

PAPER • OPEN ACCESS

Consumption behavior evaluation of college students based on PSO-DTSVM

To cite this article: Hui Chen *et al* 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* **569** 052104

View the [article online](#) for updates and enhancements.

Consumption behavior evaluation of college students based on PSO-DTSVM

Hui Chen¹, Xinze Wang¹, Yan Zhou¹, Shasha Zhao^{1,2} and Dengyin Zhang^{1,2*}

¹ School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu, 210003, China

² Jiangsu Key Laboratory of Broadband Wireless Communication and Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu, 210003, China

*Corresponding author's e-mail: zhangdy@njupt.edu.cn

Abstract. Blind and reckless consumption of the college students has become a common phenomenon. To give a correctly guiding, a decision tree support vector machine (DTSVM) based on particle swarm optimization (PSO) used to evaluate college students consumption behaviour is proposed in this paper. Firstly, the classification standard of college students consumption was derived with the investigation and data analysis. Then, the consumption ability and habits for different students are divided into four categories. Lastly, the targeted consumption index is fed back for the corresponding students. Experiment results show that a classification accuracy of 99.79% can be achieved, which will provide better guidance for the college students' consumption behaviour.

1. Introduction

Recently, the only child has become the main body of college students. Their consumption ability and level have been greatly improved and changed accordingly, which has already become a very important part of the social consumption system. However, blind, group, and excess consumption as well as other adverse behavior have become common phenomena in the college student community. Therefore, it is important to know and analyze the consumption behavior of college students. In terms of dealing with these big data problems, technologies including big data analysis[1], data mining[2][3] and machine learning[4] are favored by researchers.

Due to its complexity and differences in consumption abilities, it is not scientific or rigorous to give a same feedback to all students in the process of evaluating college students consumption behavior. Moreover, there is no denying that the same thing exists in different situations. Therefore, initially classifying college students consumption behavior is a key point.

Meanwhile, the decision tree support vector machine (DTSVM), as a new technology in data mining has been widely applied in the field of classification, which demonstrates a better performance compared to traditional classification learning algorithm[5][6]. However, due to it hardly depend on the choice of parameters, and great advantage in searching for the optimal solution with Genetic algorithm (GA), a parameter selection method based on GA-SVM had been proposed[7]. As results, the calculation time can be shorten, but the reliability of initial value selection is also reduced, and the operation is complicated. While and fortunately, particle swarm optimization (PSO) [8] as a GA evolved from the predatory behavior of birds with few mathematical constraints on the shape of the



optimization function, has superior performance, easy operation and widely used in the field of parameter optimization. As a result, it can accurately search for the optimal solution of large-scale space[9]. Thus, the PSO algorithm was chose to find the best SVM trainer parameters inhere.

Considering classification of the consumption behavior is a typical multi-classification problem, a multi-class support vector machine based on the decision tree (DT-SVM) [10] which trains different classifiers at different nodes was used to classify the consumption behavior of the college students.

In this work, a PSO-based decision tree SVM classification model was proposed and used to evaluate the consumption behavior of the college students. Basic consumption data of college students are firstly collected through iot platform and survey. Then, decision tree SVM is used for multi-classification, and PSO is used to find the optimal parameters for each classification level. At last, the student own consumption and mass consumption were feedback to guide their consumption.

2. Construction of evaluation system

2.1. Data set

Since the consumption may result from different genders, grades, and happened at different time, two data sets respectively named *basic classification information* and *consumer transaction information* had been constructed.

The basic classification information is the personal information of each college student, which can be recorded

$$q_i = \{t_i, x_i, g_i\} \quad (2.1)$$

where t_i is the purchase time mark of the i th student (working day: $t_i=1$; holiday: $t_i=0$), x_i is the gender mark of the i th student (male: $x_i=1$; female: $x_i=0$) and g_i is the grade mark of the i th student (Freshman: $g_i=1$; sophomore: $g_i=2$; junior: $g_i=3$; and senior: $g_i=4$).

In each basic category, there is consumer transaction information for each corresponding student. The consumption structures and funds for each user will be recorded in a fixed period time. In consumer transaction information, a feature vector for each transaction is also established. The i th record is recorded as

$$p_i = (s_i, c_i, cl_i, fo_i, tr_i, ga_i, mp_i) \quad (2.2)$$

where s_i and c_i is the respective theoretical and actual daily cost, and cl_i , fo_i , tr_i , ga_i , and mp_i is respectively the daily clothing, food, travel, game, and cosmetic cost, which are used to calculate students' daily spending.

2.2. Classification standard

To achieve a reliable and universal test set, basic information and transaction information of 500 students had been collected. It unobvious that these differences can be ignored within the same category. So 100 female freshmen are selected as representatives, and their consumption records during the working day are recorded. Some of them are shown in the Table 2.1.

Table 2.1 Transaction records of some samples (RMB/day)

	Theoretically	Actually	Clothing	Food	Travel	Game	Cosmetic
1	34	30.07	2.76	22.87	1.04	0.61	2.80
2	33	30.60	3.11	21.79	0.72	1.25	3.72
3	33	34.00	2.08	26.44	1.45	0.32	3.70
4	32	34.67	1.72	25.79	0.60	1.02	5.53

Obviously, there is a big gap between the theoretical and the actual consumption of each student, which means some people will have a surplus of funds, and some will have a shortage. Through the comparison of these two values, the consumption of all college students in one day is firstly divided into *excess consumption* and *normal (non-excess) consumption* two categories.

Although the consumption structure may be different due to the difference in gender, age, and consumption time, it should be stable in a certain fixed category. For individuals consistent with the stable structure, the consumption is *reasonable*. If the consumption is more or less, there is a sign of abnormality, suggesting *unreasonable* consumption. In order to find reasonable and unreasonable boundaries, firstly, we average the daily consumption of 100 college students in each aspect, and the proportion of each consumption in total consumption is shown in Figure 2.1.

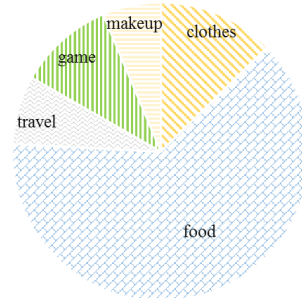


Figure 2.1 The average proportion of all kinds of consumption.

Considering the necessary of the nutritional needs for everyone, the food items will be ignored in the following analysis. Obviously, the cost of clothing and cosmetics for a girl will be much greater than that spending of games and going out. Thus, the clothing and cosmetics cost were set above the average as the unreasonable consumption. Contrary to the girls, the consumption for boys mainly focused on the cost on the games and outgoing.

Thus, on the basis of excess consumption and normal consumption, each category can be divided into two sub-categories of reasonable and unreasonable consumption. Here, all users are divided into *excess and unreasonable consumption*, *excess but reasonable expenditures*, *normal but unreasonable consumption*, and *normal and reasonable consumption* four categories.

The sample classification is shown in Figure 2.2. It can be seen that the classification trend of each type of consumption is same, indicating our classification is reasonable and effective.

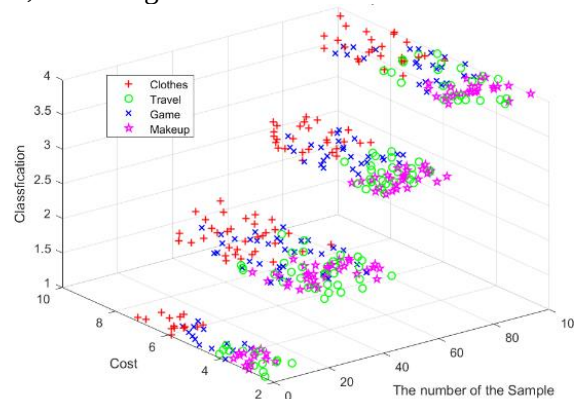


Figure 2.2 The sample classification.

2.3. Evaluation of consumption behaviour

After the $i+1$ consumption record of the student k is collected, the class branch of the classmate can be obtained with the basic classification $q_k = \{t_k, x_k, g_k\}$. Then, the daily consumption data can be formed as $p_{i+1} = (s_{i+1}, c_{i+1}, cl_{i+1}, fo_{i+1}, tr_{i+1}, ga_{i+1}, mp_{i+1})$. Using the trained classifier, when the p_{i+1} is inputted, a class of students similar to the consumption level of the classmate is obtained. Then in this category, the average theoretical consumption value is fed back as

$$\bar{s}_i = \frac{\sum_{i=1}^n s_i}{n} \quad (2.3)$$

The average actual consumption value can be deduced as

$$\bar{c}_i = \frac{\sum_{i=1}^n c_i}{n} \quad (2.4)$$

The average of the five consumer categories of clothing, food, travel, games and cosmetics will be

$$\mu_j (j = 1, 2, 3, 4, 5) = \frac{\sum_{i=1}^n \mu_{ji}}{n} \quad (2.5)$$

In the main category, whether the student has excess consumption or not in that day will be clear. In subclassification, average consumption level of the overall college students in a similar consumption structure will be derived as guidance feed backed to the students. Finally, the objective guidance to college students will be realized, realizing the goal of consumption upgrade.

3. Establishment of classification model

3.1. SVM

For university, the training sample can be described as $(p_i, y_i), i = 1, 2, \dots, n, y \in \{+1, -1\}$, where n refers to the number of samples, p_i refers to the i th consumption record. Within a linear separability, the hyperplane for completely separating the two types of samples is characterized as

$$\omega p + b = 0 \quad (3.1)$$

where, ω is the weight vector, and b is the deviation. At this point, the classification interval will be $\frac{2}{\|\omega\|}$. So, we can conclude that the smaller the value of $\|\omega\|$ is, the larger the classification interval will reach.

To make the training sample be linearly inseparable, a non-relaxed variable $\xi_i, i = 1, 2, \dots, n$ was introduced to solve the optimal classification surface problem. The problem of finding the optimal hyperplane can be transformed into a quadratic programming problem characterized as

$$\begin{cases} \min \frac{\|\omega\|^2}{2} + C \sum_{i=1}^l \xi_i \\ s.t. y_i [(\omega \cdot p_i) + b] \geq 1 - \xi_i, \xi_i \geq 0, i = 1, 2, \dots, n \end{cases} \quad (3.2)$$

where, C is a penalty function. A larger C indicates a greater penalty for misclassification. By solving the above optimization problem with the Lagrange multiplier method, the optimal decision function can be obtained as

$$f(p_{i+1}) = \text{sgn} \left[\sum_{i=1}^n y_i a_i (p_{i+1} \cdot p_i) + b \right] \quad (3.3)$$

where, a_i is the Lagrange coefficient. The category of p_{i+1} is determined by the above formula while testing. According to the KKT complementary conditions, the solution to the above optimization problem must satisfy following condition

$$a_i (y_i (\omega \cdot p + b) - 1) = 0 \quad (3.4)$$

The a_i will be null for most samples. Consequently, the sample can be correctly classified with only a small number of support vectors.

Apparently, in the college student consumption data set, overlapping area between two samples is

large. The training samples p_i can be mapped to a high-dimensional space Hilbert, which makes it linearly separable. Nonlinear linear partitioning can be achieved with different kernel functions. According to the Mercer condition, the corresponding optimal decision function becomes

$$f(p_{i+1}) = \text{sgn} \left[\sum_{i=1}^n y_i a_i K(p_{i+1} \cdot p_i) + b \right] \quad (3.5)$$

3.2. DT-SVM

The stricter of decision tree SVM multiple classification method used is shown in Figure 3.1. The whole class is firstly divided into two sub-categories, and then each sub-class is divided into two sub-subclasses again at the next level. Such process keeps going, until the leaf nodes are generated at final, at where each decision point is classified by SVM.

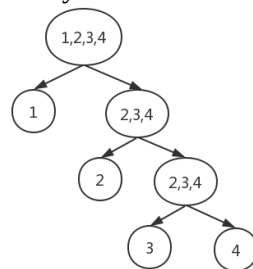


Figure 3.1 Structure of decision tree SVM.

Compared with converting single-class support vector machines into multi-class support vector machines, decision tree SVM has several benefits. It improves the classification accuracy without non-separable regions and requires fewer classifiers, which shorten the training time and reducing the number of required training samples.

While, once the classification error occurs, the error will continue along the tree structure to the subsequent nodes, leading the classification result far from the reality. Nevertheless, the PSO can optimize the input parameters of each classification. Hence, on the basis of decision tree SVM, PSO algorithm was used to further optimize the parameters.

3.3. PSO-DTSVM

(1) Basic theory

As for PSO, its basic idea is to name each solution of the optimization problem a particle and define a fitness function to measure the degree of dominance of each particle. Each particle is a point in the n -dimensional space, which is close to two points in the solution space. The first is the optimal solution get in the group, called the global optimal solution (gbest). The second is the best place ever had, called the optimal solution for the individual (pbest).

$x_i = [x_{i1}, x_{i2}, \dots, x_{in}]$ is particle i , $gbest = [g_1, g_2, \dots, g_n]$ is the global optimal solution of particle i , $pbest = [p_1, p_2, \dots, p_n]$ is the individual optimal solution of particle i , and $v_i^k = [v_{i1}^k, v_{i2}^k, \dots, v_{in}^k]$ is the flying speed of particle i . For each particle updates, their speed and position can be derived with the following formula

$$v_{id}^k = w_i v_{id}^{k-1} + c_1 r_1 (p_{id}^{k-1} - x_{id}^{k-1}) + c_2 r_2 (g_d^{k-1} - x_{id}^{k-1}) \quad (3.6)$$

$$x_{id}^k = x_{id}^{k-1} + v_{id}^k \quad (3.7)$$

where k is the k^{th} iteration, $i=1,2,\dots,m$, $d=1,2,\dots,n$; m is the number of particles, r_1 and r_2 is the random number in $0 \sim 1$, c_1 and c_2 is two commonly numbers, which is called the acceleration factor; and w_i is the inertia weight, which can change the strength of search ability.

(2) Algorithm design

SVM was used to build the model, and the parameter c and σ [11] has a large impact on the accuracy of the model. PSO is used to optimize the parameters of the classifier at each node. The corresponding optimal binary decision tree structure is displayed in Figure 3.2.

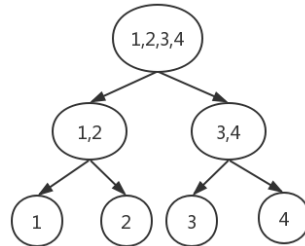


Figure 3.2 Decision tree SVM based on PSO.

The specific process of the PSO-based decision tree SVM generation algorithm is as follows

Step 1: Set the parameters of the PSO, including the number of particles m , c_1 , c_2 , ω and so on.

Step 2: Determine the search space and the approximate range of c and σ .

Step 3: Calculate the fitness value with taking the prediction accuracy of each SVM: $f(x) = accuracy_k$.

Step 4: Select the optimal fitness of the individual particles, and compare the individual fitness with its own optimal fitness. If the latter is not as good as the former, it will be replaced.

Step 5: Select the optimal fitness of the particle swarm group, and compare the individual fitness with the optimal fitness of the particle swarm group. If the latter is not as good as the former, it will be replaced.

Step 6: Determine the period. If the stop condition of the algorithm is satisfied, the cycle is stopped. On the contrary, both the speed and the position of the particle will be updated and proceed to Step 3.

Step 7: Based on the results, the PSO-DTSVM model is established.

4. Simulation and result analysis

Here, the RBF[12] is used as the kernel function of SVM, which has strong learning ability and proved to be a universal kernel function. Two main parameters c and g of RBF is adjusted. The penalty factor c is used to adjust the ratio of the confidence range between the learning machine and empirical risk. The main function of g is characterized the complexity of the distribution of sample data in the high-dimensional feature space.

In this paper, for 16 major categories of college student consumption information, 500 sets of data were captured for experiment. By the trial and error, c and g are set changing within 1~20. Select all 500 groups of data in any category as the training set and the test set. The classification accuracy affected by the effect of parameter value in cross-validation and grid search on is shown in Figure 4.1.

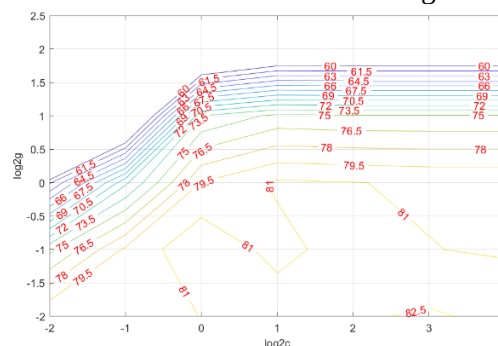


Figure 4.1 Effect of DTSVM parameter values on accuracy.

As displayed in Figure 4.1, there is a nonlinear correlation between c and g . The simulation result shows the best classification accuracy of 82.7083% under the cross-validation for $c=8$, $g=0.25$. Meanwhile, as we can see in Figure 4.2, the predicted accuracy of the trained support vector machine

reaches to 98.5417%.

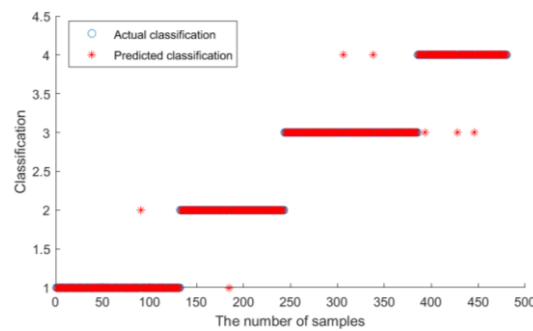


Figure 4.2 Comparison of Actual Classification and Predicted Results in DTSVM.

To further improve the prediction accuracy, we use PSO to continue optimize the two parameters c and g . The PSO local search ability and the global search ability are both set as 0.5, the maximum number of iterations is 200, the maximum number of population is 25, the inertia factor k is initially 0.6, and the range of SVM parameters c and g is $[1, 20]$. Calculating the initial fitness value of each particle, finding the optimal value of the individual and the population, and updating the speed and position of each particle until the iterations reaches to the maximum number. The detail experiment was carried out according to the algorithm in Section 3.3.

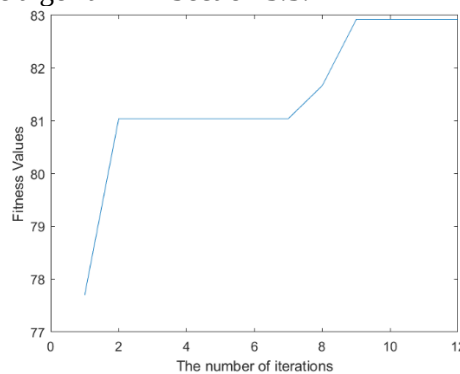


Figure 4.3 PSO fitness curve.

Iterations number dependent fitness value is shown in Figure 4.3. It can be seen that the fitness value tends to be stable after the iterations number exceeding 10 times, which suggests that the best classification accuracy under the cross-validation is 82.9167%. Compared with the previous 82.7083%, the accuracy rate was increased by 0.2%. At this time, $c=1.6903$, $g=1$, and the prediction accuracy rate is 99.7917%, which improves accuracy by 1.25% compared to the previous test. In addition, according to the prediction result shown in Figure 4.4, the PSO-based DTSVM classification successfully optimizes the penalty factor c and the kernel function coefficient g with more efficient values.

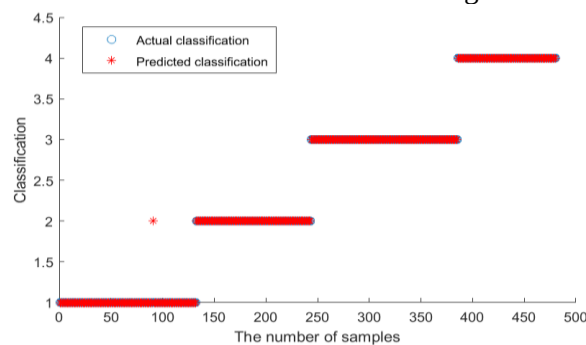


Figure 4.4 PSO-based decision tree SVM prediction results.

5. Conclusions

In conclusion, a PSO-based decision tree SVM classification method used to evaluate the consumption behaviour of the college students was proposed. The SVM multi-classifier was constructed and PSO was used to optimize the input parameters of the classifier. Experimental results imply improved classification accuracy of the consumption data with this method, which demonstrates a certain practical value. Although long the training time is needed with introducing PSO, the structural optimization and training process of decision tree is often an offline process. It is worthwhile and reasonable for better SVM classification performance in cost of the increase in offline process time.

Acknowledgements

This work comes from the research project of multi-source video compressed sensing in wireless sensor network. It was supported by Science and Technology Innovation Training Program (No. SZDG2018025), and supported by Jiangsu province key research and development - industry foresight - competition project (No. BE2017111). It also supported by the National Natural Science Foundation of China (No. 61571241).

References

- [1] K. Park, M. C. Nguyen and H. Won. (2015) Web-based collaborative big data analytics on big data as a service platform. 2015 17th International Conference on Advanced Communication Technology (ICACT), Seoul, pp. 564-567.
- [2] J. Ming, L. Zhang, J. Sun and Y. Zhang. (2018) Analysis models of technical and economic data of mining enterprises based on big data analysis. 2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), Chengdu, pp. 224-227.
- [3] S. Bhaskar, V. B. Singh and A. K. Nayak. (2014) Managing data in SVM supervised algorithm for data mining technology. Conference on IT in Business, pp. 1-4.
- [4] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal. (2016) Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, pp. 40-43.
- [5] Q. Wang, H. Chen and Y. Shen. (2008) Decision Tree Support Vector Machine based on Genetic Algorithm for fault diagnosis. 2008 IEEE International Conference on Automation and Logistics, Qingdao, pp. 2668-2672.
- [6] G. D. Liu, J. Xu. (2014) Neural network recognition algorithm of breath sounds based on SVM. Journal on Communications, pp. 218-222.
- [7] C. Sukawattanavijit, J. Chen and H. Zhang. (2017) GA-SVM Algorithm for Improving Land-Cover Classification Using SAR and Optical Remote Sensing Data. In IEEE Geoscience and Remote Sensing Letters, vol. 14, no. 3, pp. 284-288.
- [8] W. Ding and W. Fang. (2018) Target Tracking by Sequential Random Draft Particle Swarm Optimization Algorithm. In: 2018 IEEE International Smart Cities Conference (ISC2), Kansas City, MO, USA, pp. 1-7.
- [9] Y. Z. Tang and J. W. Zhou. (2015) The performance of PSO-SVM in inflation forecasting. 2015 12th International Conference on Service Systems and Service Management (ICSSSM), pp. 1-4.
- [10] X. Yu, J. Liu, Y. Zhou and W. Wan. (2010) Study of SVM decision-tree optimization algorithm based on genetic algorithm. 2010 International Conference on Audio, pp. 1079-1083.
- [11] H. Huang, Z. Wang and W. Chung. (2017) Efficient parameter selection for SVM: The case of business intelligence categorization. In: 2017 IEEE International Conference on Intelligence and Security Informatics (ISI), Beijing, pp. 158-160.
- [12] M. Duan. (2018) Short-Time Prediction of Traffic Flow Based on PSO Optimized SVM. 2018 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), pp. 41-45.