IOP Conference Series: Earth and Environmental Science

PAPER • OPEN ACCESS

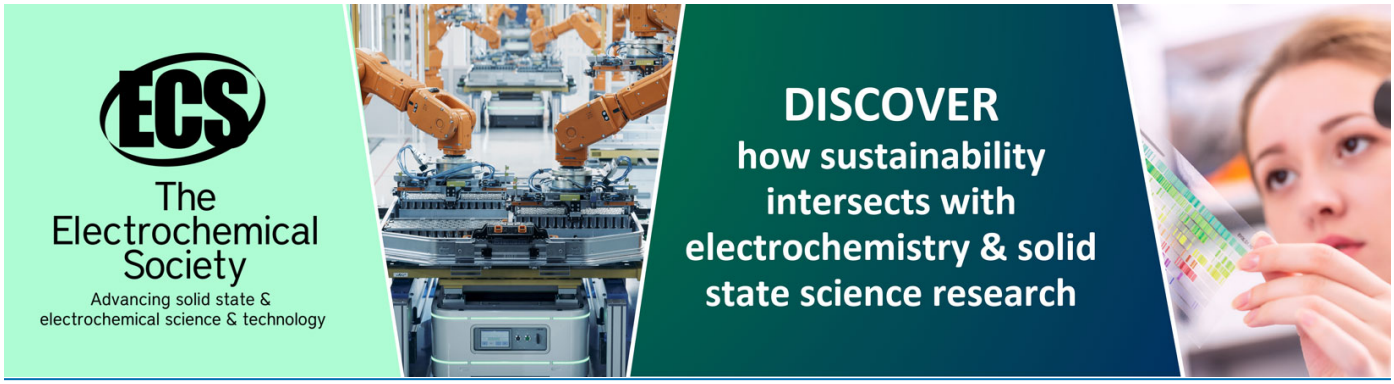
Research on Classification Mining of Tank Driving Simulation Training Data

To cite this article: Qing Deng *et al* 2021 *IOP Conf. Ser.: Earth Environ. Sci.* **693** 012063

View the article online for updates and enhancements.

You may also like

- Identifying ADHD boys by very-low frequency prefrontal fNIRS fluctuations during a rhythmic mental arithmetic task
Sergio Ortuño-Miró, Sergio Molina-Rodríguez, Carlos Belmonte et al.
- DETECTING ACTIVE GALACTIC NUCLEI USING MULTI-FILTER IMAGING DATA. II. INCORPORATING ARTIFICIAL NEURAL NETWORKS
X. Y. Dong and M. M. De Robertis
- GALAXY MODELING WITH COMPOUND ELLIPTICAL SHAPELETS
James Bosch



This content was downloaded from IP address 3.145.93.136 on 08/05/2024 at 19:16

Research on Classification Mining of Tank Driving Simulation Training Data

Qing Deng*, Yaxin Tan, Kai Zhai and Weizhen Luan

Training Center, Academy of Army Armored Forces, Beijing 100072, China
Email: 154247597@qq.com*

Abstract. Training with tank driving simulator is an important method to improve equipment operation skills. In view of the deficiency that it is difficult to find knowledge and rules from complex training data by statistical analysis method in the past driving simulation training, this paper proposes CSAGA-LSSVM algorithm to analyze tank driving simulation training data. Selecting key points to quickly generate shapelets, reducing the number of candidate shapelets; combining shapelets according to distance and time interval to enhance the ability of feature identification; designing adaptive genetic algorithm to dynamically adjust the probability of crossover and mutation to find the optimal parameter solution of least squares support vector machine. The algorithm is applied to the classification mining of shift operation data from a certain tank driving simulator to extract the operation characteristics of personnel.

1. Introduction

Tank driving simulation training is an important way for armored forces to master driving skills and realize the organic combination of personnel and equipment. It is of great significance to improve the rapid mobility ability of armored units in battlefield and maintain high-efficiency combat capability[1,2]. After tank driving simulation training, a large amount of data will be generated, including training operation data, trainees data, training subject data, training result data, etc. These data have the characteristics of large scale, high dimension and non-linearity, and there are various complex relationships among them[3]. In the traditional analysis of tank driving simulation training results, it is mainly human based statistical analysis[4,5]. It is easy to be affected by the professional knowledge and personal preference of the analysts, so the influencing factors of training are not fully considered. It is unable to accurately guide trainees to train tank driving operation skills and find valuable training rules from these complex data. In order to solve this problem, classification mining is introduced into the analysis of tank driving simulation training data in order to obtain training guidance rules.

Bayesian network[6] is an important classification mining method. The directed graph model is used to describe and calculate the probability dependence relationship between variables. The posterior probability is updated by prior knowledge, and finally the classification is solved. However, the learning process of network structure and parameters is complex, and experts often need to determine the initial value. In addition, in many cases, the prior probability of tank driving simulation training data can not be determined in advance. There is a certain coupling between attributes, which does not meet the premise of Bayesian network. Decision tree classification has the advantages of simple principle, strong noise resistance, few parameter settings, and easy to understand[7]. It is mainly applicable to the classification of small sample data sets. When faced with massive high-dimensional and nonlinear data samples, it is easy to generate a large number of invalid internal nodes and branch structures for classification, which leads to low efficiency of decision tree

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

construction and too complex form of tree. It is difficult to find the classification knowledge hidden among multiple related attributes. Deep learning is an important research direction of data classification mining in recent years, which has been widely used in the fields of image recognition and natural language understanding. Through the combination of multi-layer nonlinear modules, it realizes the representation of complex mapping relationship[8]. Least squares support vector machine[9,10](LSSVM) processes high-dimensional and nonlinear data by kernel function and feature space, realizes the mapping from sample space to feature space, transforms quadratic programming into linear equation solution. However, the training data of tank driving simulation has the characteristics of time sequence, which can not be directly input into LSSVM for classification. Based on the characteristics of LSSVM, shapelets are introduced and combined to extract the features of original data. Adaptive genetic algorithm (AGA) is designed to realize the optimal selection of super parameters. It is applied to analyze the operation data of tank driving simulation training.

2. CSAGA-LSSVM Classification Mining Algorithm

Time series and its subsequences: $TS = \{vt_1, vt_2, ..., vt_i, ..., vt_n\}$ is a real value time series, in which vt_i is the time series point, n is the time series length, can also be recorded as $|TS|$. Mostly the adjacent time series points are equidistant. Time sequence monotonically increases and is also abbreviated as $TS = \{v_1, v_2, ..., v_i, ..., v_n\}$. The subsequence $TS_{j,l} = \{vt_j, vt_{j+1}, ..., vt_{j+l-1}\}$ of time series TS is a continuous sequence whose starting position is j and length is l.

Shapelets: the feature representation of time series which can determine the class label of the sequence to the greatest extent, is also the most recognizable and interpretable local temporal pattern. Shapelets= $\{s_1, s_2, ..., s_k\}$ is a subsequence of the time series $TS = \{v_1, v_2, ..., v_i, ..., v_n\}$.

2.1. Fast Acquisition of Shapelets Based on Key points

The key points should be generated from the interior of time series points, expressing the main characteristics and change trend of time series. Based on this, the key points include the starting point, the end point, the step point and the extreme point .

The starting point and ending point are located at both ends of the time series, which clearly indicate the start time and end time of the whole time series. In the tank driving simulation training, it can reflect the time taken to complete a certain action. Step point refers to that the slope of a certain point and its adjacent two points have a large difference. The corresponding intermediate operation conversion time can be regarded as the step point.

$$(v_{i+1} - v_i^{\text{step}}) / (v_i^{\text{step}} - v_{i-1}) \geq \rho \tag{1}$$

Extremum points include local maximum point and minimum point, which satisfy the following conditions respectively

$$v_{i-k} \leq v_i^{\text{lmax}},\ v_{i+k} \leq v_i^{\text{lmax}} \quad \text{or} \quad v_i^{\text{lmin}} \leq v_{i-k},\ v_i^{\text{lmin}} \leq v_{i+k} \tag{2}$$

Where k is the size of the neighborhood near the extreme point.

2.2. Feature Generation of Combined Shapelets

Through the tank driving simulator, we can organize various speed driving exercises. They have a clear sequence and corresponding time interval during the operation of the throttle, clutch, brake and other components. How to accurately identify the driving actions of trainees and distinguish different training levels is a multi-dimensional time series classification problem. According to the single shapelets obtained in the previous section, it is difficult to achieve accurate identification effect because it is easy to ignore the operation time and logical combination between different sequences. Therefore, it is proposed to combine multiple shapelets and add the time interval between different shapelets as the basis of classification to enhance the identification ability of combined shapelets. Also retain the optimal combination shapelets through information gain evaluation, so as to improve the accuracy of classification.

The Euclidean distance between corresponding time series points is used to measure the difference between TS_1 and TS_2, as shown in Equation (3).

$$dis(TS_1, TS_2) = \sqrt{\sum_{i=1}^{n} [TS_1(i) - TS_2(i)]^2} \tag{3}$$

For the distance between shapelets subsequence s and TS_i, due to the inconsistency of their time lengths, a sliding window method is used to generate $(n-s+1)$ series $TS_{i,|s|}$ with the same length as subsequence s. The distances between these equal length sequences and s are calculated according to Equation (4), then the minimum distance is taken as the measurement between subsequence s and TS_1 combining with dynamic bending distance comparison principle.

$$shpdis(s, TS_i) = \min(dis(s, TS_{i,|s|})) \tag{4}$$

the number of class labels is C. If a certain class c_i has n_i time series in DTS, then the entropy of time series data set is expressed as follows:

$$E(DTS) = -\sum_{i=1}^{C} \frac{n_i}{m} \log_2 \frac{n_i}{m} \tag{5}$$

After obtaining the shapelets of each time series in the previous section, the shapelets are combined in pairs. s1 and s2 are randomly selected to perform the merging operation, namely $s1 \wedge s2$. According to equation (5), calculate the distance between s1,s2 and each time series in DTS. For the $s1 \wedge s2$ contains the characteristics of s1 and s2, the distance $comshpdis(s_1 \wedge s_2, TS_i)$ between $s1 \wedge s2$ and TS_i should be considered combination with the distance between s1,s2 and TS_i.

$$comshpdis(s_1 \wedge s_2, TS_i) = \max(shpdis(s_1, TS_i), shpdis(s_2, TS_i)) \tag{6}$$

Compare the difference between $SItime$ and time interval threshold θ, then the time series data set DTS_{left}, DTS_{right} are divided. Furthermore, the information gain $I(s_1 \wedge s_2, \theta)$ is calculated by equation (6) and (10), then the subsequence with the largest information gain is selected as the final shapelets by recursive solution.

$$I(s_1 \wedge s_2, \theta) = E(DTS_{left}) - \frac{|DTS_{left}{}^1|}{|DTS_{left}|} E(DTS_{left}{}^1) - \frac{|DTS_{left}{}^2|}{|DTS_{left}|} E(DTS_{left}{}^2) \tag{7}$$

2.3. LSSVM Super Parameter Optimization Based on Adaptive Genetic Algorithm

LSVM is used for classification. The value of kernel parameter σ and penalty factor *C* have important influence on the performance of classification algorithm. Therefore, adaptive genetic algorithm (AGA) is used to improve crossover and mutation operations to ensure population diversity and avoid falling into local optimum at the early stage of evolution.

Two parents were randomly selected by single point crossover operator, and the gene position of each pair of individuals was set for crossover operation, With the increase of evolutionary algebra, the crossover probability is automatically adjusted according to equation (8).

$$p_c = \frac{\exp(-0.5\tau)}{psize \cdot \sqrt{len}} \tag{8}$$

Mutation operation promotes the generation of new individuals to improve the local search ability of the algorithm. according to the change of individual fitness, the mutation probability is dynamically adjusted, and the individual with smaller fitness is given a larger mutation probability to promote the individual to evolve to a better solution.

$$p_m = \begin{cases} p_{m0} - \dfrac{\lambda(fit_{\max} - fit_i)}{fit_{\max} - fit_{avg}}, & f_i \geq f_{avg} \\ p_{m0} + \dfrac{\lambda(fit_{avg} - fit_i)}{fit_{avg} - fit_{\min}}, & f_i < f_{avg} \end{cases} \tag{9}$$

2.4. AGA-LSSVM Classification Based on Combined Shapelets

$x_k \in R^n$ is the input attribute data transformed by shapelets, $y_k \in R$ is the corresponding class label, LSSVM transforms the classification problem into an optimization problem with constraints in equation (10).

$$\min \phi(\boldsymbol{\omega}, \boldsymbol{e}) = \frac{1}{2}\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\omega} + \frac{1}{2}C\sum_{k=1}^{N} e_k^2, \quad s.t.\ y_k[\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\varphi}(\boldsymbol{x}_k) + b] = 1 - e_k \tag{10}$$

Where $\boldsymbol{\omega}$ is the weight matrix, b is the bias variable, e_k is the training error, C is the penalty factor.

The optimization problem of equation (10) is obtained by Lagrange functional:

$$L(\boldsymbol{\omega}, b, \boldsymbol{e}, \boldsymbol{a}) = \phi(\boldsymbol{\omega}, \boldsymbol{e}) - \sum_{k=1}^{N} a_k \left(y_k[\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\varphi}(\boldsymbol{x}_k) + b] - 1 + e_k \right) \tag{11}$$

According to equation (11), the classification function for any input sample $\boldsymbol{x}$ is as follows:

$$\hat{y} = \sum_{j=1}^{N} \alpha_j y_j k(\boldsymbol{x}, \boldsymbol{x}_j) + b \tag{12}$$

Through the above description of the four stages of CSAGA-LSSVM algorithm, the specific process is given below, as shown in Figure 1.

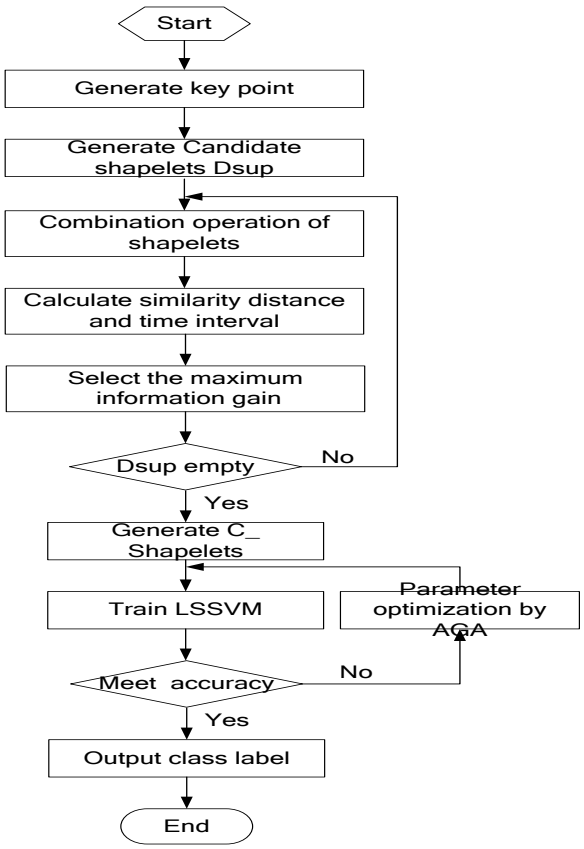


Figure 1. Flow chart of CSAGA-LSSVM algorithm

3. Operation Data Analysis of Tank Driving Simulation Training Based on CSAGA-LSSVM

3.1. Data Sources

In order to facilitate the analysis of tank driving simulation training results, through the displacement sensor and photoelectric sensor installed on a certain tank driving simulator, the operation data of the trainees are recorded in real time. The frame rate of the simulator system is 25FPS, and the data is collected once every frame running, that is, 25 groups of data are collected per second. During tank driving simulation training, in addition to the above operation data, there are also data generated by tank driving simulator system operation, including training subjects, engine speed, speed, position and other data .

In the experiment of tank driving simulation training, 120 trainees (including 30 second-class tank drivers, 30 third-class tank drivers, 30 junior tank drivers and 30 non-level personnel) were selected to carry out shift operation training in a certain type of tank driving simulation classroom. Each trainee carried out three shift operations, and collected driver's operation through sensors and simulation training system. The operation data of tank driving simulator is shown in Table. 1.

Table 1. Operation data of tank driving simulator

Number	Simulator parts	Sensor type	Value range	Output
1	gear shift	Displacement	{-1,0,1,2,3,4,5}	Switching value
2	refueling	Displacement	[0,100]	Analog quantity
3	brake pedal	Displacement	[0,100]	Analog quantity
4	clutch pedal	Displacement	[0,100]	Analog quantity
5	control lever	Displacement	[0,100]	Analog quantity

3.2. Obtaining the Optimal Classification Solution

80% of the data is randomly selected as the training sample set, and the remaining 20% of the data is used as the test sample set to input the classification model. Firstly, the multi-dimensional time series data of shifting operation are represented by shapelets feature extraction method, and all the optimal combination shapelets are solved. Then, the distances between the unitary time series data such as accelerator pedal displacement, clutch pedal displacement, brake pedal displacement and gear position are calculated in turn, so as to realize shapelets transformation of original shift operation data. Then the adaptive genetic algorithm is used to obtain the optimal super parameters of LSSVM $C = 60.5, \sigma = 0.23$. Based on the optimized classification model, the classification results of the shift operation data are obtained. For example, the classification result of the accelerator pedal displacement time series of a training sample is y_{acr}, the classification result of clutch pedal displacement time series is y_{clu}, and the classification result of brake pedal displacement time series is y_{brk}, The result of classification is y_{gear}, and the mode of statistical classification results is used as the class label y of the sample. Results are shown in Table 2.

Table 2. Result data

Class label number	Class tag value	Corresponding to the optimal shapelets value
1	qualified	$Shapelets_{acr}$={00,01,00},{00,01} $Shapelets_{clu}$={000,010,011,110,111},{111,100,011,010,000} $Shapelets_{gear}$={1,2}
0	unqualified	$Shapelets_{brk}$={00,01,11} $Shapelets_{acr}$={00} $Shapelets_{gear}$={1}
0	unqualified	$Shapelets_{acr}$={00,01,00},{00,01} $Shapelets_{clu}$={000} $Shapelets_{gear}$={1}

3.3. Result Analysis

(1) Shapelets corresponding to qualified operation results of first gear to second gear: accelerator pedal time series data $shapelets_{acr}$ [1] ∧ $shapelets_{acr}$ [2] = {00, 01, 00} ∧ {00, 01}, clutch pedal displacement time series data $shapelets_{clu}$ [1] ∧ $shapelets_{clu}$ [2] ={000, 010, 011, 110, 111}∧{111, 100, 011, 010, 000}, gear time series data $shapelets_{gear}$ = {1, 2}, It means that the accelerator pedal should be steadily pressed to the appropriate position before shifting, and then released. At the same time, the clutch pedal should be stepped from the initial position to the maximum position, and the gear lever should be pushed from the first gear to the second gear quickly. After shifting, the clutch pedal should be released to achieve fast front and stable rear, and the accelerator should be stepped down evenly.

(2) Shapelets corresponding to disqualification of first gear to second gear operation result: $shapelets_{brk}$ = {00, 01, 11}, accelerator pedal time series data $shapelets_{acr}$ = {00}, gear timing data $shapelets_{gear}$ = {1}, which means that the brake pedal was pressed in the process of shifting from first gear to second gear, but the accelerator pedal was not stepped on to rush, and the vehicle gear value did not change. The data of accelerator pedal time series data $shapelets_{acr}$ [1] ∧ $shapelets_{acr}$ [2] = {00, 01, 00} ∧ {00, 01}, clutch pedal displacement time series data $shapelets_{clu}$ = {000}, gear time series data $shapelets_{gear}$ = {1}, which means that the clutch pedal is not pressed in the process of first gear to second gear and the shift operation is unqualified.

According to the shapelets combination, the corresponding standard operation mode can be established, which can be used as the evaluation level of the operation skill level of shift from first gear to second gear, to accurately analyze the driving action of the trainees.

4. Conclusion

This paper proposes a classification mining algorithm based on CSAGA-LSSVM which is applied to analyze the tank driving simulation training data.

5. References

[1] STRACHAN I. Development trend of world military simulation training system[J]. DU F, WANG S C, translated. Foreign Tanks, 2014 (4): 34-37.
[2] JIAO K Z, CHENG P Y, LIU T, et al. Research on foreign military virtual training system[J]. Aerodynamic Missile, 2013(06):64-67.
[3] TANG Z W, XUE Q. Data mining in the simulation test of armored forces's landing attack based on decision tree[J]. Journal of Academy of Armored Force Engineering, 2013, 22 (1): 6-9.
[4] ANDREW C. Using the National Training Center instrumentation system to aid simulation-based acquisition[D]. Santa Monica, CA,, US: Pardee Rand Graduate School, 2017.
[5] DWIGHT J. An automated system for the analysis of combat training center information: strategy and development[R]. Alexandria, VA, US: U.S. Army Research Institute for the Behavioral and Social Sciences, 2015.
[6] CHAI H M, ZHAO Y T, FANG M. Parameter learning of Bayesian networks based on prior normal distribution [J]. Systems engineering and electronic technology, 2018, 40(10) : 2370-2375.
[7] YANG R, WANG P, SUO R X. Research on the application of decision tree algorithm in CRM [J]. China Management Informatization, 2019, 22 (15): 53-55. (in Chinese)
[8] WU Y X, WANG J L, YANG L. A review of cost sensitive deep learning methods [J]. Computer science, 2019, 46 (5): 1-12.
[9] YE Q L, XU D P, ZHAGN D. Remote sensing image classification based on deep learning feature and support vector machine [J]. Acta Forestry Engineering Sinica, 2019, 4 (2): 119-125.
[10] HU T, QIU Y P, CUI H J,et al. Numerical discretization-based kernel type estimation methods for ordinary differential equation models[J].Acta Mathematica Sinica, 2015,31(8):1233-1254.