)\) as the normalized mutual information NMI (Normal Mutual Information), the purpose of cluster ensemble is to make the maximum NMI between the base cluster obtained by \(T\) repetitions and the final merged cluster \(\pi_A\), that is, \(\pi_{opt} = \arg \max \sum_{i=1}^{T} \theta(\pi_1, \pi_A)\). As the NMI calculation formula is relatively complex, this project will use a heuristic algorithm, the specific ideas are:

K-means algorithm is repeated \(T\) times to get \(T\) base clusters;
Calculate the co-ordination relationship matrix, Each weight value represents the normalized number of \(N\) grids to be classified into a class;
Iterate \(L\) times, set a threshold for each iteration, the grid relationships higher than the threshold is selected, and merge them into multiple subgraphs.
If the number of subgraphs does not change in successive iterations, the number of classes in this period of time is considered to be relatively stable.

The time complexity of the K-Means method is \( O(NKP) \), \( N \) represents the total number of elements, \( K \) represents the number of cluster centers, and \( P \) represents the number of iterations (\( K, P \ll N \)). Therefore, The complexity of iteratively executes \( T \) times K-Means(\( T \ll N \)) is about \( O(KPTN) \), where \( KPT \) is a constant. The complexity of establishing a co-coordination relationship matrix is \( O(TN^2) \). The complexity of the L-time graph merging algorithm is \( O(LN) \). So the total time complexity of the above steps is \( O(NKP)+O(TN^2)+O(LN) \). Since visual analysis has high requirements for interactive real-time, in order to improve the system's computing speed, this project plans to complete the clustering integration process based on the Hadoop platform. A total of 4 MapReduce jobs are constructed. Job1 is responsible for generating the initial cluster centers, Job2 is responsible for generating \( T \) base clusters, Job3 is responsible for building the co-coordination relationship matrix, and Job4 is responsible for generating the final merged clusters. This architecture makes full use of the parallel advantage of MapReduce, can effectively improve the speed and stability of spatiotemporal clustering with different parameters.

3. Conclusions
There are many special processes that exist in reality, including mass emergencies data, complex correlations between emergencies and different data sets, huge differences in temporal and spatial distribution in different regions, similarities between emergencies, teleconnections, which pose great challenges to the method design. In this study, analysis of template exploration is classified into the category of comparative visualization.

This research adopts a grid division strategy to convert heterogeneous data into a normalized space-time grid, in order to improve the versatility of the framework. And the estimation of grid records without observed values using kernel density estimation method. The added clustering ensemble is to improve the stability of clustering results. It can avoid the influence of uncertain factors on the clustering results, as well as improving the real-time performance of clustering algorithm, by the optimization of clustering ensemble parameters, and the clustering ensemble deployed on Hadoop platform.

With the development of big data technology, it is possible to analyze and realize in-depth data mining capabilities that cannot be accomplished by previous technologies, based on the centralized storage of massive big data. In addition, it will be easier to provide valuable early warning analysis, estimation and optimization results, which can promote the intelligent linkage of information, and rise the risk pre-control level.

Acknowledgement
This project is supported by the key project of Jiangxi Provincial Department of Education Science and Technology Project "Public Security Big Data Analysis for Early Warning and Pre-control" (No. GJJ191026).

References
[1] ZhiMao Lu, Chun Li, Qi Zhang, text clustering ensemble spectral algorithm based on nearest neighbor propagation [J], Journal of Harbin Engineering University, 2012.7899-905
[2] Sen Xu, Jun Gao, XiaoPeng Hua, XianFeng Li, Jing Xu, an improved adaptive clustering ensemble selection method [J], Acta automatica Sinica, 2018.112103-2112
[3] LingChao Hu,Hong Yu,a voting based three branch decision clustering ensemble method [J], small microcomputer system, 2016.81741-1745
[4] ZiTang Ma, Jie Gou, FCM clustering ensemble algorithm based on mapred uce [J], computer application research, 2016.123554-3558
[5] JianHua Jia, research on spectral clustering ensemble algorithm [M], Tianjin University Press, 2011.8