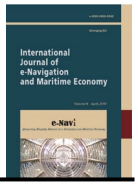
e-Navi

Available online at http://www.e-navi.kr/
e-Navigation Journal



Original article

# Application on Electronic Chart Data Management based on Data Mining Technology

*Haoran Song*

MaritimeCollege ,Shandong jiaotong university ,wei hai ,shandong,china, 264200,861328453@QQ.com

## Abstract

After decades of vigorous development, data mining technology has achieved fruitful theoretical and application results. As a highly applicable subject, data mining technology has penetrated into various fields of the national economy, and has aroused great attention from academia and industry. A large amount of chart data is stored in the electronic chart database, and its application is very extensive, providing a valuable decision basis for managers in all walks of life. It is of great significance to establish a complete data management mechanism based on data mining technology.

The traditional data analogy extraction technology, because of the data association index and the poor ability of data association, leads to the difference between the extraction data and the target data. Therefore, the application of data mining technology on electronic chart data management is studied. Data mining technology uses rough set to obtain the basic information of electronic chart data management according to similarity function, mining electronic chart data management association rules; through the comprehensive evaluanon data system of electronic chart data management, building rule base, setting up the evaluation index of electronic chart data management, achieving the similarity evaluation of the mining results. Experimental test results: compared with the traditional data analogy extraction technology, the results obtained by data mining technology have higher similarity with the target data and meet the requirements of electronic chart data management acquisition. It can be seen that this technology is more suitable for the application of electronic chart data management

*Keywords- Electronic chart; stratified network; data mining; association rules*


https://doi.org/10.52820/j.enavi.2020.14.068

## 1. Introduction

The core idea of the multi-level grid of spatial information is to divide grids of different thickness according to the size of grids with different latitude and longitude. Each level of grid has an up and down coverage relationship in scope. Each grid determines its position according to the latitude and longitude coordinates of its center point, and records basic data closely related to the grid.

The information representation, data storage and fast query technology of aeronautical information multi-level grid are drawn from the traditional GIS spatial data representation, data organization and query technology, and the information representation, data storage and fast query of the spatial information multi-level grid are studied technology. The multi-level spatial information grid can only be used as a supplement or replacement product for existing digital maps, charts, and aeronautical databases only by solving a series of technical key points, and will be truly applied. At present, the main use of maps is to store vector chart electronic charts to provide users with accurate chart information, so that the crew can perform various operations on the chart and query navigation data. The biggest problem of this electronic chart is that the chart data management is decentralized, there is no unified management mechanism, and the data redundancy is serious. In view of this, it was suggested that an electronic chart database has been established to manage and maintain chart data in a unified manner.

From the perspective of development in the field of navigation, the development of charts is an inevitable trend in the development of electronic charts. Put all vector data and information in a unified electronic chart database. For future centralized data management and update, the main idea is to establish a global electronic chart database through regional selection to complete the electronic chart data for different purposes . Due to the huge amount of global electronic chart data, fast and accurate retrieval based on the unique spatial data structure of electronic charts is particularly important. And the emergence of data mining is a good way to solve this problem. It has the characteristics of practicality of expert system, and because it is the foundation of data analysis, it can objectively excavate knowledge in data set.

## 2. Standard Data Structure Analysis of Electronic Chart

### *2.1.International Standards for the Application of the Electronic Chart*

International ships using ENC and ECDIS must have a unified international standard, therefore, the International Maritime Organization (IMO), International Hydrographic Organization (IHO) and the International Electrotechnical Commission (IEC) and other international organizations determine the unified international standard.

#### *2.1.1 IMO Performance standard*

ECDIS must comply with the ECDIS performance standard (EPS). It has been officially adopted by the International Maritime Organization in November 1995, and the IMO resolution A19 / RES817 is issued. The performance standard allows all the maritime safety administration to make electronic chart equivalent to the V / 20 rule of 1974 SOLAS convention. In other words, ECDIS (with ENC) can replace the traditional paper chart.

#### *2.1.2 IHO Data Exchange, Transmission Standard and Data Display Specification*

With the development of the IMO standard ECDIS performance, the International Hydrographic Organization has developed technical standards for digital chart formats and specifications, electronic chart content and display specifications, IHO special publication S-52 ECDIS content and chart display specifications, including Appendices, instructions, update methods, colors and symbols, IHO introduced the fifth edition of S-52 as a standard in November 1996. IHO special issue S-57 is a standard for converting IHO digital hydrological data, which includes target catalogue, mathematical model, data structure, ENC product specifications and ENC update mechanism, etc. IHO S-57 has been republished and revised many times, published in November 1996 The third edition is the current standard. This standard is a data exchange and transmissionstandard for legally effective vector electronic navigation maps

#### *2.1.3 IEC Test Standard*

At the request of the International Maritime Organization, the International Electrotechnical Commission/International Electronic Maritime Commission (IEC) Technical Committee of the

eightieth / seventh working group (IECTC80/WG7) ,in accordance with the prescribed regulations and instructions, as well as IMO/IHO ECDIS specifications, draf up necessary standards for ECDIS related equipment used for performance testing and inspection work. In December 1996, the draft standard ECDIS (IEC publication 1174) was released. In order to make necessary and appropriate revisions ,the revised version published in April 1998, the revised version provides performance requirements, test methods and required test results for electronic synthesis equipment.

*2.2 Comparison of Data Structure of Electronic Chart*

The electronic charts currently used can be divided into two types, raster and vector, according to the data structure of their spatial information.

The Raster electronic charts use bit maps to store and display the electronic chart, that is, display 1, 2, 4, 8, 16, or 24-bit pixels on the screen, ECDlS displays electronic charts in full color, so each pixel requires 24-bit storage, for example, the display resolution is 640*480, then the electronic chart on screen needs 640*480*24 / 8=921600 byte. Raster-type electronic charts can scan paper charts with a scanner, and the library can be established through proper processing (no pattern recognition required).

A vector in vector electronic chart contains its starting point coordinates (X, Y) and its displacement and direction. Therefore, the basic elements, points, lines and planes of graphics can be described by vector. Each point is given a coordinate, and two points are connected into one line segment. More than three points are connected into a polygon fold line. More than three lines are connected to a surface or polygon. As long as the density of the coordinates is large enough, one or a series of coordinates can describe the position and shape of the chart entity. In the chart legend, points such as beacons, buoys, etc. to the point of measurement points can represent vector coastline, in line contour vector representation, islands, shoals, obstacle area, anchorage could be impacted by polygon vector representation.

*2.3.Research on Data Structure of Electronic Chart*

There are two necessary conditions to realize the international standardization of electronic charts. One is that the data is produced according to the S-57 standard, and the other is that it is displayed according to the S-52 standard. The S-57 standard is a data transmission standard. The data model specified by it does not contain any rules for the graphical display of information. It only provides a method for describing the real world. The display of information is considered to be independent of its storage.

The S-52 standard is a standard for the production and application of electronic charts, especially for the display of electronic charts. The S-57 standard states that “different applications must provide their own specific display modes, and display the real world according to the purpose of special applications through a set of display rules.” The S-52 standard is based on this principle and states that “as maritime navigation The ECDIS display mode used must fully take into account the performance requirements of IMO/IHO regarding ECDIS", which requires that the performance of the chart on the ECDIS screen must depend on the navigation parameters and display options set by the navigator, taking into account the time period of navigation. Therefore, its display mode is a dynamic display mode adapted to the real world. Not only can the charts be displayed according to the default colors and symbols, but also according to user requirements and navigation status, the display effect can be dynamically changed by commands to meet the needs of navigation.

The chart consists of two data files: one is the spatial data file of the chart element, and the other is the attribute data file. In order to facilitate the retrieval of chart information based on type, hierarchy, coding and attributes, the file structure can be designed according to the chart spatial data and attribute data.

In our electronic chart system, in order to improve the response speed of the operation and reduce the size of the system database, so that it can be stored in a limited capacity device, we designed our own chart data format. In the electronic chart file format we designed, the spatial data file of each chart element is composed of three parts: title area, index area and data area.

The attribute file is composed of four parts: system header area, attribute definition area, index

area and data area. The attribute structure is defined by the attribute definition table, and the spatial data of the map element is connected to the attribute data through the element's serial number. Electronic chart data is derived from electronic chart data files in shapefile format.

Electronic chart files attempt to add units, directory structure, and directory names to digital chart numbers, which are composed of identifiers, and are composed of four files: control chart files, graphic files, index files, and attribute files. Each layer of elements in the digital chart corresponds to its own graphic file, index file and attribute file.

(1) control chart file: contains some information such as name, number, map projection, datum latitude, scale, coordinate, coordinate, datum, publication date, Department, notice etc..

(2) The graphic file is a variable-length record file, which can be accessed directly. It records the coordinate position data of chart elements, and the index file record describes the offset of the graphic file record corresponding to the starting point of the graphic file.

(3) The index file contains a 100-byte file header, followed by an 8-byte fixed-length record. The index file header is consistent with the graphic file header organization. The length of the file stored in the header is equal to the total length of the index file.

(4) an attribute file includes the attributes of the described element or the property keys that can be associated with other tables. It is a standard DBF file that can be used by many table applications based on Windows and Dos. The specific requirements are as follows:

- The attribute data file must have the same prefix and graph data file, index data file, the suffix is.Dbf;
- every graphical data record must have a corresponding attribute data record;
- Every graphic data records the attributes in the data file so that the graphic and the data file are consistent.

## 3. Methods and Techniques of Data Mining in Electronic Chart

Data mining, also known as knowledge discovery in database (KDD), is a hot topic in the field of artificial intelligence and database research. The so-called data mining refers to revealing hidden, previously unknown from a large amount of data in the database And there is a non-trivial process of potentially valuable information. Data mining is a decision support process, which is mainly based on artificial intelligence, machine learning, pattern recognition, statistics, database, visualization technology, etc., highly automated analysis of electronic chart data, make inductive reasoning, and mine potential Model to help decision makers adjust electronic chart data strategies, reduce risks, and make correct decisions.

Data mining (Data Mining) is to extract from the large amount of incomplete, noisy, fuzzy, random practical application data, hidden in it, people do not know in advance, but it is potentially useful information and The process of knowledge. This definition includes several layers of meaning: the data source must be real, large, and noisy; the knowledge found is of interest to the user; the knowledge discovered must be acceptable, understandable, and applicable; it is not required to be found everywhere Common knowledge only supports specific problem discovery.

Common methods for data analysis using data mining include classification, regression analysis, clustering, association rules, features, change and deviation analysis, Web page mining, etc., and they mine data from different perspectives.

Genetic algorithms are a type of randomized search methods that borrow from the evolutionary laws of the biological world (survival of the fittest, survival of the fittest). It was first proposed by Professor J. Holland of the United States in 1975. Its main feature is that it directly operates on structural objects, there is no derivation and function continuity limitation; it has inherent implicit parallelism and better global optimization capabilities ; The use of probabilistic optimization methods can automatically obtain and guide the optimized search space, adaptively adjust the search direction, without the need for certain rules.Genetic algorithm can produce a group of excellent offspring. These offspring strive to meet the adaptability. After several generations of inheritance, we will get the solution to the problem. As a new global optimization search algorithm, genetic algorithm is widely applied in various fields because of its simplicity, universality,

robustness, parallelism, high efficiency and practicality. It has achieved good results and is one of the most important data mining methods.

What technology the data mining system uses depends mainly on the type of the problem and the type and size of the data. For example, the cluster analysis method can be used to analyze the electronic chart. Using the association rule analysis method, the relevant information in ECDIS can be analyzed and excavated, and the rules can be found through the correlation between data. Using the decision tree method, GIS can be established by on-line monitoring, so that the staff on the ship can handle the sea depth data and the GIS data on the shore in time. In the data mining of electronic chart, the use of each method has its own unique side. Usually, the more technology we use, the higher the accuracy of the results.

*3.1 Discrete Data Mining Algorithm for Electronic Chart*

On the basis of distance based outlier detection, Sridhar Ramaswamy put forward a large and efficient outlier detection algorithm for large data sets: KNN.

The core idea of the KNN algorithm is that if most of the K nearest neighbor samples in a feature space belong to a certain category, the sample also belongs to this category and has the characteristics of the samples in this category. This method determines the classification of the samples to be classified according to the classification of the nearest sample or samples in determining the classification decision. The KNN method is only related to a very small number of adjacent samples when making category decisions. Because the KNN method mainly depends on the limited neighboring samples around, rather than the method of discriminating the class domain to determine the category, the KNN method is more than other methods for the sample set to be divided or overlapped

KNN algorithm can be used not only for classification, but also for regression. By finding the K nearest neighbors of a sample, and assigning the average value of the attributes of these neighbors to the sample, the attributes of the sample can be obtained. A more useful method is to give different weights to the influence of neighbors at different distances on the sample, such as the weight is inversely proportional to the distance.

On the basis of studying and studying two efficient KNN based outlier detection algorithms, we propose a KNN outlier detection algorithm based on two clustering.

*3.2 Algorithm Thought*

Outlier mining algorithm based on two time clustering is based on the previous algorithm, and extract their advantages, and improve the algorithm's shortcomings. Through in-depth analysis of the previous KNN algorithm, we finally get the following algorithm ideas.

The first step is to gather data into several classes, which are intended for subsequent calculations to be carried out in each class instead of the entire large dataset.

The second step: in each cluster:

To use Partition Based algorithm to calculate the K of this step for the same thought based on the division, but the clustering algorithm is not the same, because the data set has changed, so we can deal with any type of data and can eliminate the abnormal data K algorithm.

To determine the candidate division. By comparing the maximum boundary P.upper and minDkDist of each partition P, if P.upper is less than minDkDist, then this partition can not include outliers and remove directly.

*3.3Algorithmic Description*

Algorithm flow of K nearest neighbor classification algorithm

(1). Prepare the data and preprocess the data.

(2). Choose an appropriate data structure to store the training data and test tuples.

(3). Set the parameter, such as K.

(4). Maintain a priority queue of size K from large to small for storing nearest neighbor training tuples. Randomly select K tuples from the training tuples as the initial nearest neighbor tuples, calculate the distance between the test tuples and the K tuples respectively, and store the training tuple label and distance into the priority queue.

(5). Traverse the training tuple set, calculate the distance between the current training tuple and the test tuple, and get the obtained distance T from the maximum distance Lmax in the priority queue.

(6). Compare. If T>=Tmax, the tuple is discarded and the next tuple is traversed. If T<Tmax, delete the tuple with the largest distance in the priority queue, and store the current training tuple into the priority queue.

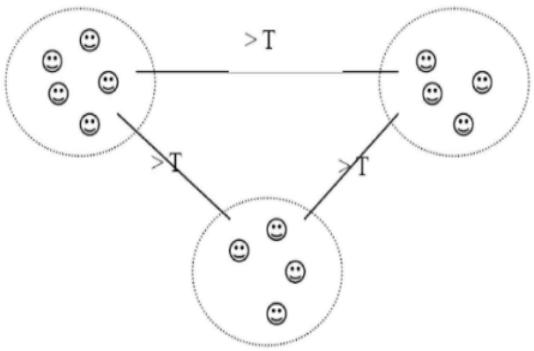
(7). After the traversal is completed, the majority class of K tuples in the priority queue is calculated and used as the category of the test tuple.

(8). Calculate the error rate after the test tuple set test, continue to set different K values for retraining, and finally take the K value with the smallest error rate.

The purpose of the first stage of clustering is to generate classes and calculate K in each class. Then, the simplest clustering algorithm can be selected here, without the high time complexity of the precise algorithm, because only the most recent in the K class Neighbors can be selected

Features: there is no need to determine the number of classes in advance. It is suitable for small intra class distance and large inter class distance, as shown in Figure 1.

Figure 1.The simplest clustering algorithm effect diagram




Here, our threshold T should be as large as possible to ensure that the K nearest neighbor of the point P is in the generated class. The class generated by this step is not directly entered into the next step, but when the class is generated, the number of data in each class is counted.

One: if all the data of clustering are larger than k, it will directly enter the second stage of clustering. 95% is the case, because our T value is very large. For some very special data sets, there may be the following situations.

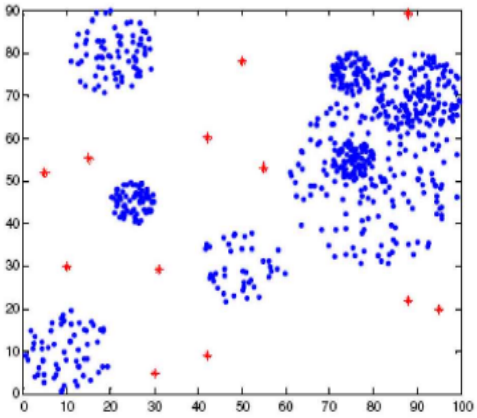
Two: if the amount of data in the class G is less than k, then the G is classified as the nearest class and then into the second phase of the clustering.

*3.4Experimental Results of Algorithm*

In order to test the effectiveness and extensibility of the KNN outlier data mining algorithm based on the two clustering, we have done a series of experiments. The experimental environment is: Intel (R) Core (TM) 2 Duo CPU 2.40GHz, 2GB RAM in Win7 environment.

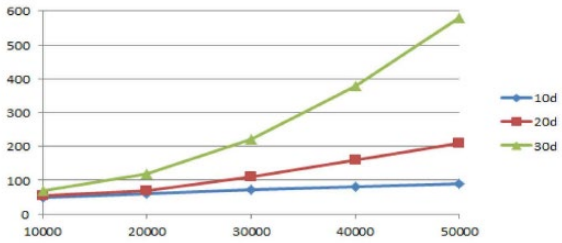
First, we do experiments in a two-dimensional synthetic dataset, using Yang Peng based clustering algorithm and our algorithm based on two times lustering. We set up a neighbor number k to 10, and we need to find out the number of outliers, N, set up to 12, and accurately find out the 12 outliers, as shown in Figure 2.

Figure 2 Outlier detection results on two bit data sets




In order to verify the efficiency and scalability of the algorithm for D, we use an analog data set, which produces the same probability of data on this dataset. The size of the data amount N is from 10000 to 50000, and the dimension is from 10 to 30. At the same time, we also give the execution time under the KNN algorithm based on a single cluster, as shown in Figure 3.

Figure 3 The experimental results of the extensibility of the two dimensional clustering algorithm for the dimension D




In order to verify this algorithm, we have done experiments on real data sets and integrated data sets, and the results show that our algorithm is accurate, efficient and scalable.

## 4. Conclusions

We use data warehouse and data mining technology to mine electronic chart system, and achieve the functions of electronic chart file management, no boundary splicing, automatic map changing, smooth

roaming, stepless scaling and so on, which improves the overall performance of the system. It can achieve the effect of efficient management and decision making, and it has a broad application prospect.

**References**

Wang Shilin , *guide to use of electronic chart display and information system,* 2013, first edition, Dalian Maritime University press

Qiu heart, Li Yong, *ECDIS's impact on shipping related industries, marine science and technology dynamics,* 2012, third edition, Dalian Maritime University press

Zhang Yingjun, *the mathematical and algorithm basis of the electronic chart,* 2011, first edition, Dalian Maritime University press

Shao Feng, Wang Jinlong, and Sun Rencheng. *Data mining principles and algorithms* [M]. Science Press, 2016.

Zhang Bin. *Application of Unbalanced Data Mining in Distributed Database*[J]. Control Engineering, 2018,25(7): 1179-1183

Liu Zhu.*Exploration on the Application of Big Data Mining in Engineering Project Management*[J].Engineering Technology Research, 2019, 4(19): 162-163.

**Received 18 May 2020**

**Revised 26 June 2020**

**Accepted 27 June 2020**