Research on Data Mining Technology Based on Machine Learning Algorithm

Shangran Li

1 Case Western Reserve University, Cleveland, 44106-5911, United States of America

Abstract: In recent years, the domestic society has made great progress, and various advanced technologies and concepts have been applied and developed continuously. Machine learning algorithm is a new algorithm which plays a great role in all walks of life. Machine learning is a process of self-improvement using the system itself, and computer programs can automatically improve performance with the accumulation of experience. It is proposed for many specific learning tasks, so that computers can extract features from many data and discover hidden rules, so machine learning has been widely used in data mining. In this paper, the application of machine learning algorithm in data mining is studied in detail. With the help of mobile terminal data, the outdoor terminal of GSM network is positioned effectively, and a three-stage positioning algorithm is proposed, which greatly improves the positioning speed and accuracy.

1. Introduction

Machine learning and statistical algorithm are two common data mining algorithms. The first is to use artificial intelligence technology to automatically find the required patterns and parameters after training and learning many sample sets. The second is to use discriminant and probability analysis, clustering and correlation analysis to carry out operations. Different algorithms also have different corresponding goals and areas. These algorithms can be independent. They can be combined with each other for their own use. Artificial neural network method in machine learning algorithm has a wide range of applications, has a good ability to deal with data and self-organizing learning ability, but also can accurately identify, which is conducive to the classification of problem data processing. Modelling can be used to work. Models are more diverse, and different needs can be met. From the overall point of view, the model of this method has higher precision, better robustness, and stronger descriptive ability. It does not need the support of experts for application, but also has some shortcomings. It is necessary to train data. More time is spent, knowledge is not very intelligent to understand, and there are limitations of scalability and openness. Machine learning is a way to improve performance automatically by making use of accumulated data. The predecessor of machine learning method is statistical learning and optimization theory. It was born with the emergence of computer. Up to now, many algorithms have been proposed for different disciplines and different problems. The representative algorithms include Bayesian estimation, decision tree, neural network, support vector machine, k neighbour method and so on. It is an important way to solve data mining problems. Data mining is an interdisciplinary, application-oriented concept. If there is a large amount of data accumulation in industries and fields such as telecommunications, finance, retail and scientific research, there are problems and needs of data mining. This paper attempts to explore the data mining problem of outdoor mobile terminal location in the network.

2. Categories of data mining tasks
Through the application of machine learning, data mining methods for data processing can be divided into: classification, regression analysis, association rules and clustering, and each mining method can be implemented by different machine learning techniques.

Classification. The training data set is used for learning, and a classification model is obtained. Then, the classification model can automatically classify the data without category labels into multiple categories, thus completing the classification. Existing machine learning classification algorithms include KNN classification algorithm, naive Bayesian classification algorithm decision tree artificial neural network ANN and support vector machine. Although the KNN method also depends on the limit theorem in principle, it is only related to a small number of adjacent samples in class decision making. KNN method is more suitable than other methods for the intersected or overlapped sample sets because it mainly depends on the neighbouring samples, rather than on the method of discriminating class domains.

Regression analysis. By analyzing the data and applying statistical methods, the relationship expression between variables and variables can be obtained. These inherent laws are used to estimate and predict future trends. Regression models can be constructed by regression tree artificial neural network ANN linear regression and logic regression. Artificial neural network is inspired by the attempt to imitate the biological nervous system. The human brain is mainly composed of neurons, connected by axons. When a neuron receives a stimulus, nerve impulses pass from one neuron to another through the axon. A neuron connects to the axons of other neurons through a dendrite, an extension of the neuronal cell body. The junction between dendrites and an axon is called synapse.

Association rules. There are association rules between transactional data, and frequent itemsets can be obtained by mining the relationship between transactional data. Based on this, we predict the probability of simultaneous occurrence of certain transactions. Apriori is a classical algorithm for mining association rules. Apriori algorithm is the most influential algorithm for mining frequent item sets of Boolean association rules. Its core is recursive algorithm based on the idea of two stage frequency set. The association rules belong to single dimension, monolayer and Boolean association rules.

Clustering. By using the mining algorithm, several data objects without class labels are clustered in several different clusters, which makes the data objects in the cluster have a high degree of similarity with each other, and the data objects among the clusters are very different from each other. K-means is a classical clustering algorithm. In addition, artificial neural network ANN and support vector SVM can also realize clustering. K-means algorithm is a hard-clustering algorithm, is a typical prototype-based object function clustering method representative, it is a data point to the prototype of a certain distance as the optimization of the objective function, using the method of function extremum to get the adjustment rules of iterative operation. K-means algorithm takes Euclidean distance as similarity measure, which is to find the optimal classification corresponding to a certain initial clustering center vector to minimize the evaluation ind. The error square sum criterion function is used as a clustering criterion function.

3. Advantages of the application of machine learning algorithm in data mining
Machine learning is a field of study on computational methods in the learning process and how to apply computer-based learning systems to solve practical problems. An important research content in machine learning is the research on the method of getting the corresponding concept description from the sample. Therefore, many machine learning methods can be directly used to solve data mining problems. Data mining is the problem of finding interesting patterns and important rules from large databases. As most of the traditional machine learning algorithms are based on memory, and TB or even PB level massive data cannot be loaded into the computer memory. Therefore, many existing algorithms cannot handle large data. With the participation of computer, the analysis and processing of large-scale data is completed. The process of extracting meaningful rules or patterns from data is data mining. Machine learning is a way to improve performance automatically by making use of accumulated data. The predecessor of machine learning method is statistical learning and optimization
theory. It was born with the emergence of computer. So far, many machine learning methods have been proposed to adapt to different disciplines. How to make machine learning algorithm adapt to the requirements of large data mining has become the main direction of industry and academia. In large data environment, the design and implementation of machine learning algorithm involves many aspects, including distributed computing data stream technology cloud technology. Machine learning algorithm combines these technologies to process hundreds of millions of data objects efficiently and train the model quickly to acquire valuable knowledge. Machine learning technology has been widely used in enterprise data mining such as intelligent speech recognition search engine in recommendation system. The development of large data and the study of key technology evaluation indicators pose new challenges and requirements for the research of machine learning methods.

4. GSM network location based on machine learning

4.1 Problem description and solution process

Mobile terminal positioning technology refers to the use of various means to determine the location information of moving objects, such as the degree of obscurity and height. This technology has a long history. In ancient times, the transmission of military alarms by igniting a beacon fire could be regarded as an original means of mobile communications, and the naked eye observation, through the approximate location of the beacon to determine military operations. Moving location can be regarded as an original location method of mobile terminal. At present, GPS is the most effective outdoor mobile terminal positioning technology, but the use of this technology has many limitations, such as high-rise dense areas of positioning performance significantly reduced, high energy consumption, mobile terminals need to have the corresponding hardware support, high computational complexity and delay and other causes not all cases. The use of global positioning system is the best choice. The outdoor mobile terminal positioning method based on machine learning makes up for the deficiency of GPS when it cannot meet the positioning requirements to a certain extent. The outdoor mobile terminal positioning method based on machine learning does not need any special hardware devices added to mobile terminals or mobile communication networks, only needs to collect a certain number of positioning areas. The mobile terminal, which is used as a training unit, can receive the level information to achieve higher positioning accuracy, and the computational complexity is lower than that of GPS. But at present, machine learning-based localization methods still have much room to improve the localization accuracy and computational complexity, and cannot fully meet the requirements of real-time localization for all concurrent communications in a large area at a lower cost of computing equipment. This paper is devoted to improving the positioning method of outdoor mobile terminal based on machine learning based on existing achievements to greatly improve its accuracy and speed performance, so that it can better adapt to practical needs.

Aiming at the problem of mobile terminal positioning in outdoor network, a three-stage positioning method based on SVM and k neighbour method is proposed. The flow chart of the method is shown in the figure. Next, we will introduce the concrete implementation of each step.
4.2 Model establishment
Modelling process is mainly based on support vector machine positioning method, the positioning position rasterization, small area raster position is an independent category, in the positioning position, we collect many terminal measurement data, and then use the computer to analyse and process the measurement report, measurement raster the distance measurement and accuracy of the mobile terminal grid are then estimated and judged. Finally, machine learning is used to analyse and solve the problem. Support Vector Machine (SVM) based localization method is to abstract small raster regions into categories by rasterizing the localization regions, and then collect a large number of mobile terminals receiving level measurement reports in the localization region, and the longitude and latitude position of the mobile terminal when sending the measurement reports to form a training data set. In positioning, by calculating the similarity measure or distance measure between the receiving level measurement report received from the mobile terminal and the receiving level measurement report received from the training data set in the partitioning grid, the decision is made on the grid to which the mobile terminal belongs, and the positioning problem is transformed into a multi-class classification problem. The classifier solves the classification problem. For multi-classification problems when the sample points in the training set to be classified belong to more than two classes, it is necessary to extend the binary classification machine learning algorithm.

4.3 Data acquisition and pre-treatment
To ensure the validity of localization by machine learning method, three batches of data measured by line are regarded as training data, and the final data set is regarded as localization data. The centralized data of the first 3 groups of training data are deleted within 10 kilometres around the batch data. After getting the data to be positioned, we need to take the interval of different time as the basis, then merge the adjacent positioning data in the same call, and get the longitude and latitude average of the same base station receiving level and multiple adjacent positioning data in the same call. This value is regarded as a new location data, because there is a strong correlation in the measurement reports adjacent to the same call, so it needs to be merged and then positioned. This can denoise, and can
greatly reduce the amount of data needed for positioning, thereby effectively improving the positioning accuracy and speed. To ensure the validity of machine learning localization method, we selected three batches of data sets with earlier road test time as training data, and the fourth batch of data sets with later road test time as localization data. We deleted the first three batches of data sets which did not exist in the range of 10 kilometres. Finally, 2034192 pieces of training data and 1428942 pieces of data were obtained from the data collected in the road survey data set. In addition to the above road survey data, we also use the numbering information and the latitude and longitude information table of all the base stations in the square area where the road survey data is located. After some data pre-processing, we store each data in the data set in the form of txt file in the computer hard disk.

4.4 Preliminary positioning based on latitude and longitude of base station
Using machine learning algorithm to locate mobile terminals is also more complex. Once the area of the region increases, the model and classification will increase correspondingly, and more complex. Therefore, the process of using machine learning algorithm to locate mobile terminals will be more expensive with the area of the location area increasing. Time. Using the longitude and latitude of the base station as the basis for early positioning, we need the following steps. First, we use the longitude and latitude of the main service cell base station of the mobile terminal to preliminary positioning of the mobile terminal, the mobile terminal may be in a square area with a side length of 2 km, so in the machine science. The learning and positioning stage only need to locate in a narrow area with a length of 2 km, instead of in a square area with a length of 10 km, which greatly reduces the computational complexity of the machine learning and positioning stage. At the same time, the initial positioning method based on the longitude and latitude of the base station in the main service cell only needs a few steps. Computation of simple expressions is almost negligible compared with reduction of machine learning localization area, so the overall computational complexity of localization method training phase and localization phase is greatly reduced. When locating the data in the positioning data set, according to the longitude and latitude of the base station of the main service cell, the side length of the base station is 1 km square grid. Because the distance between the mobile terminal and the base station of the main service cell in the urban area is generally not more than 500 meters of the base station of the main service cell of the mobile terminal can be found. After the meter square grid, it can be considered that the mobile terminal must be in the square grid of 2 km cantered on the kilometre square grid. The formats of data set are shown in Table 1.

| Field         | longitude | latitude | n-cell  | TS             |
|---------------|-----------|----------|---------|----------------|
| Interpretation| GPS longitude | GPS latitude | Number of sectors | Road test time |
| Field         | LACC1-LACC7 | Relev1-Relev7 | ser_cell_lon | ser_cell_lat   |
| Interpretation| Unique number | Sector reception | Base station longitude of main service cell | Base station longitude of main service cell |

4.5 Secondary positioning based on SVM
After preliminary positioning, a square of 2 km side length is selected, because the range of the first stage SVM is 400 meters, and the second season SVM outputs the undetermined data of 100 meters grid. The output of the positioning result is the longitude and latitude of the 100 meters grid as the core. Compared with the localization of the first-order vector machine, the computation complexity of the second-order vector machine is relatively small. The calculation of the vector machine after classification is mainly the calculation of decision function and the calculation of all kinds of vector
machines of the sample points to be classified. Many classification problems are dealt with. The positioning accuracy will be higher as the grid becomes smaller, because the total number of classification problems will be increased, and the complexity of positioning will be increased. When the size of the grid is selected by the first-level vector machine, the total number of problems in the first-level and second-level classification will be minimized to be selected according to the size of the second-level grid. Ensure that the total amount is calculated in the minimum positioning stage. In practical application, the selection of the square grid size determines the final positioning accuracy. The smaller the grid is, the higher the positioning accuracy is. However, because the total number of classes in the classification problem will increase, the computational complexity of the positioning phase will be increased accordingly, and with the decrease of the grid edge length. The increase of positioning accuracy will be smaller and smaller, so the specific second-level support vector machine positioning is the square grid size selection should be integrated with the practical application of positioning accuracy and speed requirements, the first-level support vector machine positioning square grid size selection should be determined after the second-level grid size according to the first-level classification problem. The principle of minimizing the total number of classes is chosen to minimize the total amount of computation in the positioning stage. After dividing the kilometres area of SVM into two-level grids, each grid should be numbered as the classification marker. To label the training data conveniently, the direction of longitude and latitude should be given priority, and the direction of longitude and latitude should be in order from small to large.

We substitute the longitude and latitude of each data in different training sets to the following expressions.

\[
d = \frac{2 \pi R_{\text{earth}}}{360}
\]

\[
\text{relative } X = d(\text{lon} - \text{ref_lon}) \cdot \cos(\text{lat} + \text{ref_lat})
\]

\[
\text{relative } Y = d \cdot (\text{lat} - \text{ref_lat})
\]

In the above formulas, \(d\) represents the actual physical distance corresponding to a longitude or latitude; \(\text{relative } X\) and \(\text{relative } Y\) represent the physical distances between the training data and the boundary of the square positioning area in longitude and latitude directions; \(\text{lon}\) and \(\text{lat}\) are the positioning results based on SVM.

4.6 Tertiary positioning based on k-neighbour method

In the training phase of the positioning method based on the nearest neighbour method, it is necessary to merge the data from the training data set in the 10 km square area according to a certain size of the square area. The bigger the merged area is, the lower the final positioning accuracy will be. Therefore, in practical application, it should be selected according to the positioning accuracy and speed requirements. Appropriate merge area size. After merging the training data sets, each merged data is given priority by longitude or latitude of the merged area. In the positioning stage based on the nearest neighbour method, the square area selected in the training model based on the nearest neighbour method positioning training stage is first selected when the positioning area based on the nearest neighbour method is determined. Because the combined training data in the domain are numbered in the order of longitude first or latitude first in the merged region and both longitude and latitude are numbered from small to large, the formula-to-formula can be used when numbering in the order of dimension first, longitude and latitude from small to large. The centre longitude and latitude of the merged region can be easily calculated by using the longitude and latitude of the centre point and the side length of the region based on the nearest neighbour method. Only the merged training data of the centre longitude and latitude of the merged region in the positioning model can be extracted. Then some similarity feature or distance measure is calculated between the merged data and the selected merged training data in the second localization based on support vector machine. Then the merged training data which is most like or nearest to the data to be localized is selected and the average longitude and latitude corresponding to the merged training data is output. That is the final location
result of the three-location based on the nearest neighbour method. The nearest neighbour method should be selected according to the actual simulation results. Generally, there will be an optimal value to maximize the positioning accuracy.

5. Conclusions
In this paper, a high-performance solution based on machine learning is proposed for mobile terminal positioning in outdoor network. Compared with traditional methods, it has higher accuracy and lower computational complexity. The main conclusions are as follows:

(1) The categories of data mining tasks mainly include classification, regression analysis, association rules and clustering.

(2) The process of GSM network localization based on machine learning mainly include five steps of Model establishment, Data acquisition and pre-treatment, preliminary positioning based on latitude and longitude of base station, secondary positioning based on SVM and tertiary positioning based on k-neighbour method.

(4) The practice proves that the positioning method based on machine learning greatly improves the positioning accuracy and reduces the time complexity.

References
[1] Kavakiotis I, Tsave O, Salifoglou A, et al. Machine learning and data mining methods in diabetes research[J]. Computational and structural biotechnology journal, 2017, 15: 104-116.
[2] Yosipof A, Nahum O E, Anderson A Y, et al. Data Mining and Machine Learning Tools for Combinatorial Material Science of All - Oxide Photovoltaic Cells[J]. Molecular informatics, 2015, 34(6 - 7): 367-379.
[3] Papamitsiou Z, Economides A A. Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence[J]. Journal of Educational Technology & Society, 2014, 17(4).
[4] D’Oca S, Hong T. Occupancy schedules learning process through a data mining framework[J]. Energy and Buildings, 2015, 88: 395-408.
[5] Voyant C, Notton G, Kalogirou S, et al. Machine learning methods for solar radiation forecasting: A review[J]. Renewable Energy, 2017, 105: 569-582.
[6] Joachims T, Granka L, Pan B, et al. Accurately interpreting clickthrough data as implicit feedback[C]//ACM SIGIR Forum. Acm, 2017, 51(1): 4-11.
[7] Bijalwan V, Kumar V, Kumari P, et al. KNN based machine learning approach for text and document mining[J]. International Journal of Database Theory and Application, 2014, 7(1): 61-70.
[8] Shouval R, Bondi O, Mishan H, et al. Application of machine learning algorithms for clinical predictive modeling: a data-mining approach in SCT[J]. Bone marrow transplantation, 2014, 49(3): 332.
[9] Xing W, Guo R, Petakovic E, et al. Participation-based student final performance prediction model through interpretable Genetic Programming: Integrating learning analytics, educational data mining and theory[J]. Computers in Human Behavior, 2015, 47: 168-181.
[10] Holzinger A. Interactive machine learning for health informatics: when do we need the human-in-the-loop? [J]. Brain Informatics, 2016, 3(2): 119-131.
[11] Kourou K, Exarchos T P, Exarchos K P, et al. Machine learning applications in cancer prognosis and prediction[J]. Computational and structural biotechnology journal, 2015, 13: 8-17.
[12] Esmalifalak M, Liu L, Nguyen N, et al. Detecting stealthy false data injection using machine learning in smart grid[J]. IEEE Systems Journal, 2017, 11(3): 1644-1652.
[13] Zhou L, Pan S, Wang J, et al. Machine learning on big data: Opportunities and challenges[J]. Neurocomputing, 2017, 237: 350-361.
[14] Agrawal S, Agrawal J. Survey on anomaly detection using data mining techniques[J]. Procedia Computer Science, 2015, 60: 708-713.
[15] Mullainathan S, Spiess J. Machine learning: an applied econometric approach[J]. Journal of Economic Perspectives, 2017, 31(2): 87-106.