Research on Data Management of Marine Disaster Mitigation Based on Big Data

Chen Chen*, Xunan Liu, Yao Zhang and Rui Tong
National Marine Hazard Mitigation Service, Beijing 100194, China

*Corresponding author: chenchen@nmhms.org.cn

Abstract. With the explosive growth of ocean data, ocean big data has received more and more attention. It mainly analyzes and summarizes the current research status and key technologies of ocean big data, and focuses on the examples of machine learning model prediction research in ocean big data. With the advancement of 3S technology, China's marine monitoring network has achieved all-round development, including air, ground, and sea monitoring platforms, which has promoted the rapid growth of marine big data from the original GB and TB levels to PB levels. Taking the typhoon disaster search and situation display implementation as an example, the thesis has preliminarily completed the realization and application of the intelligent ocean search engine. Combined with the design of typhoon disaster search and the overall structure and function design of the situation display, module development was carried out, and the development was initially completed to realize the business data management of marine typhoon disaster mitigation based on big data.

Keywords: Big data, marine disaster mitigation, typhoon disaster, machine learning model.

1. Introduction
Marine big data technology can provide brand-new solutions and industrialization ideas for marine disaster early warning and auxiliary decision-making, marine ecosystem dynamic monitoring and marine economic strategic research, and promote the all-round development of marine natural science and social science. For example, through the analysis of Argo data, we can find that the earth is looking for an enhancement of the global hydrological cycle; through the analysis of acoustic remote sensing data, we can obtain the distribution of biological communities and species in the ocean, providing a strong scientific reference for ensuring the ecological balance of the ocean; By analysing the observation data of seismic activity, fault activity, and mid-ocean ridge magmatic activity acquired by the "Neptune" program, it is possible to make early warning predictions of seabed earthquakes and tsunamis. Ocean data can realize prediction, early warning, and auxiliary decision-making in interdisciplinary fields such as ecology, climate, and disasters. Therefore, the research on the management of ocean big data storage, analysis, query and other aspects has important strategic and practical significance for ecosystems, human society, scientific research and other aspects.

With the large-scale layout of the marine Internet of Things, the problems of imperfect integration standards of marine information resources, unsmooth information transmission channels, and poor
data access and communication issues have been highlighted. The pressure on the second-level data processing and data access of marine monitoring systems and information platforms is increasing. Intensified, such as: marine emergency and early warning applications usually access data sources of up to hundreds of G, and high-frequency and high-intensity queries of data are difficult to guarantee; massive data processing during polar scientific research not only causes calculation pressure on shipboard equipment, but also it is difficult to confirm the efficiency of data query and extraction; in the process of strange tide decision-making assistance and numerical simulation calculation, it is necessary to provide powerful real-time processing and query capabilities in order to give full play to the benefits of rapid flood tide disaster data. Ocean data has the characteristics of decomposable tasks, decomposable data, decomposable functions, and decomposable operation granularity, making it very suitable for block storage and parallel query processing. Therefore, the use of distributed integration technology to design marine big data storage models has preliminarily solved the problems of marine data format "inconsistent", frequency of use "inconsistent", information "insecure", and query feedback "untimely". Informatization development letter must overcome technical barriers [1].

2. Overview of ocean big data

With the rapid development of current network technology and information technology, the era of big data was born. According to the four characteristics of big data, IDC defines big data as massive data scale, fast data flow and dynamic data system, diverse data types, and huge data value. In addition, the basic feature of big data is that the data has strong authenticity. The ocean area occupies a large proportion of the ground surface area. In order to further understand and understand the ocean, and play an important role in the ocean, people have invented many detection equipment and technical means to survey and investigate the ocean. So far, a large-scale observation system has been formed, and a large number of accumulations have gradually accumulated in several explorations. Data and materials, such as field survey data, marine remote sensing data, etc., this type of data is very large, and has heterogeneous characteristics. Since the development of science and technology, people have realized that the ocean can play an important role in predicting climate, and it is a key region to grasp global climate change. With the advancement of technology, the accuracy of surveying the ocean has been continuously improved, the efficiency has been effectively improved, and the network nodes have expanded rapidly. The speed and scale of data growth are much higher than other industries. The development of ocean big data can enhance the ability of data fusion and personal retrieval, the ability to calculate ocean data, the ability to visualize data in multiple dimensions, and the ability to support decision support. In addition, the work efficiency of fusion and data processing can be improved, the mining of ocean data models can be expanded, the degree of display data can be improved, and the ability to share professional experience and knowledge can be inherited. Among the basic characteristics of ocean big data, more attention is paid to diversification and high value. Among them, the diversified characteristics promote the role of ocean big data. High value as its core, mainly including the value of the data itself and the potential contained therein, is highly valued by scientific researchers worldwide. Exploring the important value of marine big data is a prerequisite for promoting the realization of the strategy of a maritime power [2].

3. Big data processing method of marine disaster

3.1. Multi-source heterogeneous data pre-processing

The pre-processing mainly carries out the steps of matching, interpolation and quality control of various data for the distribution density of ships to be predicted. The data features used here present multi-source heterogeneity, including data from 1 to 3 dimensions in different fields and feature information. It is necessary to pre-process multi-source heterogeneous data for calculating the distribution density of ships. Finally, a set of spatio-temporal matching multi-source heterogeneous fusion data sets is obtained, which lays the foundation for the subsequent training and prediction.
research. The ship density here is grid-processed using AIS data, and then summed up statistics for each grid data.

Typhoon best path data is saved in text format, which is a one-dimensional array of mixed numeric and character information. First, select the time period of transiting the South China Sea from the typhoon best path data. For these typhoon time period data, linear interpolation method is used to interpolate the positioning data once every 6h to the time resolution of 1h. Since the ship density of geographic gridding may be closely related to the distance of the typhoon centre, the absolute distance between the centre point of each grid and the centre of the typhoon must also be solved here using the Earth coordinate closest distance algorithm. N C E P reanalysis data is three-dimensional data stored in meteorological standard network common data format. Since the time resolution is not high, the time weighting method is used here for interpolation processing, and the calculation formula is as follows:

\[
\begin{bmatrix}
    P_1 \\
    P_2 \\
    P_3 \\
    P_n \\
\end{bmatrix} = w_1 \begin{bmatrix}
    P_1^1 \\
    P_2^1 \\
    P_3^1 \\
    P_n^1 \\
\end{bmatrix} + w_2 \begin{bmatrix}
    P_1^2 \\
    P_2^2 \\
    P_3^2 \\
    P_n^2 \\
\end{bmatrix}
\]

(1)

Here, P1 ~ Pn represent the first to n parameters (mainly including air pressure, air temperature, etc.) to be obtained, w1 and w2 represent the time weight of the reanalysis data at each moment, and Pn1 and Pn2 represent the parameters corresponding to the two times before and after. Finally, the time-interpolated three-dimensional data is interpolated to 0.5° × 0.5° (50km) horizontal resolution for matching. The results of the characteristic analysis show that the changing characteristics of the parameters such as temperature and relative humidity in disaster weather are not obvious and have little correlation with the behaviour of the ship, while the wind field, air pressure and precipitation have significant changes in response to disaster weather. As a characteristic parameter of disaster weather (typhoon) [3].

3.2. Machine learning algorithm model

The decision tree model is a tree structure (which can be a binary tree or a non-binary tree), based on the process of classification or regression of instances. Each non-leaf node represents a judgment on a feature attribute, each branch represents the output of this feature attribute in its value range, and each leaf node stores a category. The decision-making process using the decision tree is to start from the root node, test the corresponding feature attributes in the item to be classified, and select the output branch according to its value until it reaches the leaf node, and use the category stored in the leaf node as the decision result. Branch node selection is to find the optimal solution of the branch node. Since we are looking for the best, then we must have a measure, that is, we need to quantify the advantages and disadvantages. Commonly used measurement indicators are entropy and Gini coefficient.

\[
Entropy = -\sum_{i=1}^{n} p_i \log(p_i) \tag{2}
\]

Gini coefficient: As above, it can also be used as a measure of the degree of information confusion.

\[
Gini = 1 - \sum_{i=1}^{n} p_i^2 \tag{3}
\]

With quantitative indicators, you can measure the convergence effect of the degree of information confusion before and after using a certain branch condition. Using the degree of confusion before
branching, minus the degree of confusion after branching, the greater the result, the better the effect. Calculation of mean square error:

\[
\sigma = \left( \sum_{i=1}^{n} (x_i - \mu)^2 \right)^{\frac{1}{2}} = \left( \sum_{i=1}^{n} x_i^2 - n\mu^2 \right)^{\frac{1}{2}}
\]  

(4)

3.3. Results analysis
Four indicators of mean square error, root mean square error, mean absolute error, and regular mean square error were used to analyse the model error. Model training set error (see Table 1) and test set error (see Table 2) show that the error of random forest model is much better than the error of decision tree and bagging method.

| Table 1. Error analysis of training set. |
|-----------------------------------------|
| Decision tree | Bagging | Random forest |
| NMSE | 0.1867324 | 0.2925411 | 0.009209107 |
| MSE | 8.5260402 | 13.367168 | 0.42047998 |
| RMSE | 2.9199384 | 3.6547472 | 0.64844277 |
| MAE | 0.9282378 | 1.1732391 | 0.183384508 |

| Table 2. Error analysis of test set. |
|-------------------------------------|
| Decision tree | Bagging | Random forest |
| NMSE | 0.19811220 | 0.29928840 | 0.02246815 |
| MSE | 8.19038180 | 12.37322160 | 0.92888123 |
| RMSE | 2.86188430 | 3.51755900 | 0.96378485 |
| MAE | 0.91082880 | 1.17281490 | 0.29733624 |

Table 3 and Figure 1 show the descending order of the feature importance of the random forest model. Random forest sets two importance’s for continuous variables, one is the average mean square error reduction percentage, and the other is the average node impurity reduction. The first place in variable importance is the age of the typhoon (time after the typhoon is generated); the second place is the time of the day, indicating that the ship’s behaviour response is different during the day or night; the third place is the grid from the nearest port distance; the 4th place is the latitude of the typhoon centre; the 5th place is the longitude of the typhoon centre; the 6th place is the pressure field; the 7th place is the distance from the typhoon centre. The impact of the wind and rain fields is small. The reason may be that before the typhoon landfall brought strong winds and heavy rainfall, the ship had entered the haven and will continue to stay until the atmosphere and rainfall weather improve. In addition, the different definitions of the importance of the two features result in different ranking results.

| Table 3. Random forest model feature importance ranking. |
|--------------------------------------------------------|
| variable | symbol | % Inc MSE | Inc Node Purity |
| Typhoon age | dot | 99.98895 | 238107.37 |
| Every day | ta | 54.86966 | 46363.45 |
| Distance from the centre of the grid to the shore | distance | 51.28649 | 2091026.23 |
| Latitude of typhoon centre at a certain moment | typhoon_lat | 47.44166 | 166999.35 |
| Longitude of typhoon centre at a certain moment | typhoon_lon | 44.58938 | 174932.34 |
| Pressure field at a certain moment | pressure | 36.53857 | 874671.41 |
| Distance from grid centre to typhoon | typhoon_distance | 33.98429 | 146821.91 |
| Wind field at a certain moment | wind | 31.81796 | 822703.11 |
| Rain field at a certain moment | rain | 24.38135 | 525582.96 |
The above results indicate that under typhoons and other disaster weather, the ship’s behaviour is significantly affected by the weather.

4. Taking Typhoon as an example to study marine disaster mitigation data system

4.1. Architecture of Ocean Big Data Platform

Its software and hardware platform architecture are divided into three aspects: data layer, technology layer and application layer. The basis of the ocean big data platform is the data layer, which includes the information data collected by all platforms, such as observing the ground, sky, ships, etc., and obtaining data in remote sensing, physics, biology, and chemistry. After obtaining the data, preprocess it and organize and manage marine big data in a unified mode; the technical layer is composed of marine big data fusion and analysis, forecasting, etc., can integrate relevant technologies, and develop cloud platforms, and it has the function of personalized search for relevant information and accurately predicts marine elements; the application layer is to collect application modules under the premise of data retrieval and integration technology to improve the openness and comprehensiveness of the marine application service management system (as shown in Figure 2), combining data sharing, information processing and other technologies with scientific research to promote its development [4].

Figure 1. Typhoon prediction random forest model feature importance ranking.

Figure 2. Application layer of ocean big data platform.
4.2. Data acquisition

The construction of the experimental platform first requires massive amounts of data, but manually obtaining the monitoring data of marine buoys from the network is a tedious and boring task. Web crawler toolkits such as RCurl and RVEST included in CRAN allow users to automatically obtain desired data and resources from the network. The NOAA Marine Buoy Data Centre provides monitoring data such as wind, wind speed, temperature, air pressure, and ocean waves for download and use, and these data are presented in a table format. So, specify the website and then use the table reading function read in R language. `table()` can grab data directly from the web page to the local. The acquired data is stored in character format, which will adversely affect the subsequent processing, so it is necessary to use `as.character()` function to convert character format data to a numeric format, and uses the name (`as.numeric()`) function to specify a corresponding name for each column of data, and finally uses `write.table()` function saves the data as a CSV file to facilitate the subsequent import of the data into the Oracle database.

4.3. Data management submodule

4.3.1. Display of typhoon transit route situation.

On the case search page, select the situation display, and judge it as a typhoon case according to the matching searched case type, and automatically jump to the typhoon situation display function. The case display page remains unchanged, and the situation display home page is expanded.

1. Select "Transit Route" in the content page of the transit route situation display, the system can automatically display the real-time speed information when the typhoon reaches a certain position on the map, and the dynamic typhoon animation simulates the trajectory of the trajectory along the transit route points, discretely outline the trajectory of the typhoon, and plot the intensity of the typhoon on the arrival point in different colours (green represents tropical depression, blue represents tropical storm, yellow represents strong tropical storm, orange represents typhoon, pink represents strong typhoon red represents super typhoon);

2. Hover the mouse over an arrival position on the map for more than 1 second to display real-time information about the current time of the typhoon at this arrival point. The real-time information includes the name of the typhoon, transit time, longitude, dimension, wind force, maximum wind speed, central air pressure, moving speed, moving direction, 7-level wind circle radius, 10-level wind circle radius, and the wind direction and intensity are displayed through the rose wind direction chart;

3. After "Transit Route", continue to check the "Affect Range". When simulating the transit trajectory or hovering the mouse over an arrival position on the map for more than 1 second, the area covered by the typhoon wind field is displayed in the shadow area at the arrival point. The shadows of the 7th-level wind circle radius and the 10th-level wind circle radius are respectively shown by different colour shadows.

4. Continue to check “Storm surge increase” after “Transit route”, and display the storm surge increase volume and wave height through the buffer analysis model near the arrival point of the simulated transit track, superimposing the typhoon storm surge and the astronomical tide analysis results the strength data of storm surge or meteorological tsunami is displayed by the visual model of superposition analysis.

4.3.2. Display of the situation of disaster-bearing bodies within the typhoon transit path.

Display of the loss situation in the typhoon transit path, select the content of the transit route situation display, and display the disaster undertaking the content in the body frame.

1. Select "Injury Situation", the casualty situation in the surrounding areas on the transit path is displayed in a bar chart, the horizontal axis is the administrative area unit, and the vertical coordinate is the number of casualties. The number of casualties in the surrounding area and the percentage of the affected population. When "Injury Situation" is selected and the mouse stays in the administrative area for 1 second, the number of casualties in this administrative area is displayed in the form of a table, the
number of casualties and the percentage of the affected population; and a detailed description of the casualty content of the case;

(2) Select "Economic Loss" to display the economic loss situation of the surrounding areas on the transit path. The horizontal coordinate is in units of administrative units, and the vertical coordinate is the economic loss situation. The bar graph shows that the specific loss figures are marked on the column, and the ratio of the economic output value in previous years after showing the economic losses in the surrounding areas is compared. When "Economic Loss" is selected, and the mouse stays in the administrative area for 1 second, the economic loss of this administrative area is displayed in the form of a form, with a detailed description of the loss content of the case and showing the economic loss in this area compared with the previous year's output value. The economic losses of various disaster-pregnant bodies and the comparison of output value before and after the disaster (revenue of beaches, output value of fishing farms, and output value of marine oil and gas fields, etc.) are also compared and displayed.

4.4. Database construction
The buoy monitoring data provided by the NOAA Marine Buoy Data Centre is updated quickly and the data volume is large. If the data obtained from the network is directly processed and displayed, it not only takes up a large amount of memory, but also does not have continuity in time, which is not conducive to data mining and subsequent processing. The database can effectively organize and manage the data to facilitate the efficient access and query of the data, so the data can be stored by constructing the database. The block diagram of database structure design is shown in Figure 3. R and Oracle database can be connected through the data interface. By calling the interface program, R can be embedded in the standard SQL database management language, so as to realize data writing, querying, deleting and other functions. Oracle is an interface toolkit between R and Oracle developed by Denis Muschin and others, allowing R to establish a connection with the Oracle database. Before establishing a connection, you first need to use the dbDriver ( ) function to load the driver. After loading the driver, use dB Connect ( ) to establish the connection between R and the Oracle database. The dB Connect ( ) function has four main parameters: drv is the loaded driver; username is the login user name of the database; password is the login password; dbname is the name of the target database. The target database name consists of the database IP, port number, and database name. The general form is IP: port number / database name.

5. Conclusion
Marine disasters are the natural disasters that have the most severe impact on the natural ecology and social economy of developed coastal regions in China. The changes in the marine environment caused
by global climate change are constantly increasing the risk of marine disasters. Therefore, on the basis of in-depth research and understanding of the key and basic scientific issues of marine disaster prevention and control in the context of global climate change, it is urgent to establish a forward-looking, systematic, holistic and coordinated marine disaster risk management system, gradually build a marine disaster prevention system, strengthen the public service capacity of the marine field to cope with global climate change and the ecological environment, resources and social security, and continue to promote the construction of marine ecological civilization and the modernization of national marine governance.

References
[1] Takaesu, M. Horikawa, H., Sueki, K., Takahashi, N., Sonoda, A., & Tsuboi, S. Development of a web application system for seismic waveform data observed in real time with the seafloor seismic network donet. Marine Technology Society journal, 50 (3) (2016) 87-91.
[2] Choi, J. Kim, K., & Jun, S. C. A computational study on sensor localization for underwater acoustic sensor network in the east sea environmental model. Journal of Marine Science & Technology, 24 (6) (2016) 1146-1154.
[3] Wakamatsu, H. & Miyata, T. Do radioactive spills from the fukushima disaster have any influence on the japanese seafood market. Marine Resource Economics, 31 (1) (2016) 27-45.
[4] Liu, X. Shi, W., Yuan, S., Xu, N., Chen, W., & Wu, J., et al. Integrated management strategies analysis of marine disaster risk in china. International Journal of Offshore and Polar Engineering, 26 (2) (2016) 192-198.