Scholar Publishing Group
Machine Learning Theory and Practice
https://doi.org/10.38007/ML.2020.010101
ISSN 2790-0983 Vol. 1, Issue 1: 1-9

Cloud Data Intelligence Detection Based on Decision Tree Algorithm

Tiegang Bai*

Police Dog Technical Institute, Criminal Investigation Police University of China, Shenyang110035, Liaoning, China

baitiegang@cipuc.edu.cn

**corresponding author*

Keywords: Decision Tree Algorithm, Cloud Data, Intelligence Investigation, Big Data Technology

Abstract: At present, many countries and regions are strengthening the construction of digital cities, expecting to obtain valuable intelligence from big data by means of information technology, and the massive accumulated security video and alarm information can be analyzed to discover the laws of the data and form correlated data to provide reference for police work. In this paper, from the actual needs of intelligence analysis, cloud computing and big data technology is used to design a cloud data(CD) intelligence investigation system(IIS) to meet the requirements of information development, by building an intelligence information center, using decision tree algorithm(DTA) to analyze and classify massive data, solve the problem of insufficient human resources for on-site investigation, effectively improve the efficiency of investigation and crime solving, and provide technical support for the construction of a harmonious and stable social environment.

1. Introduction

Intelligence investigation work is an important function of relevant government departments to carry out social governance and guarantee national security. The full use of CD processing, play the role of CD village sook and analysis, help intelligence investigation work to improve efficiency, reduce labor costs, early detection of problems, proper resolution of risks, accurate grasp of the situation changes and timely adjustment of policy direction.

At present, the CD processing in the field of intelligence investigation is still in the exploration stage our government departments enjoy the convenience of CD technology at the same time, but also need to face the CD technology application innovation. Some studies point out that government departments are collecting various intelligence information all the time and CD itself is a collection

of information data, which is a natural target for intelligence work. In order to make better decisions, CD intelligence investigation should develop in the direction of standardization, security, accuracy, and efficiency, and should implement the main responsibility of government supervision of CD processing, further strengthen the top-level design of CD processing, determine the data processing The government should implement the main responsibility for the supervision of CD processing, further strengthen the top-level design of CD processing, determine the boundary of data processing, and improve the degree of specialization of CD processing [1-2]. Studies and discussions in academia have focused on how to strengthen the application of CD as a technical means of government governance, suggesting that government departments should efficiently integrate heterogeneous database data, explore the deep correlations of criminal activities in time and space factors, and achieve control of the security situation from a macroscopic perspective [3-4]. Therefore, the strategic thinking and concept of the CD era should be conformed to in order to lay the foundation for the application of CD in the field of intelligence investigation.

In this paper, we firstly elaborate the concept of DTA and propose CART algorithm; then take CD IIS as the research object and analyze the system design performance requirements; then build the system network structure and functional modules; finally analyze the implementation process of CD intelligence mining based on DTA, as well as compare CART algorithm with C4.5 algorithm in CD intelligence classification accuracy.

2. Basic Overview

2.1. DTA

The decision tree selects the next branch of the tree structure based on the conditional judgment of each branch in the tree structure. The top node of the tree is set as the root node of the decision tree. The leaf nodes are the sets of attributes and the leaf nodes are the judges of the next classification. One of the greatest features of the DTA is that its classification rules can be fully reflected by the decision tree, which is one of the most obvious features of the DTA [5].

Based on the actual problem of the requirement, the Huffman tree recursive algorithm is constructed to build the decision tree, and then the decision tree is pruned. The construction of the decision tree is completed by the Huffman tree recursive algorithm and pruning serves to reduce the noise as much as possible by training and pruning operations. The quality of node splitting in the decision tree is measured using the " information gain " metric [6].

$$Information = -\log_2(p_i) \tag{1}$$

$$Entropy = -\sum_{i=1}^{m} p_i \log_2(p_i) \tag{2}$$

Where, pi is the probability of category i, Information is the mutual information, and Entropy is the information entropy.

CART algorithm is one of the DTAs, CART algorithm starts with a single node, the root node, and for each non-leaf node and the information gain of each node is calculated [7]. If the number of leaf nodes is less than a set threshold, CART selects the one with the highest value in the node splitting criteria to split. After splitting, each newly generated node will be categorized as a leaf node or a candidate split node. The algorithm will place the possible candidate nodes into the candidate set Q. When the node is classified it will be removed from Q. This process will be

repeated until any of the termination conditions are met [8-9].

2.2. Performance Requirements Analysis of CD Intelligence Detection System

(1) Scalability

CD IIS should have strong compatibility and scalability, because the system and other business systems in the data access there is a lot of intersection, not only in the early design must be left standard and open interfaces to strengthen the data sharing and data integration functions of the IIS; at the same time should also make the system should be updatable to constantly adapt to new needs for upgrading and improvement, to ensure that the growing The system should also be updateable to continuously adapt to new needs for upgrading and improvement, to ensure that the growing real business can be realized in the system through the expansion of functional modules, secondary development, etc. [10-11].

(2) Maintainability

The system should have the ability of self-testing and self-repairing for possible faults without human intervention, and the system can be tested at the same time. The maintenance process can troubleshoot and repair faults, achieve stable system upgrades and data backup, build a system recovery mechanism, improve system maintenance capability and reduce operation and maintenance costs [12].

(3) Security

System data security is an important factor for the survival of the software, and should rely on advanced technologies such as cloud storage, cloud computing and edge computing, and strengthen the investment in hardware equipment and software technologies such as data servers, network firewalls and main in networks [13]. The CD IIS requires security management of user accounts, intelligence, cases, suspects and other types of data to ensure safe storage of data, which can be encrypted for some information to prevent external hacking. Control user operation rights through user roles, permissions, etc., to prevent users from overstepping their authority to operate, data tampering; the existence of deliberate destruction of system data information, the need for timely backup of relevant business data, encryption of data processing. Also focus on the use of operation logs for user operation of core functional modules to facilitate timely repair of abnormal problems after they occur [14-15].

3. CD Intelligence Detection System Design

3.1. System Network Deployment Design

Network architecture deployment can provide a stable operating environment for system operation. Since the amount of data used and the amount of computing required by the IIS is very large, it is necessary to use cloud computing to establish correlations between information, integrate data from different business types, improve bandwidth utilization, and achieve optimal allocation of network link resources [16]. The network architecture design process needs to focus on business response speed, reasonable allocation of network services, and guaranteeing network service quality, while improving network operation and maintenance efficiency and utilization, improving fault diagnosis capability, and strengthening application depth construction. The network deployment of IIS needs to integrate Internet technology, IT and other technologies to provide on-demand, resilient services to provide a strong guarantee for the various services of intelligence analysis and to achieve data that can reach hundreds of G scale [17]. The network deployment architecture studied

in this paper is shown in Figure 1.

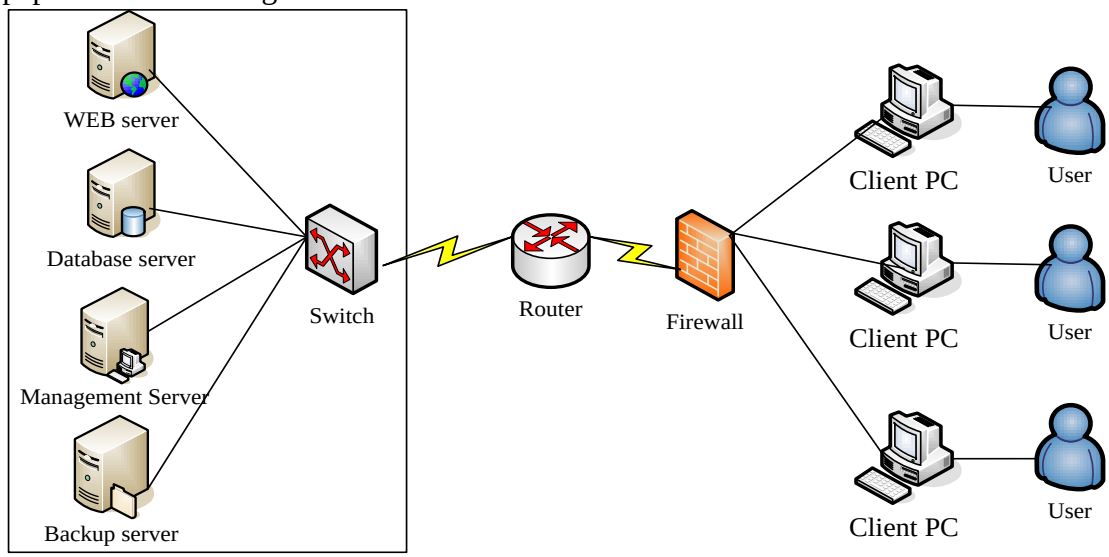

Figure 1. System network deployment architecture

3.2. CD Intelligence Investigation Function Module Design

Combined with the business process of intelligence work and the specific division of labor, the system is mainly divided into two parts in the design of functional modules: the foreground interactive system and the background analysis system, the foreground and background for different user objects, respectively, to achieve their different business needs. Front-line users and leading decision makers initiate investigation topics, submit and approve investigation requests, and receive investigation feedback results through the foreground interactive system, while intelligence department users perform data extraction, analysis and modeling, selection algorithms and other operations on investigation topics through the background analysis to derive investigation results and ensure the normal operation of the intelligence investigation business model [18]. The two major parts are relatively independent and interact with each other to ensure the security and confidentiality of the system while taking into account the convenience and efficiency. The specific functional decomposition of the system is shown in Figure 2.

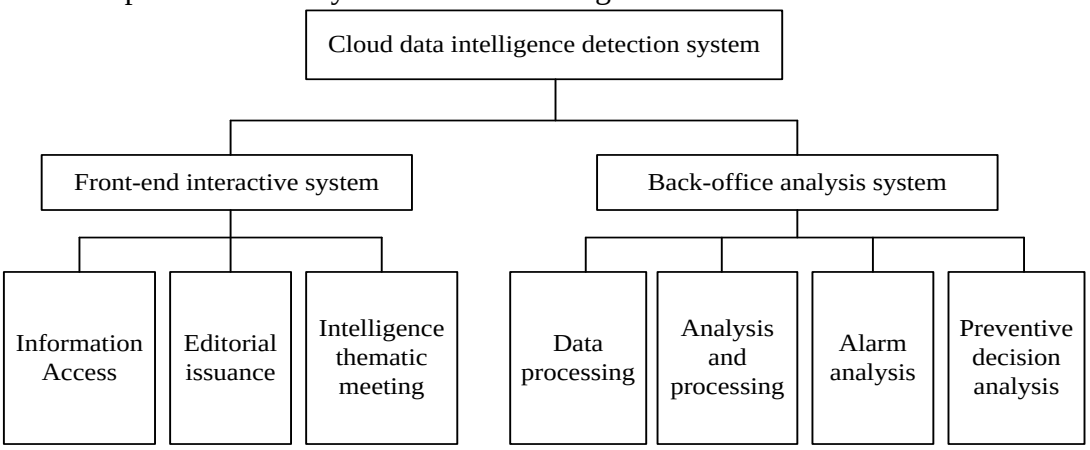

Figure 2. Functional module structure

(1) Front-end interactive system

Information access module: can be classified statistics of each unit to initiate the topic, has been feedback topics and investigation results, etc., the history of information can be screened by keywords, or use the built-in search engine for full-text search.

Editorial issuance module: receive the intelligence department feedback analysis of the results of the investigation, the results of the evaluation of the merits or according to the value of clues for the rating; leading decision-makers can make decisions based on intelligence recommendations whether the need to issue intelligence instructions to direct front-line units to carry out operations.

(1) intelligence thematic meeting: through the platform of intelligence information reported around the automated classification and aggregation, statistical analysis, combined with manual research and analysis and targeted investigation advice and work measures to enhance the connection between the police departments, improve the timeliness of information transfer, can effectively enhance the macro control of social security dynamics, to prevent major instability and the occurrence of major cases and incidents. Provide an efficient and convenient channel to realize the online synthetic operation mode led by comprehensive intelligence department, with the joint participation of multiple police departments and mutual collaboration.

(2) Background analysis system

Data processing module: the social data and other information collected from other systems in the public security network or other channels are extracted, imported, cleaned and transformed into standardized data that meet the analysis requirements of IIS.

Analysis and processing module: According to different application requirements, the corresponding algorithms are selected, such as association rule algorithm for police analysis, key personnel analysis, etc., and Bayesian algorithm for high-risk origin analysis. And the pre-processed data are analyzed and investigated, and the results of the system's operation are processed again by human power, and finally the intelligence products of practical value are derived as expected.

Police analysis: mainly used to grasp the changes in the security situation in the jurisdiction, from the changes in the trend to find the content that can predict the prediction. For example, the association rule algorithm is selected to dig the intrinsic connection between the police situation and some specific factors, and then get the key factors that cause the change of the security situation in the jurisdiction and other investigation results, according to which the intelligence department can issue early warning tips for specific areas, places and time periods; in addition, by building a more complex investigation model, the police data can be combined with other non-public security data and refined through processing to become valuable intelligence investigation results for reference.

Preventive decision analysis: classify or time series analysis of the characteristic individuals in the database to get the characteristics of the characteristic groups, as well as the illegal and criminal phenomena of different groups and their re-violation characteristics, etc. Through the analysis of the characteristic groups, we can carry out different work for different groups of people to effectively prevent crime in future public security affairs, and can provide the identification of suspicious elements, relevant social information and other effective clues for the investigation of public security cases. It can provide effective clues for public security case investigation, such as the identification of suspicious elements, relevant social information, etc., and visualize the analysis results.

4. Implementation and Application of CD IIS Based on DTA

4.1. Intelligence Mining Implementation

Using the learning method in data mining technology, a large amount of data can be used to learn the model and get a learned model, which can be processed for later data, prediction and decision making for later data. In this paper, the decision tree method is used to implement intelligence mining. The decision tree method mines the relationships of people, cases, and affairs in intelligence information to find out the relationship between elements such as time, people, and events and the target of investigation. The classification rules can be obtained easily and quickly through decision trees, and their tree results can visually represent the results of classification. The generation of decision trees includes two steps: the growth process and the pruning process.

The general process of decision data mining model is to examine the objective establishment, data set selection, selection of learning method, model learning completion, data analysis, result saving or deletion, etc. After the training and testing of the mining model is completed by mining the temporary residence information and the illegal crime information of the people involved in the case, the temporary residence and case learning model is obtained, and these important parameters can be used for predictive analysis. As Figure 3 shows the flow chart of DTA.

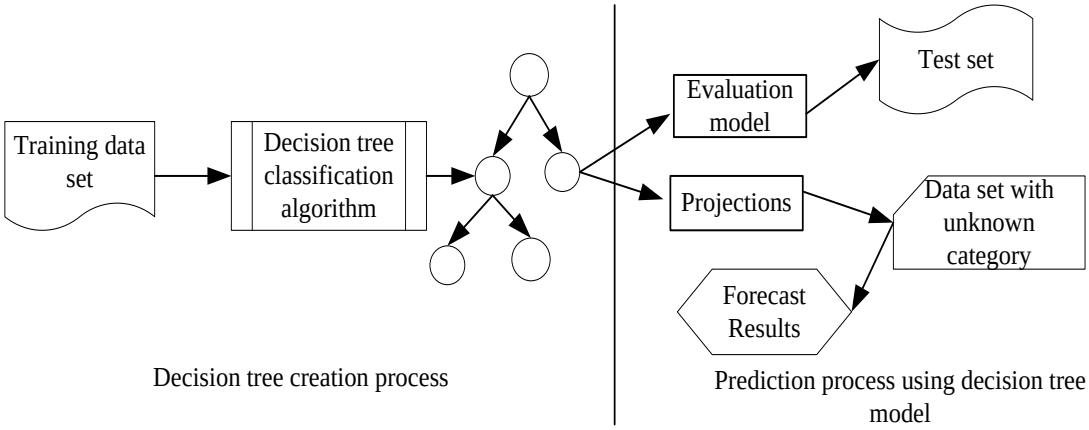

Figure 3. Basic flow of DTA

4.2. Application of CD Intelligence Data Classification Based on DTA

Since decision tree has classification function, DTA is widely used in data classification. In order to verify whether DTA is applicable to intelligence data classification, this paper compares the classification accuracy of CART algorithm and C4.5 algorithm. The experimental idea is to divide the intelligence dataset into 10 equal parts without limiting the number of leaf nodes, of which 9 parts are used to train the decision tree and the remaining one is used to test the trained decision tree. Based on the training dataset, decision trees are constructed using CART algorithm and C4.5 algorithm, and then the test dataset is used to verify the accuracy of the constructed decision trees. Ten experiments were conducted, and finally the average of these 10 experiments was used as the final experimental result. The comparison graph between the classification accuracy is shown in Fig. 4, and the average values of the final classification accuracy and the number of leaf nodes are shown in Table 1.

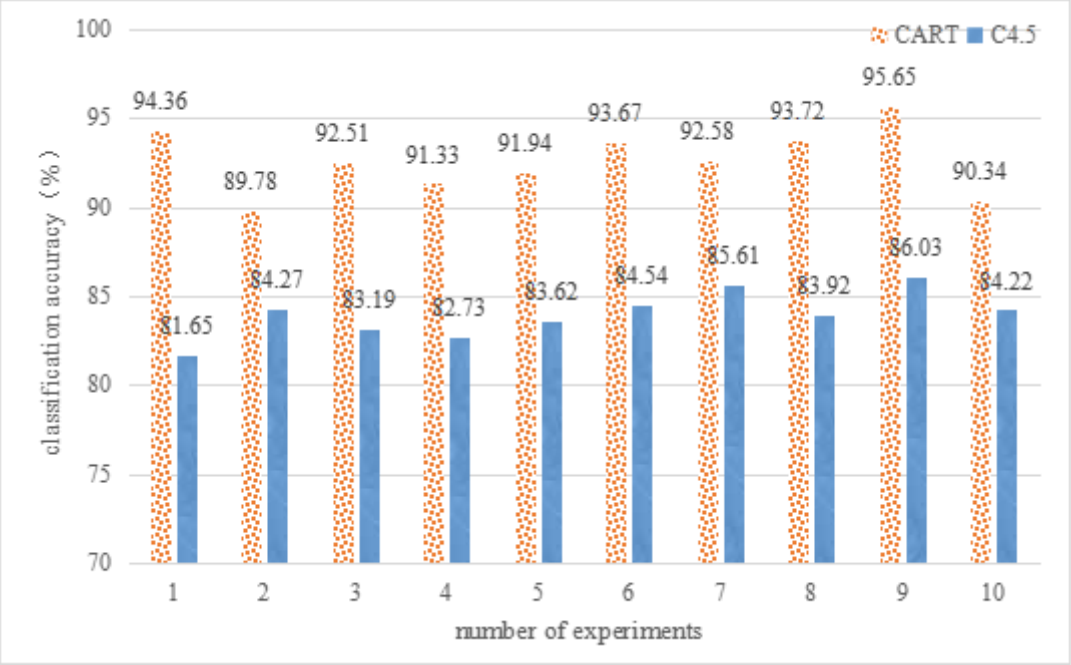

Figure 4. Classification accuracy of 10 experiments

According to the experimental results shown in Table 1, the decision tree constructed using the CART algorithm is more accurate than the decision tree constructed by the C4.5 algorithm, and the decision tree constructed by the CART algorithm generates fewer leaf nodes. the excessive branches of the C4.5 algorithm not only do not increase the classification accuracy, but also overcomplicate the generated decision tree. By merging the generated branches, the CART algorithm succeeds in making the constructed decision trees more concise and with better performance.

Table 1. Experimental results of CART and C4.5 algorithms

Algorithm	Average number of leaf nodes	Average accuracy(%)
CART	34	92.588
C4.5	88	83.978

5. Conclusion

Intelligence investigation is inseparable from the development of computer technology and network technology. Nowadays is the era of cloud computing, the resources of public security intelligence are not only limited to the internal public security organs, but also a large amount of social information is an important source of intelligence, and it is almost impossible to fully integrate and extract intelligence from the massive information without the support of strong computing power. Therefore, this paper constructs the CD IIS with the help of DTA, verifies the practicality of DTA in intelligence classification by comparing the accuracy of CART and C4.5 decision classification, and realizes the cluster parallel intelligence mining of CD intelligence investigation by DTA.

Funding

This article is not supported by any foundation.

Data Availability

Data sharing is not applicable to this article as no new data were created or analysed in this study.

Conflict of Interest

The author states that this article has no conflict of interest.

References

[1] Marziye Narangifard, Hooman Tahayori, Hamid Reza Ghaedsharaf, Mehrdad Tirandazian: Early Diagnosis of Coronary Artery Disease by SVM, Dtas and Ensemble Methods. Int. J. Medical Eng. Informatics 14(4): 295-305 (2020).

[2] Chandrashekhar Azad, Bharat Bhushan, Rohit Sharma , Achyut Shankar, Krishna Kant Singh, Aditya Khamparia: Prediction Model Using SMOTE, Genetic Algorithm and Decision Tree (PMSGD) for Classification of Diabetes Mellitus. Multim. Syst. 28(4): 1289-1307 (2020).

[3] Ferdinand Bollwein, Stephan Westphal: A Branch & Bound Algorithm to Determine Optimal Bivariate Splits for Oblique Decision Tree Induction. Appl. Intell. 51(10): 7552-7572 (2020).

[4] Hsu-Che Wu, Jen-Hsiang Chen, Pei-Wen Wang: Cash Holdings Prediction Using DTAs and Comparison with Logistic Regression Model. Cybern. Syst. 52(8): 689-704 (2020).

[5] Firoozeh Karimi, Selima Sultana, Ali Shirzadi Babakan, Shan Suthaharan: Urban Expansion Modeling Using an Enhanced DTA. GeoInformatica 25(4): 715-731 (2019). https://doi.org/10.1007/s10707-019-00377-8

[6] Muhamad Hasbullah Mohd Razali, Rizauddin Saian, Yap Bee Wah, Ku Ruhana Ku-Mahamud: An Improved ACO-Based DTA for Imbalanced Datasets. Int. J. Math. Model. Numer. Optimisation 11(4): 412-427 (2020).

[7] Evin Sahin Sadik, Hamdi Melih Saraoglu, Sibel Canbaz Kabay, Mustafa Tosun, Cahit Keskinkilig, Gonol Akdag: Investigation of the Effect of Rosemary Odor on Mental Workload Using EEG: An Artificial Intelligence Approach. Signal Image Video Process. 16(2): 497-504 (2020).

[8] Ekaterina Jussupow, Kai Spohrer, Armin Heinzl, Joshua Gawlitza: Augmenting Medical Diagnosis Decisions? An Investigation into Physicians Decision-Making Process with Artificial Intelligence. Inf. Syst. Res. 32(3): 713-735 (2020). https://doi.org/10.1287/isre.2020.0980

[9] Jjishnu Bhattacharyya, Manoj Kumar Dash: Investigation of Customer Churn Insights and Intelligence from Social Media: A Netnographic Research. Online Inf. Rev. 45(1): 174-206 (2020). https://doi.org/10.1108/OIR-02-2020-0048

[10] Anastasia Kioussi, Anastasios . Doulamis, Maria Karoglou, Antonia I. Moropoulou: Cultural Intelligence-Investigation of Different Systems for Heritage Sustainable Preservation. Int. J. Art Cult. Des. Technol. 9(2): 1 6-30 (2020). https://doi.org/10.4018/IJACDT.2020070102

[11] Steve Edwards: Heart Intelligence: Heuristic Phenomenological Investigation into the Coherence Experience Using Heartmath Methods. Al Soc.34(3): 677-685 (2017). https://doi.org/10.1007/s00146-017-0767-7

[12] Stefania Costantini, Giovanni De Gasperis, Raffaele Olivieri: Digital Forensics and Investigations Meet Artificial Intelligence. Ann. Math. Artif Intell.86(1-3): 193-229 (2019). https://doi.org/10.1007/s10472-019-09632-y

[13] Shahriar Akter, Katina Michael, Muhammad Rajib Uddin, Grace McCarthy, Mahfuzur Rahman: Transforming Business Using Digital Innovations: The Application of AI, Blockchain, Cloud and Data Analytics. Ann. Oper. Res.308(1): 7-39 (2020). https://doi.org/10.1007/s10479-020-03620-w

[14] Elena Verdu, Yuri Vanessa Nieto, Nasir Saleem: Call for Special Issue Papers: Cloud Computing and Big Data for Cognitive loT: Deadline for Manuscript Submission: August 15, 2020. Big Data 10(1): 83-84 (2020).

[15] Amanpreet Kaur Sandhu: Big Data with Cloud Computing: Discussions and Challenges. Big Data Min. Anal. 5(1):32-40 (2020).

[16] Cameron K. Peterson, David W. Casbeer, Satyanarayana G. Manyam , Steven Rasmussen: Persistent Intelligence, Surveillance, and Reconnaissance Using Multiple Autonomous Vehicles With Asynchronous Route Updates. IEEE Robotics Autom. Lett. 5(4): 5550-5557 (2020). https://doi.org/10.1109/LRA.2020.3008140

[17] Ayesha Bhimdiwala, Rebecca Colina Neri , Louis M. Gomez: Advancing the Design and Implementation of Artificial Intelligence in Education through Continuous Improvement. Int. J. Artif. Intell. Educ. 32(3): 756-782 (2020).

[18] Irene-Angelica Chounta, Emanuele Bardone, Aet Raudsep, Margus Pedaste: Exploring Teachers' Perceptions of Artificial Intelligence as a Tool to Support their Practice in Estonian K-12 Education. Int. J. Artif. Intell. Educ. 32(3): 725-755 (2020).