Application of Machine Learning Algorithms in Analysis of Learners' Behaviour Data

Jinjin Liang[#1,2], Yong Nie [*1]

[#]1.*Associate Professor & School of Education & Shaanxi Normal University*
2. Associate Professor & School of Sciences & Xi'an Shiyou University
Address 1. Chang'an South Road，Xi'an，Shaanxi, China, 710062
Address 2. East No. 2 Dianzi Road，Xi'an，Shaanxi, China, 710065

Abstract

Education Informatization is conducive to obtaining the Learners' behaviour data both from the offline traditional classroom by the educators and the network classroom by the online platform. If machine learning algorithms can be designed to reveal the information underneath these behaviour data, it will provide scientific evidences for educators to make wise decisions and design effective teaching strategies. A framework is constructed for applying machine learning algorithms into the Learners' behaviour data, which includes analysing learners' characteristics by Clustering algorithm, constructing a risk assessment model by Support Vector Machine (SVM) and designing an outlier detection model by Support Vector Data Description (SVDD). Utilizing the results derived from those algorithms, the educators can design effective teaching process to match the learners' practical situation, carry out the teaching interventions. Machine learning algorithms provide theoretical foundation for the realization of learner-cantered, individualized, precise and intelligent teaching process.

Keywords —*Education Informatization, Learners' behaviour data, Machine learning algorithms, Analysing learners' characteristics, Risk assessment, Outlier detection.*

I. INTRODUCTION

Education Informatization gives impetus to the innovation of the teaching mode and environment, which brings about new ways of network learning, such as Massive Open Online Courses (MOOCs), mobile learning and blended learning [1]-[2]. Network learning has the merits of low cost and high efficiency, and is convenient for personalized and fragmented learning. However, most of the courses are mainly based on video, and lacks the learning experiences of the interaction and communication with teachers, which is not suitable for deep learning [3]. Blended teaching is proposed based on the "learner-centered" principle, which integrates the dominant advantages of supervising and controlling the offline classroom for the educators, as well as the dominant roles of self-regulated and individualized leaning through online classroom for the learners; it is conducive to better exploring and sharing the value of high-quality educational resources from network, and has become a predictable educational ideological reform trend in the international education circle.

Under the blended teaching mode, a large number of Learners' behaviour data are generated and become part of education data, which includes not only the subjectively behaviours observed by educators, but also the objectively network learning behaviours recorded by the online platforms. Existing researches have been done to improve the teaching quality. Literature [4] introduces MOOC online class to replenish and serve the traditional class in the campus. Literature [5] proposes Small Private Online Course (SPOC), which combines offline answers from teachers with online independent learning and testing from students. Reference [6] designs a SPOC course to carry out the teaching practice, which combines the online and offline teaching to match the actual development of the school. Literature [7] analyses the application basis of learning needs, learning plans, content preparation, measurement and evaluation in typical SPOC case. Existing researches on learning behaviour data include finding out the relation with the learning effects [8], designing data mining models [9] or adaptive learning systems models [10]. Although some accomplishments has been achieved, the researchers are still in the start-up step and the information beneath the data needs to be further exploited.

Excavating the hidden information among those behaviour data is of application study and practical exploration use, it will help the educators to make much wiser and scientific decisions. The development of machine learning provides profound theoretical supports for analysing and classifying those behaviour data. This paper constructs a frame of applying the machine learning algorithms into Learners' behaviour data with respect to the leaners' characteristic, risk assessment and outlier detection. Specifically, the C means clustering algorithm is used to analyse and divide the learners into various and disconnected clusters based on the similarity [11]; Support Vector Machine (SVM)[12]-[13] is introduced to derive the risk assessment model with good generalization and high precision; Support

Vector Data Description (SVDD) is applied to find a hypersphere with minimal radius and reject samples falling outside as outliers[14][15]. The derived results can be used as a scientific basis for matching the educational resources, designing teaching activities and carrying out the teaching intervention.

II. ANALYSIS OF THE LEARNERS' CHARACTER

Learners' behaviour data include knowledge foundation, emotional attitude, learning needs and interaction behaviours etc., they differ significantly from various learners, which have distinct characteristics of Isomerism, highly nonlinear correlation and multi-mode. Analyse learners can help educators better understand the learners' actual situation, and recommend suitable recourses and teaching strategies. The target variables are characterized by learners' emotion, attitude, scores, etc.

Clustering analysis is carried out on education data; it classifies the learners into various groups based on similarity among the education data. Clustering uses the principle of "Intra-group Distance Minimization" and "Inter-group Distance Maximization", It classifies the samples into various groups, where samples in the same class has high similarity and samples in the different class has high differences. The commonly used clustering algorithms are classified into division method, hierarchy method, density method, grid method, and probability model, etc.

C Means (CM) clustering is a widely used density based method, which has the merit of simplicity and speed, and only the initial center is required. CM is used to analyse the learners with respect to the knowledge foundation, learning habit, general characteristic and emotion attitude.

Denote by $X = \{(x_i)\}_{i=1}^{l} (x_i \in R^n)$ the Learners' behaviour data, where x_i is the i-th training data of n attributes. C Means (CM) clustering algorithm aims to divide all the data into c disconnected subsets $X_i (i = 1, \cdots, c)$, satisfying

$$X = \bigcup_{i=1}^{c} X_i, X_i \cap X_j = \Phi (i \neq j) \quad (1)$$

C Means (CM) minimizes the following objective function with a minimal square distance.

$$\min \sum_{i=1}^{c} \sum_{x_k \in X_i} u_{ij} d_{ij}^2$$
$$s.t. \sum_{i=1}^{c} u_{ij} = 1, \forall j = 1, 2, \cdots, n \quad (2)$$

Here, $U = (u_{ij})_{c \times l}$ is the membership matrix, where $u_{ij} = 1$ means the j-th data is assigned to the i-th cluster $Z = \{z_1, z_2, \cdots, z_c\}$ and $u_{ij} = 0$ means the j-th data is not assigned to the i-th cluster; d_{ij} is the Euclidean distance from x_j to the i-th cluster center z_i, which is computed by

$$d_{ij} = d(x_j, z_i) = \| x_j - z_i \| = \sqrt{\sum_{p=1}^{n} (x_j^p - z_i^p)^2} \quad (3)$$

where v_i^p is the p-th attribute of any column vector $v_i = (v_i^1, v_i^2, \cdots, v_i^p, \cdots, v_i^n)^T$.

Select c samples as the initial cluster center; C Means clustering assigns the samples to the nearest clusters; compute the new cluster center and repeat the whole process. Until the objective function is minimized.

III. DATA DRIVEN LEARNING RISK ASSESSMENT MODEL

The learning risks of the learners are influenced by the knowledge foundation, emotional attitude, interaction behaviour. In the blended teaching mode, the information data consists of the offline classroom and the online platform, while the former are collected by the educators based on some predefined system indicators and the latter are recorded by the online platform. If an assessment model can be designed to evaluate the risk of the learners based on the information data of learners' behaviour, which gives dynamic assessment and timely diagnosis based on each factors related with learning, the educators will have objective basis to carry out the teaching intervention to improve the teaching influences.

Support Vector Machine is used to formulate the risk assessment model, and the risk levels will be divided into two or three categories by the educators, such as high - low or high- middle - low. It can even be divided into more categories according to the educators' own needs. SVM is proposed by Vapnik with profound statistic learning theory (SLT) and optimization theory. It has the merit of high generalization ability, high precision and few parameters. Without loss of generality, we use the binary classification case to explain the principle of SVM, with the circles as the positive class and the rectangles as the negative class. When the two classes are separable, SVM figures out the dashed line as the separating plane and uses the solid line to predict the class label for any new data. If the test sample lies above the solid line, it is treated as the positive class; and if the test sample lies below the solid line, it is treated as the negative class. Usually, the data cannot be separated completely; SVM introduces slacks to allow tolerances or called errors for the misclassified points. Figure 1 demonstrates the classification principle in the linear separable and non-separable case.

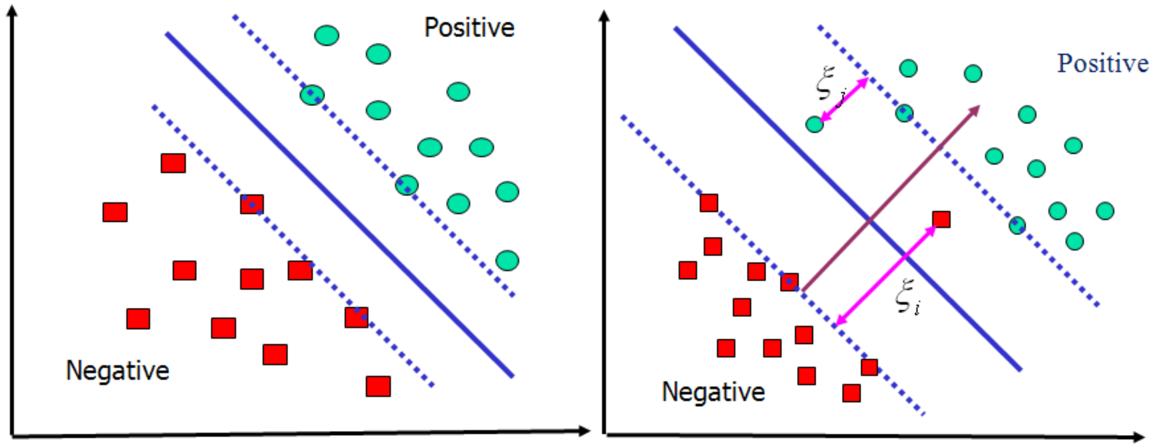

Fig. 1. Classification principle of SVM in the linear separable and non-separable case

In real cases, the data may be more complex, no matter how hard we try, and the data cannot be separated by the linear techniques. SVM introduces the technique of Mercer kernel by using a nonlinear map $\phi : x \to \phi(x)$ to transform data in the original space to a high feature space, so that linear techniques can be used. Figure 2 demonstrates the classification principle in the linear separable

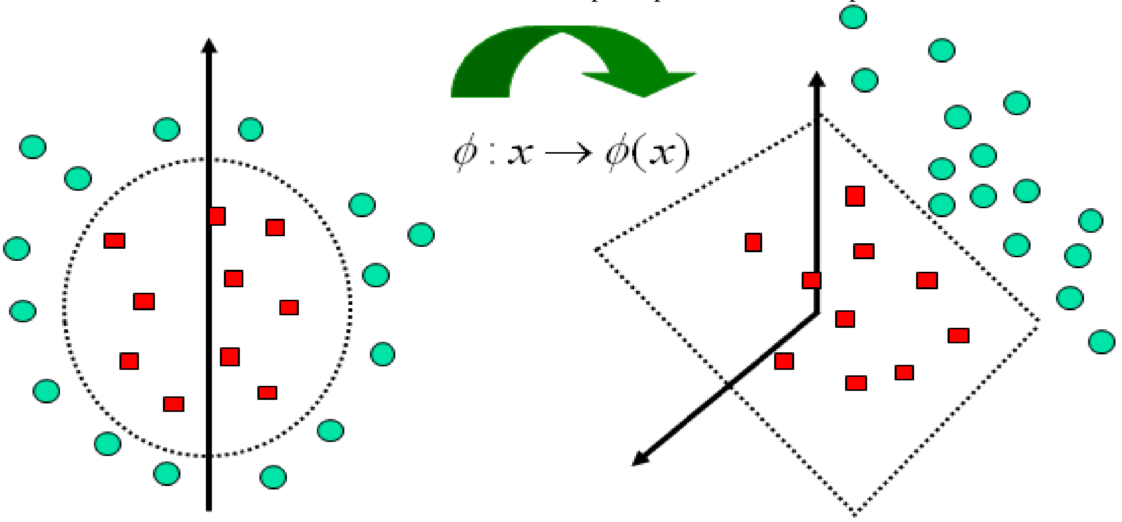

Fig. 2. Classification principle in the non-linear case

Denote by $T = \{(x_1, y_1), (x_2, y_2), ..., (x_l, y_l)\}$ the learners' behaviour data as the training set, where $x_i \in R^n$ is the column vector collected from the teaching process for the i -th learner; y_i is the risk label indicating the level: $y_i = 1$ indicates the high risk and $y_i = -1$ indicates the low risk.

SVM aims to find the optimal separating hyperplane and separates data into two disconnected classes. Introducing a set of slacks $\{\xi_i\}_{i=1}^n (\xi_i \geq 0)$ for each misclassified data to allow errors, and pay a penalty $C > 0$ proportional to the amount of constraint violations; SVM minimizes the following program with "Structural Risk Minimization" principle to get a decision function with good generalization and high precision. The linear separable cases can be formed in the same frame work, and we will illustrate the nonlinear separable case without generality.

$$\min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i$$
$$s.t. \; y_i[w^T \phi(x_i) + b] \geq 1 - \xi_i, i = 1, \cdots, l; \quad (4)$$
$$\xi_i \geq 0, i = 1, \cdots, l.$$

The optimal separating hyperplane to be defined is $w^T x + b = 0$, where $w \in R^n$ is the normal to the separating hyperplane, $b \in R$ is the bias, "T "is the transpose operation, $p_1^T p_2$ is the inner product of p_1 and p_2; $\xi \in R^l$ is the constraint violations or called the error. The linear separating hyperplane is exactly one straight line in the two dimensional space for $x \in R^2$; and it will be one flat surface in the three dimensional space for $x \in R^3$.

The dual technique in optimization theory is utilized to transform the above program into one dual program and figure out the solution. Please refer to reference [16] for detail. Denote the Lagrange multiplier as $\alpha_i \geq 0 (i = 1, \cdots, l)$, the dual program is minimized as follows in SVM.

$$\max_{\alpha} \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$
$$s.t. \ \sum_{i=1}^{l} \alpha_i y_i = 0 \qquad (5)$$
$$0 \leq \alpha_i \leq C, i = 1, \cdots, l.$$

Here $K(x, y) = (\phi(x) \cdot \phi(y)) = \phi(x)^T \phi(y)$ is the kernel function. The commonly used kernels are the linear kernel $K(x, y) = x^T y$, polynomial kernel $K(x, y) = (x^T y + 1)^d$ and Radial basis kernel $K(x, y) = \exp(-\| x - y \|^2 / \sigma^2)$

For the new input data, we use the sign function to predict the risk label for any new data x

$$g(x) = \mathrm{sgn}\{\sum_{i=1}^{l} \alpha_i^* y_i K(x_i, x) + b^*\} \qquad (6)$$

IV. INTRODUCTION

Learners' behaviour data reflect the learners' learning status, which contains a variety of factors, such as knowledge foundation, emotional attitude, learning style, and interaction status. The factors are further differentiated into many attributes such as learners' metacognitive level, motivation interest, learning style, individual differences, self-discipline, question feedback, extended learning, score ranking, preview and review, learning plan, note-taking. All those data are collected and transformed into the column vectors. Analysing those kinds of education data using machine learning techniques, we can detect the abnormal models as outliers, and provide us proof to make wiser decisions.

Support Vector Data Description is a widely used method in outlier detection, which finds out a hypersphere with minimum radius to cover all the target data, and reject the others as the outlier data falling out of the hypersphere.

Denote by $\{x_i\} (x_i \in R^d, (i = 1, ..., n)$ the target data in R^d, which contains the education information. Respectively denote by R and a the radius and center of the sphere, $\{\xi_i\}_{i=1}^{n} (\xi_i \geq 0)$ the slack for the i-th data, SVDD introduce $C (C > 0)$ the penalty to make compromise between the errors and the complexity, and minimizes the following program

$$\min F(r, a, \xi_i) = R^2 + C \sum_{i=1}^{n} \xi_i$$
$$s.t. \ (\phi(x_i) - a)^T (\phi(x_i) - a) \leq R^2 + \xi_i, i = 1, ..., n; \ (7)$$
$$\xi_i \geq 0, i = 1, ..., n.$$

Similar as that in SVM, kernel function is introduced to get more flexible description model without extra computation.

The optimal solution can be obtained through figuring out its dual by introducing Lagrange function and differentiating it with respect to R, a and ξ_i.

$$\max \sum_{i=1}^{n} \alpha_i K(x_i, x_i) - \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j K(x_i, x_j)$$
$$s.t. \ \sum_{i=1}^{n} \alpha_i = 1 \qquad (8)$$
$$0 \leq \alpha_i \leq C$$

Denote its optimal solution as $(\alpha_1^*, \alpha_2^*, ..., \alpha_n^*)$; the sphere center $a = \sum_{i=1}^{n} \alpha_i x_i$ is a linear combination of input training samples, and the radius is computed by

$$R = \frac{1}{n} \sum_{i=1}^{n} \| x_i - a \| \qquad (9)$$

For any input education data x, if the distance to the sphere center is less than R, we accept the learners as the target class, and the learners' status is normal; if the distance to the sphere center is bigger than R, we detect the learners as the outlier class, and the learners' status is abnormal. Figure 3 illustrates the classification principle of SVDD.

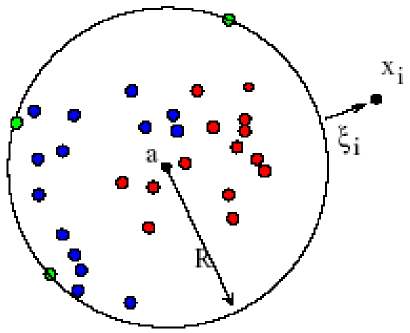

Fig. 3. Classification principle of SVDD

SVDD uses (10) to judge the status of a new learner.

$$f(x) = \mathrm{sgn}\{R^2 - (\phi(x) - a)^T (\phi(x) - a)\} \quad (10)$$

V. CONCLUSIONS

Denote the Learners' behaviour data in the vector form, this paper designs a frame for applying the widely used machine learning algorithms for analysis and classification. Using the obtained scientific evidence, the educators can teach more effectively and provide the learners needed education, which will increase the speed and quality of education informatization. Future research will include applying these algorithms into real education data and mining the hidden information, which will be beneficial to designing the teaching strategy and carrying out intervention for the teachers.

ACKNOWLEDGMENT

Thanks for all those who helped me during the writing and reviewing process of the paper. Deepest thanks goes first and foremost to Professor Yong Nie, who gives me consistent and illuminating instruction. Sincere thanks goes to the editors for their industrious hard work. Special thanks goes the anonymous reviewers for their valuable suggestions and comments, which make the paper more complete.

REFERENCES

[1] Mittag, H. J. . Blended Learning in Practice: An overview on Recent Developments [J]. Journal of Lifelong Learning Society, 2016, 1(5): 171-186.

[2] Chen, H. J., Dai, Y. H., & Feng, Y. J. et al. Construction of Affective Education in Mobile Learning: The study Based on Learner's Interest and Emotion Recognition [J]. Computer Science & Information Systems, 2017, 14(3): 685-702.

[3] Shu-Fen Tseng，Yen-Wei Tsao, Liang-ChihYu， Chien-lung Chan, K. & Robert Lai. Who will pass? Analyzing learners behaviors in MOOCs. [J]，Research and Practice in Technology Enhanced Learning, 2016, 11 (1)：1-11.

[4] Armando Fox. From MOOCs to SPOCs[J]. Communications of the ACM, 2013, 51(12): 38-40.

[5] Tim Goral. Make way for SPOCS：Small Private Online Courses May provide What MOOCs can't [J]. University Business, 2013, 16(17): 45-46.

[6] Sun Yujie. Research of design and practice based on SPOC College Basic Computer Courses in the post MOOC [D], Hebei Normal University, 2016.

[7] Unal Cakiroglu. Analyzing the effect of learning styles and study habits of distance learners on learning performances：A case of an Introductory Programming Course [J]，International Review of Research in Open and Distance Learning, 2014, 15(4).

[8] Vasesen B E, Prins F J, Jeuring J. University students's achievements goal and help seeing strategies in an intelligent tutoring system [J]. Computers & education, 2014 (72): 196-208.

[9] Peng Shaodong. A mining method model and its application of online learning behavior research in the era of big data [J]. E-education Research, 2017, 38(1): 72-81.

[10] ZHAO Huiqiong and so on. Empirical Research of Predictive Factors and Intervention Countermeasures of Online Learning Performance on Big Data-based Learning Analytics [J]. E-education Research, 2017, 38(1): 64-71.

[11] Li Kai, Gao Yan, Cao Zhe. Fuzzy clustering algorithm based on the automatic variable weights of samples and features[J]. Journal of Harbin Engineering University, 2018,39(09):1554-1560.

[12] Ye Qi-xiang, Han Zhen-jun, Jiao Jian-bin. Human detection in images via piecewise linear support vector machines [J]. IEEE Transactions on Image Processing, 2013, 22(2): 778-789.

[13] Zhou Su, Hu Zhe, Wen Zejun. A K Means/ Support Vector Machine based self-adaptive online fault diagnosis method for fuel cell system [J]. Journal of Tongji University,2019, 47(2): 255-260.

[14] Zhao Yang, Wang Sheng-wei, Xiao Fu. Pattern recognition-based chillers fault detection method using support vector data description (SVDD)[J]. Applied Energy, 2013, 112(1): 1041- 1048.

[15] Liang Jinjin, Wu De. Clustering piecewise double support vector domain classifier [J]. Control and Decision, 2015, 30(7): 1298-1302.

[16] Zou Bin， Peng Zhiming， Xu Zongben. The learning performance of support vector machine classification based on Markov sampling[J]. Science china information sciences, 2013， 56(3): 1-16.