International Journal of Database Theory and Application
Vol.8, No.4 (2015), pp.19-30
http://dx.doi.org/10.14257/ijdta.2015.8.4.03

The Research of Data Mining in Traffic Flow Data

XuLuhang

Shandong University
xijiesd@126.com

Abstract

Data mining technology is one of the fastest growing areas of the information age of the 21st century, many scholars in different fields , such as database experts , statisticians , have gained a new breakthrough and development , which makes data mining increasingly become hot topic. Intelligent transport, traffic flow data analysis is very important, then how quickly and intelligently analyze traffic data flow has been a problem. Firstly, the data necessary preprocessing, making the data in the form of data mining algorithms can be used directly, followed by K-means algorithm for clustering processing toll stations , and clustering algorithm for clustering results were verification , and finally the use of Microsoft time Series algorithm to predict traffic flow data , the final results can clearly see the exact rate reached a staggering fifty percent or more.

Keywords: *Traffic flow data; data mining; clustering; time series*

1. Introduction

With the development of Internet and computer technology, access to data has been very easy, but for the large amount of data, involves a broader data, according to a summary of the past that simple statistical methods to analyze the specified mode of analysis of such data cannot be completed . So A comprehensive utilization of a variety of statistical analysis, intelligent , smart language database to analyze very large data technology on the consequent , which is currently the international statistical hottest topic " data mining "The market demand for technology and its technical background support. One area of data mining technology is very broad range of applications , including traffic data stream is also its applications in recent years , along with the country's economic development and the ongoing construction of the city 's rapid modernization forward , highway construction and tunnel construction has made leaps and bounds. In addition, in the transport sector there are also a lot of traffic data stream , then the data mining[1] technology in the field of highway tunnel traffic areas and provides a powerful data support. But at the same time , the complexity of traffic data and real-time data mining technology development also presents formidable challenges.

The introduction of new data mining techniques to replace the traditional methods of data interpretation and analysis is very meaningful and necessary for randomness transportation systems and traffic information , based on traditional knowledge , database[2], and model base of decision support systems on the use of on-line analysis of the theory and techniques of data warehouse , data mining and expert system , a method of forming (clustering , classification algorithm , time series algorithms , decision tree algorithms, neural networks and innovative data analysis systems, applications of data mining) , study the establishment of a specific model for traffic information mining algorithms , and then deal with the traffic flow data. Data flow information includes a variety of sensors (light intensity detector , CO / VI detectors, vehicle loop detectors , wind direction and speed detectors, etc.) to dynamically collect information, which also includes a traffic, speed , lane share other information. However, in the past, these large amounts of data has not been effective development and utilization , further in-depth processing, in which the relationship between the potential variation of traffic flow , as well as between each detector data is often ignored, or did not cause Note that the data which resulted in a waste of resources. In order to more quickly process vast amounts of data traffic flow, derive useful information, it is necessary to choose a more efficient method of knowledge discovery, data mining technology is a good method for knowledge discovery, data mining techniques to provide mankind with a species found an effective way to rule, knowledge and information, through data mining , information can be found in the direction of development and future existence of data, scientific guidance to provide a theoretical basis for the relevant departments. Now, with the rapid development of data mining in a variety of fields, it was discovered that the way knowledge is constantly changing. Currently, the traffic flow data is very large, but large amounts of data is considered to be "data rich, but information poor ", the rapid growth of a large number of traffic data usually exists in the database. So how do these large amounts of data through data mining techniques to find useful information and how to find the linkages is very important and urgent issues between these data, however, application of data mining technology in the transport sector, will promote the future of high-speed development of highways .So data mining technology in the traffic flow data[3] becomes a very meaningful and very practical work, so along with the rapid development of data mining, data mining programs science effectively applied to traffic data which will be able to efficiently and scientific services in the transport sector.

2. Related Works

In recent years, As information technology continues to rapid development of science and economy have made tremendous progress, on the other hand, in different areas also produced vast amounts of data , for example, banks with large daily transaction data and human of large amounts of data generated by space exploration . Obviously, these data are included in a large and effective information, then how can the service of humanity find confidence in these data, people have a useful exploration. With the rapid development of computer technology, thus promoting the great development of database technology, but the human face of the growing mass of data , people are no longer content merely to review the performance of the database, further proposed a deeper problem , which can be from data to extract useful information and knowledge to make decisions for the progress of a service. However, the database program, its already seem weak, the same traditional statistical techniques is also facing unprecedented challenges. That there is an urgent need for innovative ways to deal with these massive data like. Eventually, people combine databases, statistics and machine learning [4], data mining techniques proposed to solve this problem.

Data mining technology makes database technology has been improved and new development, it can not only past data search and traversal, and be able to identify the intrinsic relationship between the individual data, thereby promoting interaction between the data. Three basic techniques of data mining has now developed a very mature, namely: data mining algorithms, massive data traversal and powerful multi- processor computers.

Data mining technology is a combination of interdisciplinary research field, which combines artificial intelligence [5], machine learning, database technology, knowledge engineering, statistics, information retrieval, object-oriented research methods, data visualization and the latest technology, high-performance computing, after years of exploration, resulting in a number of new methods and theories, database development and information technology, the Internet, the evolution of computers to improve performance and advanced architecture, artificial intelligence and statistical methods in data analysis. Development of data mining technology is very promising, it is known as one of the ten key technologies affect the future and now at home and abroad main development of data mining technology is mainly around the following questions:

1. A particular type of data storage and data mining technology adaptation ;
2 .Specific business logic and data mining techniques smooth integration issues ;
3.Interactive mining techniques and data mining system architecture development ;
4 .Large data standardization and selection ;
5 .Data mining theory and algorithms;
6 .Visualization of data mining systems and languages.

Today's data mining techniques are used in clustering, classification and prediction that the existing data generated by the corresponding rules and other clustering or classification data mining method, followed by these rules to predict the development of the data, obtained for knowledge of human services.

Data mining functions are divided into clustering, classification [6], association rules and sequential patterns, deviation detection and prediction, etc., data mining function is complete, despite the strong performance of data mining, but also faced some unexpected challenges, this means for the future development of data mining provides a wider space.

You can see that data mining technology is only a means of service data for knowledge, it is not a panacea, it can be found in some of the future potential users , but does not tell the user the reason, we cannot guarantee that these users will become a reality. Data mining techniques to be successful it must require to solve the problem areas have a very deep understanding, understanding data, and understand the process, in order to ultimately arrive at a reasonable interpretation of the data mining results.

Data mining techniques of time series data mining algorithms is a very important program and widely used method, which is often used to predict traffic flow data. Time series data mining is a temporal data mining, data mining and other temporal data mining techniques in different categories, the data it deals with a time dimension of real data. Time series data mining is mining in time series analysis of a universal application.

Time series is an important and common form of data found in the time series of massive knowledge hidden behind the analysis of time series of changes in the way, when the right to make decisions is very important significance. Therefore, data mining techniques proposed shortly after, there are many scholars thought to use data mining to time series analysis. Sequential pattern mining is mining sequence is very important part, R.Agrawal first proposed mining sequence patterns, which are then put forward AprioriAll and AprioriSome algorithms.

The results of data mining technology in the transportation system is mainly to predict and deal with traffic flow data, the future direction of the analysis of the data obtained by the data, the development trend forecasting traffic flow and use the predicted traffic guidance. In addition you can also contact by studying different factors affect the corresponding relationship between traffic flow data, obtained the degree of influence of different factors on the traffic flow, and use the results of these studies properly instruct transport services for the transportation system.

Researches on dynamic traffic flow forecasting method is still in the preliminary stage, there is no mature theoretical prediction. Establishment of key aspect of this model is similar to working with a classification of historical trends. If the real-time traffic data is not collected or the detector does not have reliability data, then the historical trend model may be the only option. Despite the historical trend model can be solved within a certain range of different periods, the traffic flow at different times of change, but often do not have a static prediction science, because it does not solve the sudden change in traffic conditions, such as traffic accidents and so on.

Data mining techniques in many countries, intelligent transportation systems in many cities has been a very wide range of applications, now use data mining techniques to analyze traffic data, has been very common, so the data mining technology in traffic is promising.

3. Proposed Scheme

3.1 Data Mining

Data mining technology from massive, noisy, incomplete, randomized, simulated data to extract hidden in there, people do not know previously, but it has practical knowledge processes and information. As the economy continues to develop, it is the accumulation of data at an alarming rate, often with TB measure, how to extract data from a large number of data can be used is very much needed. Data mining technology is developed in response to this demand and data processing technology, is a key step in knowledge discovery.

Data mining technology is one of the high-tech field of rapid technological development, there are many scholars in different fields, as the database of experts, statisticians, etc., have been a number of new results in this field, which makes data mining techniques become more and more discussions the hot topic. With the rapid development of information technology, the method of data collection more people high and rich, thus becoming more and more data is accumulated, the data has reached the level of GB or TB and high-dimensional data has become more mainstream, these large amounts of data and its high-dimensional data analysis features make traditional means powerless. Increasingly powerful computer's performance, making it possible to help people expect computers to analyze and understand the data, helping to enrich the data we were able to make the correct reference guide, so the data mining technology that combines a variety of analytical tools, from the mass of data discover useful knowledge on the ensuing program and the rapid development in use.

The main task of data mining is to cluster analysis, correlation analysis, classification, prediction, error analysis and timing mode.

Association rule mining is proposed by RakeshApwal others come. Between two or more variables in a certain regularity, it is called association. The data is in the database associated with a class can be found, the existence of important knowledge. Association into causal association, temporal association and simple association. The purpose of the analysis is to identify the database associated with the closed network hidden. General credibility and support with two thresholds to measure the correlation between association rules, but also continue to introduce relevant, interest and other parameters, making the rules more in line with the requirements of mining .

Clustering [8] is the data summarized in the form of a plurality of similarity in accordance with the data the same as the same class of comparison, different types of data are quite different. Cluster analysis can establish macro concept, different relationships found to exist in the presence of the data pattern, as well as possible data attributes.

Classification is to find a description of the concept of a category, which represents the macroscopic information of such data, that the connotation of the class description and to use this description to construct the model, can generally be represented by a decision tree model and rules. Classification is obtained using a training data set classification rules through a certain algorithm. Classification can be used to predict and rule description.

Prediction is to use past data to identify changes in the laws and the establishment of the model and use this model for future features and types of data to predict. The most important is the forecast uncertainty and precision, generally predicted variance measure.

Timing mode refers to the probability of searching through the time series model is large and recurring. And regression Similarly, it is also known to use the data to predict future data, but the difference is only the data of the different variables in time.

There are many useful knowledge in the deviation, the data in the database, there are many unusual , abnormal conditions exist in the database data is very important. The basic method for checking the bias is to find differences between the reference and the observations.

In today's real-world database vulnerable to interference or noise data inconsistencies and other issues, because the database is too large, it is often up to several gigabit even larger. So how to preprocess the data and improve the quality of the data, thus improving the results of data mining becomes even more critical.

There are many methods of data preprocessing, data cleaning can remove the noise in the data, correct them inconsistent. Data integration of data from multiple sources into a consistent data storage, such as data cubes and data warehouses. Data conversion (e.g. normalization) may also be used. Such as standardization can improve data mining involves the effective distance metric and accuracy of the algorithm. Data statute by removing redundant features, aggregation or clustering methods to compress data. These data processing techniques in data mining before use, you can greatly improve the quality of data mining models, reducing the time required for the actual excavation.

Data mining method generally includes: data integration, data cleaning, data reduction and discrete data.

The existence of noisy, incomplete and inconsistent data is real-world database, a common feature of large or data warehouse. Appear incomplete data may be caused by a variety of causes. Some interesting properties, such as sales transaction data in the customer's information is not always available. Other data is not included, the input may simply be considered unimportant. Relevant data is not recorded due to equipment failure or due to misunderstanding, in addition to modify data or record history may be ignored. Inconsistent data and other data can be deleted. Missing data, in particular the lack of certain elements characteristic value may need to push out.

Noisy data (with incorrect property values) may have a variety of reasons. Data collection device may be faulty; or human error that may occur when a computer data entry; data transmission errors may occur. These may be due to technical limitations, such as for synchronous data transfer limit the size of the buffer. Incorrect data may also be used by the data code or by naming inconsistencies caused. Duplicate tuples also need to clean up the data.

3.2 Clustering Algorithm

The clustering algorithm is a collection of abstract or physical grouping of objects into a plurality of classes by the process similar objects. Generated by the clustering of the cluster is a collection of a set of data objects, these objects and objects in the same cluster are similar to each other , with the other objects in the cluster is not the same .

Cluster analysis can be used as a separate tool to obtain the distribution of the form data , the characteristics of each cluster was observed , the concentration of certain specific cluster further analysis , in addition , cluster analysis can be used as other algorithms (such as classification and characteristics preprocessing step) preprocessing algorithm , that is, before the implementation of other algorithms to use clustering algorithm to find some hidden relationships , you can also make use of the appropriate treatment of these algorithms on the generated clusters. Clustering properties acquired directly affects the results of the analysis, data mining algorithms for clustering feature requirements are as follows:

1. Constraint-based clustering: in practice may need to be based on different constraints clustering, it is necessary to find a good clustering properties of a data packet is a challenging task, but also found that satisfies certain conditions constraints.
2. Scalability : In many clustering scheme, the data objects on a small data set is often satisfied robustness , and for containing data on millions of clustering large-scale databases , may result in different deviations result This requires a high degree of clustering program scalability.
3. For the input order of insensitivity: Some clustering algorithms for data input sequence is very sensitive, for example, the same set of data , submitted in a

different order to the same algorithm, it is possible to produce a big gap clustering results.

4. Clusters with arbitrary shape can be found: the use of many clustering algorithms Manhattan distance or the Euclidian distance to determine the clusters, tend to find the size or density having a near spherical clusters, but a cluster may be any switch a. Thus, any switch can be found proposed algorithm clusters is important.
5. Anti-interference ability: In the vast majority of real-world applications contain isolated points, vacancies, unknown data or erroneous data. Therefore, clustering algorithm should be able to have this ability to resist noise data, otherwise the quality of clustering results cannot be ensured.
6. High -dimensional data processing: a database may contain a number of attributes or dimensions, many clustering algorithms adept at handling uni-dimensional or low- dimensional data, however, is difficult to obtain a low-dimensional clustering quality assurance. Typically in the case of multi-dimensional well judge the quality of clustering. Requiring clusteringalgorithm can handle high-dimensional data.

Clustering Algorithms:

K- averaging Algorithm

K- mean clustering algorithm is an iterative algorithm in an iterative process constantly moving objects in a cluster of clusters until you get the desired date, each cluster using the mean cluster object to represent. Use K average algorithm clusters, the cluster is very similar to the object, the object in different clusters are sometimes quite different. Then the process can be as follows:

1) 1) From the object data in the x k randomly selected objects as the initial cluster center;
2) Next, calculate the average value of each cluster and with this average represents the corresponding cluster;
3) Calculating the distance of each object and the center of the object and re- divide the corresponding object according to the principle of minimum distance;
4) Turning next to the second step, calculates the average value of each cluster again. This process will continue to be repeated until there is no longer one criterion function object or a significant change in the clustering of no significant changes so far:

   Typically, the criterion function algorithm uses k- average mean square error criterion , can be defined as :

$$E = \sum_{i=1}^{k} \sum_{q=c_i} \left| q - w_i \right|^2$$

Where, E is a data set of all objects with the corresponding cluster centers and the mean square error, q is the object of the given data, w_i is the mean of clustering c_i (q and w are multidimensional) .k- averaging algorithm for large databases is relatively efficient and scalable, the time complexity of the algorithm is O (ktn), where t is the number of iterations . Typically, at the end of the most solvable locally. However, k-average algorithm must be used in order to sense the situation at the average value obtained. For categorical variables are no longer practical, but also given the number of pre-generated clusters, when more sensitive to outliers and noise, will not be able to non-convex shape of the data accordingly.

k-center Point Algorithm

PAM ((Partitioning Around Medoids) algorithm may also be referred k- center algorithm, close to the center of each cluster can be used to represent an object . Object beginning of each cluster to be a representative of an arbitrary choice , the remaining according to its object to his latest one cluster and the representative object distances size distribution , then repeated with a non- representative objects instead represents the object, and ultimately improve its performance and the quality of clustering process can be represented as follows :

1) Select the k initial objects representative poly (center) of the n randomly from the object data in ;
2) Depending on the object represented by each cluster center, as well as various objects and distances between objects such centers, based on the minimum distance of the corresponding object is divided again.
3) Randomly selects one object Orandom non- center, calculating exchange to make the entire distance which the center of the object and the amount of change of the size of the exchange.
4) If the amount of change in the distance as a negative cost , then they would exchange Orandom and the center of the object , then the object k centers constitute a new cluster.
5) Next, a second turning step , there is a change to update each cluster center . This process is repeated continuously until the clustering object is no longer a criterion function change or no longer has significant changes. Among these, the criterion function can k- same average algorithm. When there are outliers in the data , or noise , k- center algorithm to be better than average k- algorithm, but the computational complexity of the algorithm k- center is much higher , so you cannot very well be applied to large databases to the top .

3.3 Time Series Algorithm

The so-called time-series data [9] is a chronological arrangement of the observed and recorded for each data set. Time sequence database refers to the database by the time-varying with time, or time series values . Trends of time series data mining problem is to determine given the two time series is similar in nature, or refers to a long sequence of time -series data source , find meaningful patterns edge timing data here.

Time series data mining is to find out from the mass of time-series data in which people do not know in advance, but it can be very useful. Knowledge and information related to the nature and time, and can be short, medium and long-term forecasts to guide the behavior of people in the military, social, economic and life.

From the mathematical aspects, if a variable $G(t)$ of a process to observe measurement data set $G_1, G_2, ..., G_n$ in a series of discrete moments $t_1, t_2, ..., t_n$ where you can get called discrete time series. Assumed $G(t)$ to be a random process, $G(t)(i=1,2,...,n)$ is called a sample in advance, also known as time series.

Mining step sequence typically can be expressed as follows:

1) Sorting stage appropriate for database sorting, sorting the results of sucked the beginning of the sequence data into the database (this method is often achieved by means of preprocessing).
2) Large projects set the stage, to find all the frequent item sets consisting of a collection. In fact, the whole is greater than the 1- to get synchronized sequence combination that is.

3) The conversion stage, when finding sequential patterns, we must continue to carry out large set of sequences to detect whether or not included in a given sequence among customers.
4) Sequence stage, the use of the database to find frequently transformed sequence, which is great sequence.
5) Select the biggest stage, the longest sequence found in the sequence resulting in the collection.

3.4 Data Mining in Traffic Flow Data

Refers to traffic flow at a particular point in time, the number of pedestrians through a road section or of a location, the vehicle, generally refers only to vehicle traffic. This is the change over time, often as a representative of the traffic peak hour traffic volume with an average traffic volume and design -hour traffic. Traffic, lane occupancy and traffic speed is three elements, these three elements of mutual restraint, mutual contact [10].

Assuming the free flow of traffic flow. N car length is performed on sections M, make speed is V, it can be defined based on these three parameters

Traffic density on the road for M is $\lambda = \frac{N}{M}$

N number of cars through a section of the time is $T = \frac{M}{V}$

N number of car traffic is defined by a cross-section of $Q = \frac{N}{T}$

You can get all kinds of finishing above

$$Q = \frac{N}{T} = \frac{N}{\frac{M}{V}}$$

$$Q = \frac{N}{T}V$$

$$Q = KV$$

Where : Q is the flow rate , volume / h
V is the speed range , km / h
λ is density , vehicle / km

The relationship between these three variables is the relationship between a three-dimensional , three -dimensional projection coordinate system can be expressed as the volume of traffic - density, speed - traffic volume and speed - the relationship between density.

4. The Experimental Results and Analysis

In this paper, because the collected raw data traffic obtained at the toll station, because the original data is not available for the final form of data mining, data mining, so before the initial data on the first proper cleaning, primarily data attribute, type, etc. into the type of data mining properties and necessary, so that through a number of conversion and calculation in order to get.

By observing the data, can be clearly seen in the collection of six months from the toll booths to the statistics of the various models in the amount of traffic information 1-6 months. For the time of data collection may encounter some problems here, so the data collected in the table for each month is relatively chaotic. So you want to do first is to get accurate each month, various models of traffic flow. Then, after each month in the oracle to get traffic in each table coupled together via a query to get the total traffic for the month. The results obtained are shown in Table 1.

Table 1. Toll Station Traffic Statistics for Each Vehicle in the First Half

Month	Bus traffic			
	Type1	Type2	Type3	Type4
January	25521	45254	52203	22042
February	22015	42036	78550	24563
March	14526	7883	11454	4521
April	9865	7895	12547	7862
May	9765	14562	12536	14562
June	14256	8532	9762	9658

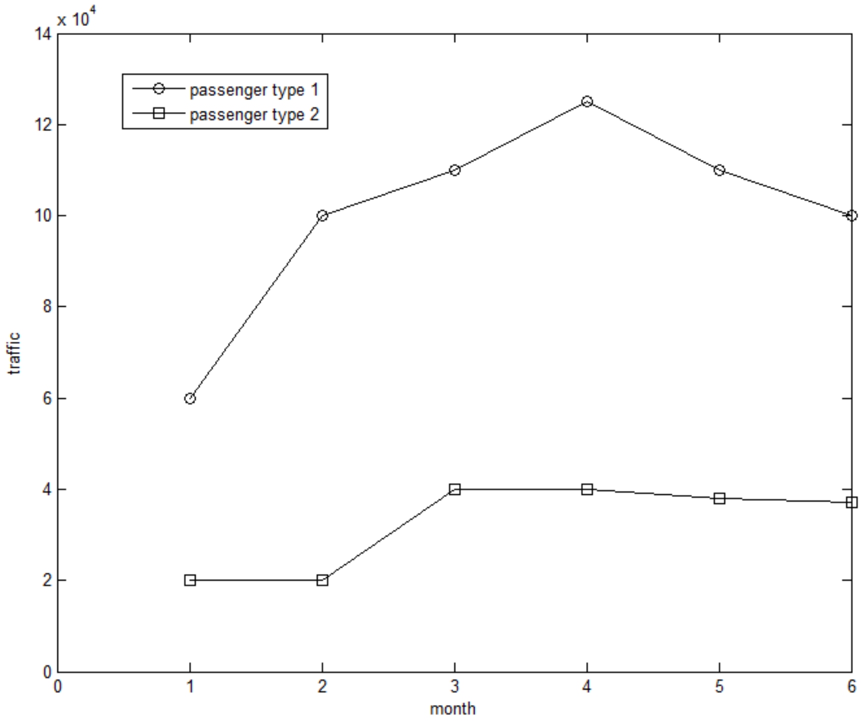

Figure 1. A Car each Month Traffic Graphs

As can be seen from Figure 1, no matter what type of car , their traffic trend in the first half are very identical, are rising from the beginning of January to reach around April began to slowly decline, so no matter What kind of car, traffic at the peak point of the first half will occur between March and April.

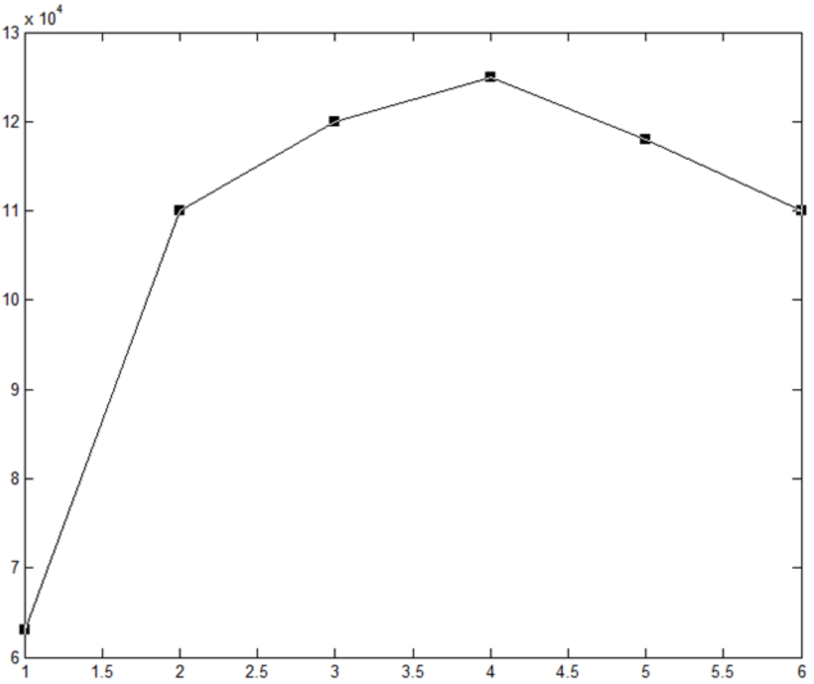

Figure 2. February Traffic Forecast Map

Of total passenger traffic five lanes of a certain type of prediction can be drawn in July passenger traffic for type 1 108 603 , as shown in Figure 2 , while the actual traffic in July was 112,046 , deviation 3443 , the correct rate of 97 %. Then select the number of steps will be even greater forecasting, you can find almost predicted curve is a straight line, so when used to predict six months of data, best not to forecast data for many months, and this will lead to larger deviations. So one day when the time is accurate to predict may be more inaccurate, because that is the predicted time may be too short, too sensitive to make changes in the curve generated due to interference prediction. However, when the exact same time, lane -to-day research, you can get a time series function model, which is extremely important.

5. Conclusion

This paper analyzes the data mining algorithms and time series clustering algorithm, and specifically describes the principle and application of the algorithm. How to time-series data mining algorithms and clustering algorithms applied to the problem of dealing with traffic flow. Through experiments , the algorithm can be seen for the traffic flow data mining is very useful , and can be certain prediction, of course, according to further aspect of the present paper some room for improvement, for example, a number of complex factors not fully taken into account, this is also the focus of future work needed research.

References

[1] D. T. Larose, “Discovering knowledge in data: an introduction to data mining[M]”, John Wiley & Sons, **(2014)**.

[2] T. K. Attwood, A. Coletta and Muirhead, “The PRINTS database: a fine-grained protein sequence annotation and analysis resource—its status in 2012[J]”, Database, **(2012)**.

[3] G. Chang, Y. Zhang and D. Yao, “Missing data imputation for traffic flow based on improved local least squares[J]”, Tsinghua Science and Technology, vol.17, no.3, **(2012)**, pp.304-309.

[4] A. Frank and A. Asuncion, UCI machine learning repository, 2010[J].URL http://archive.ics. uci. edu/ml, **(2011)**, pp.15: 22.

[5] M. N. Huhns, “Distributed artificial intelligence[M]”, Elsevier, **(2012)**.

[6] D. Brauckhoff, X. Dimitropoulos and A. Wagner, “Anomaly extraction in backbone networks using association rules[J]”, IEEE/ACM Transactions on Networking (TON), vol.20, no.6, **(2012)**, pp.788-1799.

[7] R. Zhu, Y. Zhao and Y. Li, “Optimal linear precoding for opportunistic spectrum sharing under arbitrary input distributions assumption”, EURASIP Journal on Advances in Signal Processing, **(2013)**, pp.59.

[8] A. Raftery, “Fast Inference for Model-Based Clustering of Networks Using an Approximate Case-Control Likelihood”, **(2011)**.

[9] A. S. Banks, “Cross-National Time-Series Data Archive, 1815-[2011][J],” **(2011)**.

[10] Y. Song and H. J. Miller, “Exploring traffic flow databases using space-time plots and data cubes”, Transportation, vol.39, no.2, **(2012)**, pp.215-234.

Author

Xu Luhang, Shandong University, undergraduate course grade three University studies: Computer science and technology. He is interested in intelligent algorithm. He is a student body President of the computer science and Technology College of Shandong University, and has been the first prize in national mathematical modeling competition in Shandong province.