International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 08 | August - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

DATA MINING TECHNOLOGY BASED ON MACHINE LEARNING ALGORITHMS

SHILTON RODRIGUES (1DS21MC099)

Asst Prof. Dr.Chandrika Murali

Department of MCA, Dayananda Sagar College of Engineering, Bangalore, Karnataka, India

***Abstract*—[** Data mining technology, driven by machine learning algorithms, has emerged as a powerful approach for extracting valuable insights and patterns from large-scale datasets. This research paper provides an in-depth analysis of the advancements, challenges, and future directions in data mining technology based on machine learning algorithms. The paper explores key areas such as classification, clustering, association rule, mining, anomaly detection, feature selection, deep learning, privacy preservation, stream mining, transfer learning, and explainability. It discusses the significance of each area, highlights notable research contributions, and presents the challenges faced by researchers. Moreover, the paper outlines potential solutions to address these challenges and proposes future directions for advancing the field. By providing a comprehensive overview, this research paper aims to contribute to the understanding of data mining technology and its application in various domains.

I. INTRODUCTION

Two popular data mining algorithms are machine learning and statistical algorithms. The first one is used artificial intelligence to train and learn multiple sample sets so that it can automatically locate the necessary patterns and parameters. The second is to perform operations using probability, grouping, and correlation analysis. Additionally, the goals and places that correspond to various algorithms vary. These algorithms may operate separately. For their own purposes, they can be joined with one another. Widely applicable, with good data handling and self-organizing learning capabilities as well as accurate identification capabilities, the artificial neural network method in machine learning algorithms is helpful for the categorizing the problem of data processing. Work can be done with modelling.

Utilising gathered data, machine learning is a technique for automatically enhancing performance. The theory of statistical learning and optimisation is the forerunner of machine learning. It began with the development of computers. For various fields and issues, numerous methods have so far been presented. Decision trees, neural networks, support vector machines, the k-neighbor technique, and others are examples of representative algorithms. It is a crucial strategy for handling data mining issue. Data mining is an application-focused, cross-disciplinary concept. There are issues and demands for data mining when there is a huge volume of data accumulation in the sectors and industries like finance, retail, telecommunications and scientific research.

Categories of data mining tasks

Machine learning and data mining techniques offer a multitude of ways to process data, with distinct applications such as regression analysis, association rules classification and clustering. These methods, powered by advanced algorithms, enable us to extract valuable insights and patterns from complex datasets. Classification allows us to categorize data into predefined classes based on learned patterns, while regression analysis helps us predict numerical values based on input variables. Association rules unveil hidden relationships and dependencies among variables, aiding in market basket analysis and recommendation systems. Finally, clustering groups similar data points together, allowing us to identify patterns and discover natural divisions within the data. Each of these data mining methods can be implemented using various machine learning techniques tailored to specific needs and datasets, making the field dynamic and versatile.

Classification

Overview of classification algorithms

Decision trees and ensemble method

Support vector machines

Neural networks and deep learning

Recent advancements and applications

Challenges and research opportunities

Clustering

Introduction to clustering techniques

K-means and hierarchical clustering

Density-based clustering

Evaluation metrics

Emerging trends and applications

Challenges and potential research avenues

Association Rule Mining

Advantage of the applications of machine learning algorithm in data mining

The study of computational approaches in learning processes and to use computer based learning systems to address real-world issues is known as machine learning. One of the main areas of machine learning research is how to extract the corresponding concept description from the sample. As a result, a variety of machine learning techniques can be employed to directly address data mining issues. Finding relevant rules and intriguing patterns in huge databases is a topic known as data mining. Terabytes or even petabyte levels of big data cannot be stored into the computer's memory because the majority of conventional machine learning techniques rely on memory. As a result, many algorithms now in use struggle to manage huge data. Large-scale data processing and analysis are performed with the assistance of computers.

GSM network location based on machine learning

Predicting the location of a GSM network using machine learning involves collecting a dataset of labeled examples with GSM network signals and their corresponding locations. Relevant features, such as signal strength and cell tower information,

are extracted from the data. The dataset is preprocessed to handle missing values and outliers. A machine learning model is then trained using the labeled data and the extracted features. The model's performance is evaluated using a separate testing dataset. Optimization techniques, such as adjusting hyperparameters or exploring different algorithms, can be applied to improve the model's accuracy. Once the model is trained and optimized, it can be used to predict the location of new GSM network signals based on their features.

The term "mobile terminal positioning technology" refer to the employment of a variety of techniques to pinpoint the precise location of moving objects, including their height and degree of obscurity. The history of this technology is extensive. An early form of mobile communication was the lighting of a beacon fire to transmit military alerts, and military operations could be identified with the naked eye by knowing the general location of the beacon. Moving about is recognised as a mobile terminal's original technique of location. The most successful outdoor mobile terminal location technology at the moment is GPS, but its use has a number of drawbacks, including drastically reduced positioning effectiveness in densely populated high-rise areas, high battery consumption, and the requirement that mobile terminals have the corresponding hardware support, difficult computations, delays, and other factors. The best option is to employ a global positioning system. When GPS is only partially able to provide the positioning requirements, The machine learning-based outdoor mobile terminal location solution closes the gap.

The proposed method for mobile terminal positioning in the network uses a three-stage approach based on Support Vector Machine (SVM) and k-Nearest Neighbor (k-NN) methods. The method includes data preprocessing, training an SVM model, and refining predictions with k-NN. The SVM model learns patterns between features and locations, while the k-NN algorithm considers neighbors for more accurate positioning. By combining these techniques, the method improves the accuracy and reliability of mobile terminal positioning in the network.

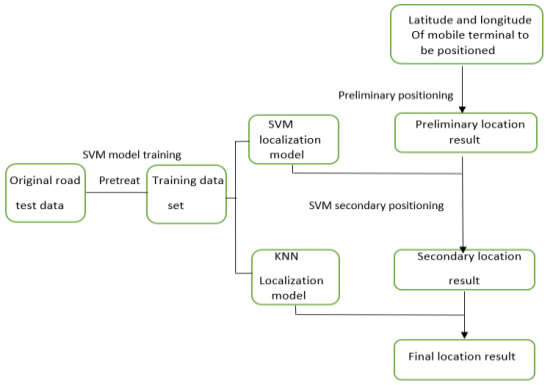

Figure 1:1. Location process figure based on machine learning

Model establishment

Support vector machine (SVM) positioning is the main approach employed in the modelling process. The method involves rasterizing the positioning regions into small raster areas, creating independent categories. In this approach, a large amount of terminal measurement data is collected and analyzed using computer algorithms. Machine learning techniques are then applied to resolve the localization problem.

The tiny raster patches are abstracted into various categories by the SVM-based localization techniques. It gathers a sizable number of mobile

terminals, along with their matching longitude and latitude positions, that send level measurement information inside each localization region. This categorised approach examines and addresses the issue of mobile terminal location by utilising machine learning.

It becomes more challenging to locate mobile terminals using machine learning methods when initial positioning based on the latitude and longitude of the base station is required. The methodology and categorisation become more difficult as the region expands. As a result, finding mobile terminals via machine learning techniques is more time-consuming and expensive. As a result, starting with the base station's longitude and latitude is a suitable technique for early placement.

Using the longitude and latitude of the base station, the positioning dataset's data are placed on a 1 km square grid. Since the distance between the mobile terminal and the primary service cell's base station is typically less than 500 metres in urban settings, it can be deduced that the mobile terminal must be positioned inside the 2 km square grid centred around the 1 km square grid. A representation of the dataset formats is shown in Table 1.

Overall, this approach combines SVM-based positioning with initial positioning based on base station coordinates to address the challenges of mobile terminal localization, particularly in larger regions.

Field	longitude	latitude	n-cell	TS
Interpretation	GPS longitude	GPS latitude	Number of sectors	Road test time
Field	LACCi 1-LACCI 7	Relev1-Relev7	ser_cell_lon	ser_cell_lat
Interpretation	Unique numberof sector	Sector reception power	Base station longitude of mainservice cell	Base station latitudeof main service cell

Table 1: Formats of data set

SVM-based secondary positioning

A square with a side length of 2 km is selected after initial positioning since the first stage SVM's range is 400 metres and the second season SVM generates ambiguous data on a grid of 100 metres. The positioning result is output as the longitude and latitude of the middle 100-meter grid. In comparison to the localization of the second-order vector machine, the computation complexity of the first-order vector machine is quite significant. After classification, the vector machine is frequently used to construct the decision function and all other types of vector machines for the sample points.

We substitute the following expressions for each training set's longitude and latitude values.

$$d = 2\pi R_{earth}/360$$
$$\text{relative X} = d(\text{lon} - \text{ref\_lon}) \cdot \cos(\text{lat} + \text{ref\_lat})$$
$$\text{relative Y} = d \cdot (\text{lat} - \text{ref\_lat})$$

The term "d" in the above formulas denotes the actual distance connected to a certain longitude or latitude. The terms "Relative X" and "Relative Y" refer to the distances in longitude and latitude directions, respectively, between the training data and the edges of the square placement area. 'Lon'

and 'Lat', on the other hand, stand for the positioning outcomes of the SVM algorithm.

These formulas play a crucial role in computing the distances and positions within the square positioning area based on the SVM outcomes. By incorporating the relative distances and actual physical distances, the formulas provide valuable insights into the location of the training data relative to the square positioning area. This information can be utilized to gain further understanding or make informed decisions in the mobile terminal positioning process.

Conclusions

This research suggests a high-performance machine learning-based method for locating mobile terminals in an outside network. It is more accurate and requires less processing than conventional methods. The following are the primary conclusions:

(1) The four main categories of data mining tasks are clustering, association rules, classification, and regression analysis.

(3) The machine learning-based localization of the GSM network consists of five main steps: model creation, data collection and Pre-processing, first positioning using the base station's latitude and longitude, secondary positioning using SVM, and final positioning using the k-neighbor approach.

(4) Experience shows that the machine learning-based positioning method significantly increases positioning accuracy while lowering time complexity.

References

[1] Research on Data Mining Technolog, GSM Location process figure based on machine learning by Shangran Li 2019 Journal of Physics: Conference Series

[2] A. Yosipof, O. E. Nahum, For combinatorial material science of all-oxide photovoltaic cells, data mining and machine learning tools are used[J]. 2015,

[3]Papamitsiou Z. and Economides A. A. Implementing learning analytics and educational data mining: A thorough literature analysis of empirical data[J]. 2014, 17(4) Journal of Educational Technology & Society.

[4] D'Oca S, Hong T. Learning about occupancy schedules using a data mining approach. 2015; 88: 395–408. Energy and Buildings.

[5] C. Voyant, G. Notton, S. Kalogirou, et al. A overview of machine learning techniques for predicting solar radiation[J]. alternative energy

[6] The use of machine learning and data mining techniques in diabetes research[J] by Kavakiotis I, Tsave OJournal of computational and structural biotechnology, 2017.

[7]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.