



Research on Data Mining Technology of Online Intelligent Education Under Collaborative Learning

Zhi Zhang^(✉)

Technique Center of Modern Education, Jiangxi Science and Technology Normal University,
Nanchang 330013, China
lky02560@tom.com

Abstract. The educational data mining technology is delayed and has limitations in the process of operation. It cannot effectively deal with the characteristics of behavioral sequence, which leads to the problem of low accuracy and recall rate. This paper proposes a data mining technology for online intelligent education under collaborative learning. Construct the learner portrait under collaborative learning from the four dimensions of curriculum learning characteristics, task completion characteristics, interpersonal interaction characteristics and learning performance, and extract the characteristics of educational data. The decision tree method is used to cluster the students' behavior characteristics, and the online intelligent education data mining algorithm is designed to analyze the correlation between students' behavior habits and comprehensive quality. The experimental results show that compared with the data mining technology based on model driven and homomorphic encryption privacy protection, the accuracy and recall of the data mining technology in this paper are improved, which can analyze educational data more sensitively and provide reference for decision-making.

Keywords: Collaborative learning · Online learning · Wisdom education · Data mining · Data characteristics · Learning behavior

1 Introduction

Educational informatization has had a revolutionary impact on educational development. The national informatization development strategy has classified educational informatization as a key issue. Smart education is a new educational form to promote the development of students' smart ability. The support of big data technology has laid a solid foundation for the development of smart education. With the maturity of online learning education, the increase of network teaching resources and the popularization of data application, a variety of special types of education data have been generated and accumulated in various education systems. These make education data mining an independent research field. The impact of big data on education is reflected in the following five

aspects: Educational big data directly reflects the current situation of teaching and effectively promotes teaching; Education big data can correctly grasp the individual needs of learners and accelerate the development of personalized education; Educational big data objectively presents teaching reality and helps educational decision-making to be scientific; Education big data has changed the technology and method of education evaluation and provided thinking support for teaching evaluation; Education big data promotes intelligent education and positively affects the realization of intelligent education. Most colleges and universities have built campus information systems, so that schools can not only obtain basic data related to students' campus learning, life and daily management, but also analyze valuable information in big data, study students' behavior characteristics and behavior laws, and reflect the improvement of school education and teaching quality, students' learning efficiency and ability training. Analyze and summarize the ways, methods and measures conducive to students' personal development, help students improve their living conditions, improve their academic achievements, assist teaching management, speed up the process of "double first-class" construction in Colleges and universities, and improve the decision-making and management ability of talent training in Colleges and universities. Using educational data mining technology, educators can process and analyze a large number of educational data, and study students and how to learn more deeply, so as to find the value hidden behind the huge data. At present, the academic community has carried out extensive research on online intelligent education data mining and achieved some research results. Educational data mining applies the methods and techniques of machine learning, statistics, data mining psychopedagogy, information retrieval, cognitive psychology and recommendation systems to various educational data sets to solve educational research problems. Based on the characteristics of educational big data mining, CHEN et al. [1] summarizes the purpose and process of educational big data mining based on deep learning. Rawlings et al. [2] proposes expanding OT's role in the mental health treatment of parkinson's disease through high-quality online education, when the data set scale is large, the mining speed increases gradually. PANG et al. [3] proposed an efficient data mining simulation method based on homomorphic encryption privacy protection. Although the above methods can improve the quality of educational data mining, they still cannot effectively deal with the characteristics of behavior sequence. To solve this problem, this paper studies the data mining technology of online intelligent education under collaborative learning. The innovation of this technology is to make use of data mining technology to expand educational data in breadth, depth and subdivision, and to use diversified, multi-level and unstructured educational data to reflect the educational background, process and results in a more comprehensive and true way.

2 Data Mining Technology of Online Intelligent Education Under Collaborative Learning

2.1 Establishing Learner Profiles Under Collaborative Learning

Collaborative learning is a strategy to organize students to learn through groups or teams. The collaborative work of group members is an organic part of achieving the learning

goals of the class. Students in group collaboration activities can share the information and learning materials they explore and discover in the learning process with other members of the group, or even with other groups or the whole class. Learner model was first proposed in the field of intelligent computer-aided instruction. It is expressed as a group of complex data structures to describe the characteristics of learners that can be understood by computers. As a tool for analyzing learners, learner portrait is of great significance to optimize the data mining technology of online intelligent education. Online learner portrait can provide information for teaching stakeholders, so that the online learning environment can meet the learning needs of learners. Learner model is to establish a mathematical model and store data through structured representation of learners' basic information, knowledge and ability level, interest preference, learning style and emotion, so as to predict and evaluate learners and reflect their learning state. The learner portrait model based on E-learning behavior can describe its individual characteristics, Including the basic information of learners (name, gender, occupation, etc.), learning style, learning interest, cognitive level, etc. The more comprehensive the data acquisition of individual characteristics and online learning behavior of online learners, the greater the coverage of the construction of learner portrait model, and the more conducive it is to provide accurate online learning support services for teaching and learning stakeholders. Under collaborative learning, students can independently set up Plan the curriculum research program, fully cultivate and give full play to students' awareness of autonomous learning, and have more learning space for students who have spare power. Especially by participating in quality development courses, students' practical ability has been significantly improved. According to the types of available data and combined with the behavior characteristics of learners under collaborative learning, this study constructs a learner portrait model based on collaborative learning behavior from four dimensions: curriculum learning characteristics, task completion

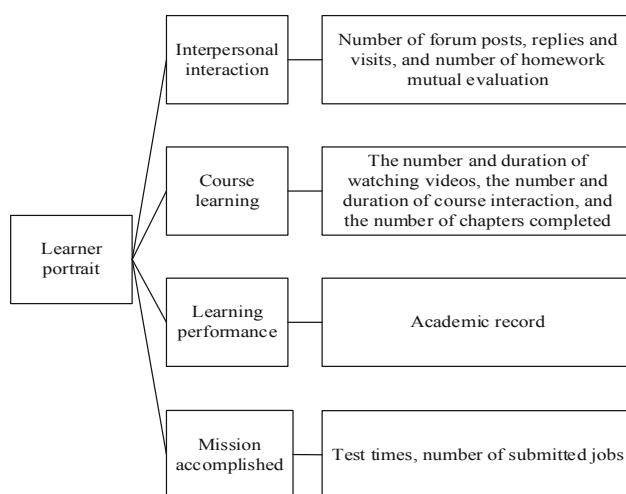


Fig. 1. Learner portrait

characteristics, human-computer interaction characteristics and learning performance. The portrait model is shown in Fig. 1.

Under the operation mechanism of cooperative learning, free course selection makes students no longer study in class as a unit, the time and space of students' communication and joint learning are broken, the binding force and cohesion of group activities are strengthened, and the function of cooperative learning group as an objective carrier for the cultivation of collective sense of honor is also strengthened. Each part of the learner portrait model based on collaborative learning behavior constructed in this paper corresponds to the specific learning behavior of xapi standard, which provides teaching and learning stakeholders with a more complete grasp of the learning situation of network scholars and make more scientific and effective educational decisions.

2.2 Extracting Educational Data Features

Online intelligent education platform is an important part of the construction of digital campus in Colleges and universities, including public data exchange platform, unified identity authentication platform and unified service portal platform. Through the establishment of three platforms, standards and specifications, form a basic platform for horizontal and vertical interconnection, aggregate the scattered and heterogeneous applications in the school, integrate high-quality resources, establish a sharing mechanism, realize the seamless access and business integration of various application systems, and solve the remaining problems of information island, data inconsistency and difficult sharing. Each learner has its own characteristics. When using the intelligent learning system, it is necessary to establish its own learner model for each learner, and the model is constantly self modified with each learner's learning; At the same time, the intelligent learning system under the background of big data stores a large number of characteristic information of other learners. Therefore, we should also consider the impact of group information on each learner model, so we should constantly feed back to the learner model for correction [4]. Students' learning habits have been formed for a long time, and their tendencies and behaviors are not easy to change with external factors. Good and regular life behavior habits are beneficial to academic performance. These habits are closely related to students' ability of self-control and self-discipline. By extracting the characteristics of educational data, it provides the basis for student behavior classification and data mining. Most of the traditional education data features are extracted according to students' scores, basic information of students' enrollment and other contents. These features lack certain comprehensiveness and flexibility in front of big data. For education big data, this paper uses principal component analysis to extract the main characteristics of education data, and reflects the main information contained in the overall indicators through the linear relationship between multiple indicators. Establish the initial matrix of education data indicators and calculate the correlation coefficient. The correlation coefficient matrix needs to meet the following conditions:

$$|P - \chi| = 0 \quad (1)$$

In formula (1), P represents the correlation coefficient matrix; χ represents the feature root. Extract the corresponding feature vector and convert the standardized index variable

into the main component, which can be expressed as:

$$B = \mu_1 A_1 + \mu_2 A_2 + \cdots + \mu_n A_n \quad (2)$$

In formula (2), B represents the principal component feature; A represents the education data index; μ represents the feature vector; n represents the number of feature vectors. College students with different academic achievements have differences in campus card consumption, which reflects college students' different consumption needs and consumption psychology. The extraction of learning features can better explain the behavior data and have a more comprehensive understanding of the data in advance. Then it analyzes the campus behavior and performance ranking data, and converts the original behavior data into behavior characteristics related to academic performance.

2.3 Clustering of Students' Behavior Characteristics in Online Wisdom Education

Intelligent education and intelligent learning are the inevitable products of the rapid development of educational informatization. Wisdom is the comprehensive application of knowledge, intelligence and creativity and the transformation of mental operation to external practice. We pay attention to the leading role of ethics and values in learning, production and life practice. The intelligent learning environment takes students as the main body, intelligently pushes teaching resources for students, supports students' inquiry learning methods, and cultivates students' creative thinking ability. The consumption data of students' all-in-one card, card swiping and attendance data in class, book borrowing record data, network authentication data and record data of participating in various competitions and community activities during school reflect the learning and life rules of students' real-time activities during school. These data can dynamically and accurately map students' school behavior characteristics, And the behavior behind these data. By constructing the decision tree, the clustering rules of online intelligent education students' behavior characteristics are extracted. Feature selection is to find features with strong ability to identify training sample data, which can greatly improve the efficiency of decision tree classification [5]. In this paper, information gain and Gini coefficient are used as the criteria for selecting classification features. The calculation formula of information gain can be expressed as:

$$w = h(a) - h(a|\lambda) \quad (3)$$

In formula (3), w represents information gain; h represents information entropy; a represents training data set; λ represents feature. It can be seen from the formula that information gain is the difference between information entropy under feature conditions. The greater the change of information gain, the more orderly the information changes, and the stronger the classification ability of the feature to data samples [6]. The calculation formula of Gini coefficient can be expressed as:

$$g = 1 - \sum_{i=1}^c \vartheta_i^2 \quad (4)$$

In formula (4), g represents Gini coefficient; c represents the total number of categories; i represents a single category; ϑ represents probability. Like information entropy, Gini coefficient is used to measure the uncertainty of information [7]. Taking the cluster analysis of students' learning habits as an example, the comparison results of learning habits can be obtained through the analysis of four indicators: class attendance rate, book borrowing volume, number of times in and out of the library and length of learning time. For such students whose class attendance is not good, the amount of books borrowed is not much, the number of library access is also small, and the length of learning time is also low, the school tutor should focus on the learning situation of such students, and timely urge them to strengthen learning discipline and cultivate good learning habits. This paper applies decision tree classification algorithm to students' behavior habit data and learning achievement data, constructs the corresponding decision tree classification model, extracts classification rules, and analyzes the correlation between students' behavior habits and students' learning achievement, so as to provide scientific basis for teachers' teaching and scientific guidance for students' learning.

2.4 Design Online Intelligent Education Data Mining Algorithm

The main purpose of data mining is to extract new and valuable information from various data sources, analyze these information, find the hidden laws or associations between data, and establish a model to support decision-making [8]. According to the structure and properties of the data obtained from data preprocessing, select the appropriate data mining algorithm to mine the data and build the corresponding model [9]. For some data mining algorithms, it is necessary to divide the data set into training data set and test data set, build the model with training data set, and test the performance of the model with test data set. According to the cluster analysis of students' behavior characteristics in the previous section, students' behavior has been classified. The following uses data mining algorithm to analyze the correlation between students' behavior and comprehensive quality. Aiming at the problem that the characteristics of behavior sequence cannot be effectively processed in data mining, the prior knowledge of the properties of frequent itemsets is used for pruning, so as to control the increase of candidate itemsets [10, 11]. Apriori algorithm uses the iterative method of layer by layer search to find frequent itemsets, and uses k -itemsets to search $(K + 1)$ - itemsets, so as to traverse all frequent itemsets in the whole data set. The flow of online intelligent education data mining algorithm is shown in Fig. 2.

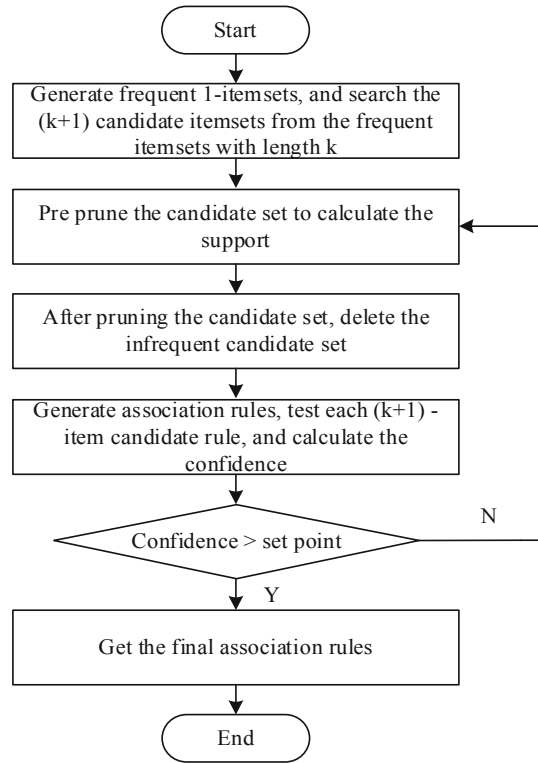


Fig. 2. Data mining algorithm flow of online intelligent education

Due to the complexity of iterative conditions of frequent itemsets, this paper uses the sequence model of cyclic neural network to model the characteristics of students' behavior, and explores the internal relationship between students' behavior characteristics and comprehensive quality. The sequence model exists not only between the adjacent input layer and the hidden layer, between the hidden layer and the output layer, but also between the hidden layer and the hidden layer in the time dimension. The update value formula of the hidden layer is:

$$\eta = f(\delta y + \varpi \tau + \nu) \quad (5)$$

In Eq. (5), η represents the hidden layer state; f represents the activation function; δ represents the hidden layer state weight matrix; y indicates neuronal status; ϖ is the state weight matrix from the hidden layer to the input layer; τ represents the input data at the current time; ν represents the offset term. The weight parameters are updated by time-varying back-propagation algorithm. The weights of the three parameters of the network model are shared in the time dimension, that is, the model parameters at different times are completely consistent, which reduces the training parameters. The core of intelligent learning environment is to use intelligent information technology and virtual environment to assist students' learning and life, provide different teaching situations, record and track students' learning process in the whole process, for in-depth

exploration and analysis of teaching and learning modes, and explore learning paths suitable for different students. The student behavior data sequence is used as the input of the encoder, and then transformed into a series of high-dimensional implicit representations. The potential expression of hidden layer, namely behavior characteristics, is mined through deep RNN, so as to understand students' behavior. By dividing the grade degree of the explicit learning behavior dimension of online intelligent education data, the corresponding grades of learners are converted, so as to obtain the results of the learning behavior of learners on the network teaching platform, so as to provide reference for improving the learning evaluation of learning users and making corresponding intervention measures for teachers' teaching decision-making in the later stage. Based on the above process, the research on data mining technology of online intelligent education under collaborative learning is completed.

3 Experimental Analysis

3.1 Experimental Data Set

This paper proposes a data mining technology for online intelligent education under collaborative learning. The effectiveness of this technology is verified by experiments. The practical research object selected in this experiment is the course a of China University MOOC network. Course a is one of the teacher education courses on the "China University MOOC" platform, which provides online e-learning services for learners. At present, the methods of learning behavior data collection of learning platform are mainly to track and record learning behavior through computer technology, such as log access, web server and so on. This paper makes use of the background database of the existing network teaching platform and the learners' learning behavior data recorded in the database. Firstly, this paper divides students' learning logs into different sessions. If the time interval between two consecutive questions is more than 24h, the record will be divided into two sessions. Therefore, his learning data can be divided into different session records. Obviously, if a conversation contains more records, it reflects that students spend more time studying. In order to collect the learning behavior data on the network teaching platform, when mining the learning behavior, the learning behavior data of learners are mainly counted from three functional modules: curriculum statistics, learning statistics and traffic statistics. Do some simple processing for the collected data, such as data format confusion, inconsistent form, partial data missing and so on. There is confusion in the data structure of students' data, and there is also a lack of data in the collected data due to factors such as curriculum and students' suspension. For example, some students' course scores are failed, some are zero, and there are blank values. In order to ensure the integrity and availability of data, this paper processes the metadata by merging tables, unified types and peer-to-peer comparison of student information, deletes the redundant data and modifies and adds the missing data. The data logs downloaded through the background interface are all Excel forms. The data in the form are large, including useless information on this research, such as basic information such as course number, college, specialty, class, ID number, mobile phone number, etc. Considering that learning behavior data involves the privacy of learners, data desensitization has been anonymized.

3.2 Experimental Setup

In order to improve the validity of experimental results, ensure that the control group and the experimental group of the same environment Settings. In this paper, python language and pycharm compilation platform are used as data mining tools to mine and analyze students' behavior habits and academic performance data, and study the correlation between students' behavior habits and academic performance. Six servers are used in the whole process of student behavior data analysis, which are virtual machines of VMware virtualization platform in cloud data center. These servers are in the same LAN, which constitutes the environmental foundation of Hadoop cluster. The specific configuration is shown in Table 1.

Table 1. Experimental configuration

Host name	Memory/G	Hard disk/G	Describe
Master	64	1000	Install Hadoop cluster deployment
Slave A	32	800	Rsyslog and other services
Slave B	32	800	Zookeeper and other services
Mysql	16	900	Database
Front end	32	500	Install NginxRedis
Back-end	32	500	Install redis

As the main operating system, Centos system has a 4-core Intel(R) Xeon(R) E5-2660 v2 2.20 GHz CPU. Three servers are used for big data analysis, one is used as the master node, and the other two are slave nodes. The other one is used as a MySQL database server to store the results of data analysis, and the two deploy front-end applications and back-end services.

3.3 Results and Analysis

In order to quantitatively evaluate the performance of the method, 70% of the data sets are randomly selected as the training set and the rest as the test set. This paper adopts two popular evaluation methods, namely, accuracy and recall evaluation experiment. The accuracy is calculated as follows:

$$\alpha = \frac{N_1 + N_2}{M_1 + M_2} \quad (6)$$

In formula (6), α represents the accuracy rate; N_1 represents the number of actually positive samples in the predicted positive samples; N_2 represents the number of actually negative samples in the predicted negative samples; M_1 and M_2 represent the total number of positive and negative samples. Similarly, the calculation formula of recall rate can be expressed as:

$$\beta = \frac{N_1}{N_1 + N_3} \quad (7)$$

In Eq. (7), N_3 represents the number of predicted negative samples but actually positive samples. The accuracy rate represents the proportion of students' scores correctly classified in all sequence behaviors, while the recall rate represents the proportion of students' scores classified as positive in all actual scores. This experiment compares the output results of educational data mining technology with the results of educational data mining technology based on model driven and homomorphic encryption privacy protection, so as to verify the effectiveness and superiority of online intelligent educational data mining technology proposed in this paper based on collaborative learning. The comparison results of the accuracy of various data mining technologies are shown in Table 2.

Table 2. Comparison results of accuracy (%)

Number of experiments	The educational data mining technology proposed in this paper	Educational data mining technology based on model driven	Educational data privacy protection based on homomorphic encryption Mining Technology
1	96.07	89.44	88.43
2	94.58	88.68	85.68
3	95.29	87.26	86.29
4	97.66	91.06	84.06
5	96.32	90.39	85.82
6	95.24	88.52	86.64
7	94.81	89.81	85.31
8	95.65	87.95	81.42
9	96.23	85.63	82.85
10	96.12	89.24	85.93

According to the test results in Table 2, the accuracy of the educational data mining technology under collaborative learning proposed in this paper is 95.80%, which is 7% and 10.56% higher than the educational data mining technology based on model driven and homomorphic encryption privacy protection, respectively. Therefore, the data mining technology proposed in this paper has relatively high accuracy, has certain advantages in dealing with various types of educational data, and can analyze the potential relationship between students' behavior. The comparison results of recall rates of various data mining technologies are shown in Table 3.

According to the test results in Table 2, the accuracy of the educational data mining technology under collaborative learning proposed in this paper is 90.69%, which is 9.65% and 8.51% higher than the educational data mining technology based on model driven and homomorphic encryption privacy protection, respectively. From the comparison

Table 3. Comparison results of recall rate (%)

Number of experiments	The educational data mining technology proposed in this paper	Educational data mining technology based on model driven	Educational data mining technology based on homomorphic encryption privacy protection
1	91.47	81.52	84.44
2	90.88	78.36	82.68
3	89.99	78.42	83.26
4	88.65	79.83	85.02
5	90.32	82.66	81.31
6	91.09	81.22	82.95
7	92.14	80.05	80.28
8	91.48	83.11	83.64
9	90.52	84.42	78.41
10	90.33	80.85	79.82

results of recall rate, the data mining technology in this paper has better performance, can capture more purpose features on the basis of ordered behavior features, and may achieve better prediction. Based on the experimental results of accuracy and recall, the data mining technology proposed in this paper can more sensitively analyze students' daily activities and more appropriately reflect their life and learning conditions.

4 Conclusion

Big data is gradually penetrating into all aspects of human society, not only changing people's way of thinking, work and lifestyle, but also changing social productivity and production relations. Educational data mining has become a main direction of big data application in the field of education. In collaborative learning, data mining technology can not only help to improve the school management and decision-making ability, but also predict and warn students' abnormal behavior and academic achievement by analyzing the data of various application systems of digital campus. This paper proposes a data mining technology for online intelligent education under collaborative learning. Its innovation point is to construct the learner portrait under the cooperative learning from four dimensions and extract the characteristics of educational data. The decision tree method is used to cluster the behavioral characteristics of students, and the online intelligent education data mining algorithm is designed to analyze the correlation between students' behavioral habits and comprehensive quality. This technology can improve the accuracy and recall of educational data mining analysis, and record the relationship between student behavior sequence and achievement in digital form.

However, there are still some limitations in this paper. The constructed learner learning behavior data mining model adopts the existing explicit learning behavior data, and does not analyze the implicit learning behavior data. In addition, the weight value can be determined according to the learning behavior of different learners, so as to enhance the internal correlation between learning behavior and achievement.

References

1. Chen, D., Zhan, Y., Yang, B.: Analysis of applications of deep learning in educational big data mining. *E-education Research* **40**(2), 68–76 (2019)
2. Rawlings, D., Winsall, M., Yin, H., et al.: What is a compassionate response in the emergency department? Learner evaluation of an End-of-Life essentials online education module. *Emergency Medicine Australasia: EMA* **33**(6), 983–991 (2021)
3. Pang, J., Sui, M.: Homomorphic encryption privacy protection data efficient intelligent mining simulation. *Computer Simulation* **36**(6), 316–319 (2019)
4. Shen, G.: Application research of big data analysis in college wisdom education. *Modern Electronics Technique* **42**(04), 97–100 (2019)
5. Zurawski, S.A., Pickett, K.A., Widmer, M.: Expanding OT's role in the mental health treatment of Parkinson's disease through high-quality online education. *Am. J. Occup. Ther.* **75**(2), 1–11 (2021)
6. Weber, M.M., Larkin, D.J., Patrick, M.: Creating informed consumers of aquatic invasive species management programs through online education for nonprofessionals. *Invasive Plant Science and Management* **15**(1), 41–48 (2022)
7. Networks, J.: Online education services selects juniper networks to elevate student experiences in the digital era. *Commwire Magazine: Incisive, Intormed, Independent, Objective* **8**(11), 13–14 (2022)
8. Li, P., Jia, L., Li, W.: Research on a book recommendation model based on data mining. *Journal of Zhejiang University of Technology* **47**(1), 80–85 (2019)
9. Zhang, P., Ding, L., Jiang, N., et al.: Data mining algorithm of the automation system of the distribution network based on the support-confidence-lift framework and its application. *Electrical Measurement & Instrumentation* **56**(10), 62–68 (2019)
10. Gao, Y.: Homogeneous vulnerability mining simulation of online network heterogeneous fault tolerant data. *Computer Simulation* **37**(3), 377–380 (2020)
11. Mohamed, N., Salama, M.M.A.: Data mining-based cyber-physical attack detection tool for attack-resilient adaptive protective relays. *Energies* **15**(12), 1–15 (2022)