Research on Collaborative Classification of E-Commerce Multi-attribute Data Based on Weighted Association Rule Model

Yi-huo Jiang(✉)

Fuzhou University of International Studies and Trade, Fuzhou 350202, China
hbgv96012@126.com

Abstract. Because the association between multi-attribute data of e-commerce is not obvious, the traditional collaborative classification method of e-commerce multi-attribute data has the problem of low classification accuracy. Therefore, the weighted association rule model is introduced to realize the optimal design of collaborative classification method of e-commerce multi-attribute data. Firstly, the weighted association rule model is built, and the multi-attribute data is mined and cleaned under the e-commerce platform. Taking the processed e-commerce data as the sample, the multi-attribute data classification index of e-commerce is determined. Through setting project weight, e-commerce data attributes and calculating multi-attribute relevance, multi-attribute data collaborative classifier is obtained. In the weighted association rule model, the collaborative classifier is used to get the multi-attribute data collaborative classification results of e-commerce. Compared with the traditional collaborative data classification methods, it is concluded that the accuracy of collaborative data classification is improved under the e-commerce platform of clothing and food 24.22%.

Keywords: Weighted association rule model · E-commerce · Multi-attribute data · Data collaborative classification

1 Introduction

E-commerce usually refers to a new type of business which is carried out by the buyer and the seller without meeting each other in a wide range of business activities all over the world, under the open Internet environment, to realize online shopping of consumers, online transactions and online e-payment among merchants, as well as various business activities, transaction activities, financial activities and related comprehensive service activities Business operation mode [1–3]. In the new business operation mode of e-commerce, one of the most important problems faced by merchants is how to obtain the market demand information of commodities in time and actively, develop the purchasing potential of customers, and find the corresponding superior commodities as soon as possible, so as to adjust the business plan as quickly as possible [4–6]. E-commerce, as the business model of information society, is developing at a faster speed than people expected. The U.S. government decided in the morning that e-commerce is a major information construction project, and formulated the "global

S. Liu and L. Xia (Eds.): ADHIP 2020, LNICST 347, pp. 377–388, 2021.
https://doi.org/10.1007/978-3-030-67871-5_34

e-commerce framework". From the perspective of China's domestic situation, the overall situation of Internet development determines the success or failure of e-commerce marketing. According to the report of China Internet Network Information Center, as of June 30, 2018, the number of Chinese Internet users reached 80200. However, with the increasing scale of e-commerce websites, the types of goods stored in websites and the relationship between these types are becoming more and more complex. In order to facilitate the users' query and respond to the requirements of the market, e-commerce websites must first classify a large amount of commodity data, and then achieve the goal of information extraction and intelligent market decision-making [7–10].

E-commerce multi-attribute classification is one of the most important technologies in the application field of e-commerce voucher, and many algorithms have been proposed so far. E-commerce multi-attribute classification is a technology which constructs a classifier according to the characteristics of data set and assigns a class to the unknown class samples. The process of constructing classifier is generally divided into two steps: training and testing. In the training stage, the characteristics of the training data set are analyzed to produce an accurate description or model of the corresponding data set for each category. In the test phase, we use the description or model of category to classify the test and test its classification accuracy. E-commerce website can use the classification algorithm in data mining, through training and testing, construct the classifier of commodity information and category, and realize the automatic classification of commodities [11–13].

At present, multi-attribute data classification methods in e-commerce platform include decision tree classification method, support vector machine classification method and K nearest neighbor classification method, but there are some problems in current classification methods, such as low classification accuracy and long classification time. In order to solve the above problems, a weighted association rule model is proposed. General association rule mining assumes that all items in the database have the same importance. When calculating frequent item sets, only the frequency of items is considered. But in some application areas, users pay different attention to different projects, that is, the importance of projects is different. In order to reflect the importance of the project, project weighting is introduced. After project weighting, when mining the weighted association rules, the frequency of project set appearing in transaction database and the weight of project should be considered comprehensively. At the same time, the weighted process can not only distinguish the importance of the project, make the mining results more reasonable, but also greatly improve the efficiency of the algorithm. Because in the association rule algorithm, the main machine operation time will be consumed in the stage of generating frequent item sets. If the irrelevant items with small weight are cut off in the early stage of generating frequent sets, the time complexity of the algorithm can be effectively reduced. Through the introduction of weighted association rule model, classification collaboration can be realized, so as to improve the ability of e-commerce multi-attribute data classification.

2 Design of Collaborative Classification Method for Multi-attribute Data in E-Commerce

From a mathematical point of view, e-commerce multi-attribute data classification is a process of mapping. It maps the data of unspecified categories to the original categories. The mapping can be one-to-one or one to many, because in some cases, a product can be associated with multiple categories, which can be expressed as follows:

$$f : A -> B \tag{1}$$

Where a is the set of commodities to be classified and B is the set of categories in the classification system. The mapping rule of e-commerce multi-attribute data classification is that the system summarizes the classification rules according to the information of each classification in the analyzed samples, and then establishes the discrimination formula and rules. When the new data comes, according to the discrimination rules, determine the commodity related categories.

2.1 Building Weighted Association Rule Model

The essence of association rule mining is to find hidden patterns or causal relationships between projects in a large number of complex information data carriers. Because the theoretical basis of association rule technology is easy to understand and accepted by people, and the extensive and far-reaching application space in the future, a large number of researchers have conducted extensive research and improvement on it, and algorithms with representative ideas of the times have emerged. The weighted association rule model can be described as: set D as transaction database, transaction number as N, I as all item sets in the database, and the weight set corresponding to I as W, where w_j indicates the importance of project i_j. $Support(X)$, $Confidence(X)$ support and confidence of data sample x, respectively. w min sup is the weighted support threshold. The form of the weighted association rules discussed is:

$$X \Rightarrow Y \tag{2}$$

If there are association rules in formula 2 in e-commerce multi-attribute data set D, the support degree is the percentage of the number of data containing $(X \cup Y)$ in the total number of transaction database d, that is, the probability of occurrence of event $(X \cup Y)$ is:

$$Support(X \Rightarrow Y) = P(X \cup Y) = \frac{N_{X \cup Y}}{N_{total}} \tag{3}$$

Where and respectively represent the number of tuples and the total number of tuples containing X and y. Then the weighted support degree of weighted association rules is:

$$W\ \sup(X \Rightarrow Y) = \left(\sum_{i_j \in X \cup Y} w_j \right) \times Support(X \Rightarrow Y) \tag{4}$$

If the weighted support of itemset x satisfies the condition in formula 5, it is called weighted frequent itemset.

$$W\ \sup(X) \geq w\min \sup \tag{5}$$

In addition, external confidence is a determinacy measure used to express the validity of association rules. If rule $X \Rightarrow Y$ of transaction data set D is used, confidence *Confidence* is defined as the ratio of the number of transactions X and Y occur at the same time in D to the number of transactions x only occur, that is, conditional probability $P(Y|X)$ is:

$$Confidence(X \Rightarrow Y) = P(Y|X) = \frac{N_{X \cup Y}}{N_{total}} \tag{6}$$

The principle and steps of the proposed model algorithm for weighted association rules are as follows: Taking the user's one-time login as the transaction division, the user resource download transaction table is generated according to the recent download record table. Recently, the storage structure of the download record table is user ID, login time, user questions, query time, downloaded documents, download time. Then, the resource weight table is generated according to the recently downloaded resource scale and article description table, and the coverage of frequent itemsets set by users is accepted to generate the minimum support min*Support*. According to resource weight table and user resource download transaction table, weight association rules are generated by frequent item set discovery based on Apriori algorithm. In the first step, scan the database, search the maximum transaction length size in the database and return the value. In the second step, access resource weight table and generate resource weight sequence in descending order. In the third step, according to the returned result size and Weight set w calculation:

$$\Delta = \frac{1}{\sum_{j=1}^{Size} w_j} \tag{7}$$

Minimum weighted support is generated from min*Support* and Δ. In the fourth step, Apriori algorithm is used to generate frequent sets with min*Support* as the minimum support degree, and the weighted support degree of item set is filled. *w*min*Support* is used as the minimum weighted support to filter the weighted frequent item set, and weighted association rules are generated based on the weighted frequent set.

2.2 Mining E-Commerce Data

According to the user's behavior, such as user registration information, user rating data, and user browsing behavior, the collaborative classification method of e-commerce multi-attribute data establishes the user's behavior interest model. Therefore, it is necessary to mine the corresponding e-commerce user information data. The data mining process is shown in Fig. 1.

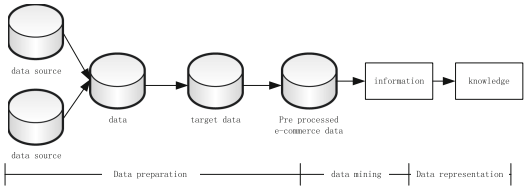

Fig. 1. Data mining flow chart of e-commerce

The user data used mainly includes the following three categories: user profile, user browsing record and user behavior characteristics. A real-time acquisition program is installed on the e-commerce platform, which stores all the running data in the platform into the memory, and takes it as the data sample of e-commerce multi-attribute data collaborative classification.

In order to ensure the effectiveness of the collaborative classification results of e-commerce multi-attribute data, and reduce the time consumed by classification work, the preliminary e-commerce data is cleaned. According to the different data quality problems, the cleaning of e-commerce data includes three parts: the cleaning of removing advertising words and commodity titles. The rule-based method is used to find out the advertising words in the product information and remove them. This step is an offline processing step. We design a rule-based method to clean the advertising words. Therefore, we establish a rule base of advertising words and clean the advertising words in the product information based on the rule base. All the rules in the rule base are described in the form of regular expressions. According to the different categories of advertising words, we divide the rules in the rule base into three categories, which are characteristic word rules, specific part of speech combination and rules of relations between goods. For goods in different categories, except for advertising words, other similar rules are relatively small, among which the rules in the first and second categories are added to the rule base by artificial settings, The third kind of rules are obtained by machine statistical learning. There are still some data quality problems in the product title after removing the advertisement words, such as the emergence of special punctuation, the abnormal segmentation and combination of useful information of the product, the repetition and contradiction of the product information, etc., After removing the advertisement words, the result is "10 times light changing 920000 screen Nikon digital camera s8100v/s9100 is better than s8000s8200", and the symbol "V" should replace s8000s8200 with a space, which is an abnormal combination phenomenon and contradicts the description of s8100s9100. Regular expression is used to

eliminate the problems of special punctuation and abnormal segmentation and combination of commodity information. At the same time, the relative position of words in commodity title is recorded, so that the attribute tuples matching in part of speech tagging have different weights, so as to reduce the interference of commodity information conflict on entity recognition.

2.3 Determine the Multi-attribute Data Classification Index of E-Commerce

There is no unified pattern and absolutely effective classification method for multi-attribute data classification. According to different target enterprises, multi-attribute data classification can be carried out. In e-commerce enterprises, the most prominent characteristics of data are the massive storage of data information, the fast update of data, and the dynamic change of database. In addition, the data behavior data in e-commerce model often has the characteristics of high dimension, many variables and incompleteness. Based on the operation and storage characteristics of e-commerce multi-attribute data, the corresponding classification index is determined, and the classification index system is shown in Fig. 2.

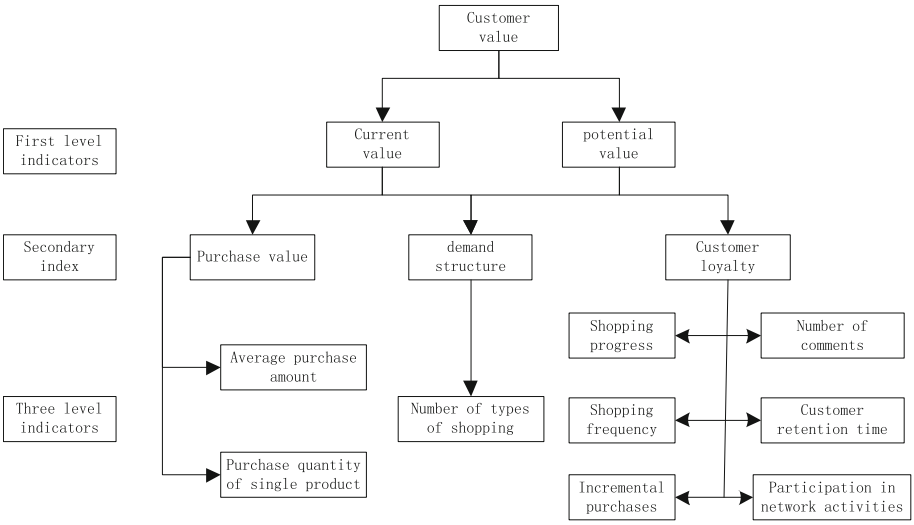

Fig. 2. Multi attribute data classification index system

In the multi-attribute data classification index system of e-commerce, network activity participation is a qualitative index, which is measured by two dimensions of high and low network activity participation, and other indexes are quantitative indexes.

Project refers to the document resources in e-commerce platform. The novelty of resources is the most important parameter to attract users. That is to say, the description time of resources is selected as the parameter to generate the weight of resources. The specific project weight is set as follows: set the resource sequence in the resource recent

download scale as $(d_1, d_2, \cdots, d_n)$, the corresponding description time series is $(t_1, t_2, \cdots, t_n)$. Suppose $T_{\max}$ is the latest time in the description time series, $T_{\min}$ is the oldest time,Then the calculation formula of the weight value of resource d_n is:

$$\eta_j = a + (1+a) * \frac{(t_j - T_{\min})}{(T_{\max} - T_{\min})} \tag{8}$$

Among them, α is a constant, and the specific value can be determined according to the requirements of e-commerce platform.

E-commerce data sets usually have different kinds of attributes, including character attributes and numerical attributes. The numerical attributes can be divided into sequential attributes, discrete value attributes and continuous value attributes. Character class attributes are usually external categories of key values. Numerical attribute is the quantitative record of variables, in which the order attribute is to arrange the key values in order and express the order with numbers, and the discrete value attribute is the discrete value key value without operational significance, while the continuous value attribute is the most common numerical attribute. Table 1 shows the list of attribute words of e-commerce data.

Table 1. List of attribute words of e-commerce data

E-commerce data	Attribute word			
	Acer	Adapter	……	Zoom
X_1	valueser [1]	valueser [2]	……	valueser[n]
X_1	valueser [1]	valueser [2]	……	valueser[n]
……	valueser [1]	valueser [2]	……	valueser[n]
X_1	valueser [1]	valueser [2]	……	valueser[n]

Improve the relevance between projects in order to predict users' rating of projects more accurately. The revised formula for calculating the project forecast score of users is:

$$P_{a,p} = \bar{R}_p + \frac{\sum_{n \in MAI} AC(p,n)\left(R_{a,n} - \bar{R}_n\right)}{\sum_{n \in MAI} AC(p,n)} \tag{9}$$

$AC(p,n)$ is the confidence level of association rules between item P and item n. Mai is the most recently associated set of items for item P.

2.4 Install Data Collaborative Filtering Classifier

Because there are some differences in the number of positive and negative samples in the small classifier, the classifier chooses the class weighted association rule model classifier. According to the classification principle of weighted association rule model,

a set of training samples is given l, Training sample (x_i, y_i) with space dimension D, according to the data attribute association relationship of e-commerce represented by formula 10, two kinds of data are classified based on the mining data samples.

$$H : \omega \cdot x + b = 0 \tag{10}$$

The specific data classification process can be expressed as follows:

$$\begin{cases} H_1 : y = \omega \cdot x + b = +1 \\ H_2 : y = \omega \cdot x + b = -1 \end{cases} \tag{11}$$

Where ω is the reciprocal of the distance from H_1 to h, when the value meets condition 12, the sample can be separated accurately, that is to say, it meets the following requirements:

$$y_i[(\omega \cdot x_i) + b] - 1 + \xi_i \geq 0 \tag{12}$$

ξ_i is the relaxation factor. Penalty factor C is introduced into the mathematical model of the classifier, and the classification function is as follows:

$$f(x) = sign\left(\sum y_i C(x_i, x) + b\right) \tag{13}$$

In the data collaborative filtering classifier, the penalty parameters are divided into C_+ and C_-, the corresponding penalty factors are positive and negative.

2.5 Implement Collaborative Classification of E-Commerce Multi-attribute Data

The mining and processing of e-commerce data is input into the weighted association rule collaborative classifier, and the similarity between the electronic data and each attribute is calculated respectively, so as to get the multi-attribute data collaborative classification results of e-commerce. The similarity calculation formula is:

$$sim(u, v) = \frac{\sum_{i \in I_{uv}} (R_{ui} - \bar{R}_u)(R_{vi} - \bar{R}_v)}{\sqrt{\sum_{i \in I_{uv}} (R_{ui} - \bar{R}_u)^2} \sqrt{\sum_{i \in I_{uv}} (R_{vi} - \bar{R}_v)^2}} \tag{14}$$

Where $sim(u, v)$ represents the similarity between u and V of e-commerce data, R_{ui} and R_{vi} represent the attributes of data u and V to e-commerce data I respectively, $\bar{R}_u$ and $\bar{R}_v$ set of items for data u and V. Set the threshold value of e-commerce multi-attribute data collaborative classification as χ, if the calculation result in formula 14 is greater than the threshold value, then the e-commerce data u and V have the same attribute and can be divided into the same category, otherwise, calculate the similarity of the next group of data until all the e-commerce data mined are classified.

3 Comparative Experimental Analysis

The purpose of the experiment is to test the performance of the algorithm and algorithm under a single minimum support degree, the performance of the classification method under the multi weighted association rule model, and compare the test results under different environments, and analyze the data classification results before and after the application of the multi-attribute data collaborative classification method in e-commerce.

3.1 Experimental Environment and Preliminary Preparation

The operating environment of the experiment is window xp operating system, inter (R) core (TM) 2 Duo T6500 (2.10 GHz) CPU, 2G memory, written in C++ language. In the experiment, IBM data generator was used to generate different data sets with different transaction number, different project number and different average transaction width under XP system. The parameters of each group of experiments are the same except that the contrast parameters are variable. IBM is a classic data set synthesis tool, which is used to generate standard experimental data in association rule mining research. Due to the real-time change of e-commerce data, in order to reflect the collaborative design of classification methods, the implementation environment of e-commerce multi-attribute data collaborative classification method based on weighted association rule model is different e-commerce platforms.

Because the weighted association rule model is applied in the design of e-commerce multi-attribute data collaborative classification method, the relevant operation parameters of the model need to be set, and the model setting interface is shown in Fig. 3.

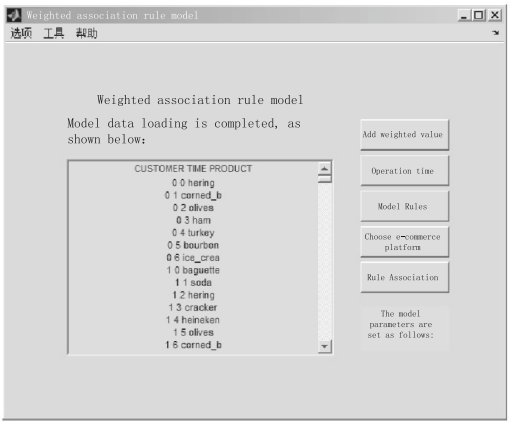

Fig. 3. Parameter setting interface of weighted association rule model

3.2 Experimental Process

Set the accuracy rate as the measurement index of the classification method, and the solution method of the index is the coincidence rate of the set e-commerce classification data and the classification method output data. In order to form the experimental comparison, the traditional e-commerce multi-attribute data collaborative classification method is set as the experimental comparison method and applied to the same e-commerce platform. Through the real-time collection and classification of e-commerce multi-attribute data, the classification results are output, and the experimental results about the accuracy rate are calculated.

3.3 Analysis of Comparative Experimental Results

Under the environment of clothing e-commerce, the experimental results about the accuracy rate of classification obtained through the statistics and statistics of data are shown in Table 2.

Table 2. Experimental results of clothing e-commerce environment classification

Data set	Class	Dataset size/MB	Traditional e-commerce multi-attribute data collaborative classification method			Design Collaborative classification method of multi-attribute data in E-commerce		
			Wrong score/MB	Accuracy %	Total accuracy %	Wrong score/MB	Accuracy %	Total accuracy %
1	A	25	0	100	76	0	100	100
	B	25	12	52		0	100	
2	A	100	1	99	63.5	3	97	98.5
	B	100	72	28		0	100	
3	A	300	2	99.3	68.2	6	98	99.0
	B	300	189	37		0	100	
4	A	750	6	99.2	66.2	19	97.5	98.7
	B	750	501	33.2		0	100	
5	A	1500	7	99.5	66.4	32	97.9	98.9
	B	1500	1000	33.3		0	100	

It can be seen from the data in Table 2 that there are certain differences in the accuracy of the two e-commerce multi-attribute data collaborative classification methods under different data sets. The average classification accuracy of the traditional classification method is 68.06%, while that of the design method is 99.02%, which is 30.96% higher than that of the design method.

In the same way, the collaborative classification results of e-commerce multi-attribute data are obtained under the food e-commerce platform, as shown in Table 3.

Table 3. Classification experiment results of clothing e-commerce environment

Data set	Class	Dataset size/MB	Traditional e-commerce multi-attribute data collaborative classification method			Design Collaborative classification method of multi-attribute data in E-commerce		
			Wrong score/MB	Accuracy %	Total accuracy %	Wrong score/MB	Accuracy %	Total accuracy %
1	A	25	8	68	84	0	100	100
	B	25	0	100		0	100	
2	A	100	36	64	82	3	97	98.5
	B	100	0	100		0	100	
3	A	300	126	58	79	6	98	99.0
	B	300	0	100		0	100	
4	A	750	274	63.5	81.7	19	97.5	98.7
	B	750	0	100		0	100	
5	A	1500	571	62	81	32	97.9	98.9
	B	1500	0	100		0	100	

Through the calculation of the data in Table 3, the average classification accuracy of the two multi-attribute data collaborative classification results is 81.54% and 99.02% respectively, compared with the classification accuracy of the design method increased by 17.48%. By synthesizing the classification results of multi-attribute collaborative data in different e-commerce environments, it is found that the design method can stabilize the classification accuracy of data above 99%, so it has high application performance.

4 Conclusion

It can be seen from the above that with the growing e-commerce market, the increasing number of commodities and the increasingly diversified levels of participants, it is increasingly difficult to provide valuable information for users. At this time, it is necessary to classify all kinds of goods in different levels to extract information intelligently from e-commerce market. It can quickly query the corresponding commodities for both sides of the market transaction, determine the purchase scheme of commodities, and then complete the placement strategy of commodities and recommend the commodities that may be of interest to users. The extraction of these information is completed on the premise of classification. The above classification methods will effectively improve the classification efficiency and accuracy of e-commerce market commodity data, and better serve businesses and customers.

References

1. Cheng, C.H., Chen, C.H.: Fuzzy time series model based on weighted association rule for financial market forecasting. Expert Syst. **4**(35), 110–115 (2018)
2. Cagliero, L., Garza, P., Kavoosifar, M.R., et al.: Discovering cross-topic collaborations among researchers by exploiting weighted association rules. Scientometrics **116**(2), 1273–1301 (2018)
3. Subbulakshmi, B., Deisy, C.: An improved incremental algorithm for mining weighted class-association rules. Int. J. Bus. Intell. Data Min. **13**(3), 291–308 (2018)
4. Murugan, I., Nabhan, A.R., Subramanian, A.: A weighted association rule mining method for predicting HCV-human protein interactions. Curr. Bioinform. **13**(1), 73–84 (2018)
5. Liu, S., Yang, G. (eds.): ADHIP 2018. LNICST, vol. 279. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-19086-6
6. Fernandes, D.S.F., Domingues, M.A., Vaccari, S.C., et al.: Latent association rule cluster based model to extract topics for classification and recommendation applications. Expert Syst. Appl. **112**(6), 34–60 (2018)
7. Liu, H., Yang, S., Gou, S., et al.: Terrain classification based on spatial multi-attribute graph using polarimetric SAR data. Appl. Soft Comput. **68**(24–38), S1568494618301510 (2018)
8. Fu, W., Liu, S., Srivastava, G.: Optimization of big data scheduling in social networks. Entropy **21**(9), 902 (2019)
9. Gou, J., Hou, B., Ou, W., et al.: Several robust extensions of collaborative representation for image classification. Neurocomputing **348**(5), 120–133 (2019)
10. Liu, S., Lu, M., Li, H., et al.: Prediction of gene expression patterns with generalized linear regression model. Front. Genet. **10**, 120 (2019)
11. Xiao, H.-g., Deng, G.-q., Wen, T.A.N., et al.: A weighted association rules mining algorithm based on matrix compression. Meas. Control Technol. **37**(3), 10–13 (2018)
12. Gupta, K.O., Chatur, P.N.: Gradient self-weighting linear collaborative discriminant regression classification for human cognitive states classification. Mach. Vis. Appl. **31**(3), 1–16 (2020)
13. Liu, S., Liu, D., Srivastava, G., Połap, D., Woźniak, M.: Overview and methods of correlation filter algorithms in object tracking. Complex Intell. Syst. 1–23 (2020). http://doi.org/10.1007/s40747-020-00161-4