The sports performance based on variable weight and support vector machine

Wang Qingbin[1], JIA Ling[1]

1.Zhuhai College of Jilin University, Zhuhai, china 519000

Email: qingicedowa@sina.com

Key words: sports; 1000m running; prediction; screen; weight

Abstract. as sports performance is affected by many factors, it's crucial to choose effective divisors. A sports performance prediction model based on weight-SVM is proposed for factor selection and evaluation. Firstly, support vector machine is used to eliminate some invalid factors by use of cross validation MSE minimization principle; then mandatory screening is adopted to weight retention factor and quantize importance degree of retention factor; finally, SVM sports performance prediction model is constructed and applied to 1000m running performance prediction. Simulation result shows, compared with reference model, weight-SVM prediction has higher precision and it can screen and weight factor precisely and provide a new research idea for sports performance prediction.

Introduction

Prediction performance of sports performance prediction model is related to factor selection, factor importance evaluation (weight) and modeling method [1-2]. From perspective of modeling, traditional multiple linear regression and stepwise linear regression belong to linear regression model, which cannot reflect uncertain and non-linear relation between sports performance prediction and factor. What's more, the prediction result of this modeling method deviates from the actual requirement; neutral network, principal component regression and partial least squares regression are based on principle of minimum experience risk so they have over fitting phenomenon easily and have relatively poor generalization ability [3-5]. Support vector machine (SVM) is a non-liner modeling method based on structural risk minimization (SRM). It can well solve over fitting and local minimization problems of neutral network and other traditional machine learning algorithms. It has outstanding generalization ability and achieves good prediction effect [5] in many fields. From the perspective of factor selection, most methods at present select sports performance-related factors according to subjective consciousness and experience. They do not evaluate the effectiveness and weight of factors and do not treat factors equivalently. In fact, different factors have different impacts (contributions) on sports performance and they shall be weighted differently.

Weight-SVM model

SVM algorithm

Support vector machine (SVM) is a machine learning method based on statistical learning theory. It solves local extremum that cannot be avoided in the neutral network method and prevents over fitting and it has outstanding generalization ability. Estimate function of SVM is only determined by minor support vectors and computation complexity depends on the number of support vectors and is not related to space dimension of the sample so "curse of dimensionality" [9] is avoided.

Suppose there are n learning samples {xi,yi}, i=1,2,…,N，where xi is sample input, yi is expected value of model output. SVM describes estimate function in the following formula:

$$f(x) = w \cdot \varphi(x) + b \tag{1}$$

Where, w is weight vector and b is offset vector.

Optimize the target value by use of optimization function and then

$$\min J = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i^* + \xi_i) \tag{2}$$

Constraints are:

$$\begin{cases} y_i - w\cdot\varphi(x) - b \le \varepsilon + \xi_i \\ w\cdot\varphi(x) + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0, i = 1, 2, \cdots, n \end{cases} \tag{3}$$

Where, ξ_i, ξ_i^* are relaxing factors and C is penalty factor.

Introduce Lagrangian multiplier and the optimization problem above becomes a typical convex quadratic optimization,

$$\begin{aligned} L(w,b,\xi,\xi^*,\alpha,\alpha^*,\gamma,\gamma^*) &= \frac{1}{2}\|w\| + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \\ &-\sum_{i=1}^{n}\alpha_i(\xi_i + \varepsilon - y_i + f(x_i)) - \sum_{i=1}^{n}\alpha_i^*(\xi_i^* + \varepsilon - y_i + f(x_i)) \\ &-\sum_{i=1}^{n}(\xi_i\gamma_i - \xi_i^*\gamma_i^*) \end{aligned} \tag{4}$$

Where, α_i and α_i^* are Lagrangian multipliers.

To accelerate solving speed, transfer formula (3) into antithetical one and then

$$\begin{aligned} W(\alpha,\alpha^*) &= -\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)(\varphi(x_i), \\ &\varphi(x_j)) + \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)y_i - \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\varepsilon \end{aligned} \tag{5}$$

Constraints are:

$$\begin{cases} w = \sum_{i,j=1}^{n}(\alpha_i - \alpha_i^*)x_i \\ \sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0 \\ 0 \le \alpha_i, \alpha_i^* \le C \end{cases} \tag{6}$$

For linear regression, SVM regression function is:

$$f(x) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)(\varphi(x_i), \varphi(x)) + b \tag{7}$$

For non-linear prediction, kernel function $k(x_i, x)$ operates instead of $(\varphi(x_i), \varphi(x))$ to prevent "curse of dimensionality" [10]. Finally, SVM regression function is:

$$f(x) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)k(x_i, x) + b \tag{8}$$

Initial selection of factors

The object of the research is 1000m running performance. Before, 1000m running performance prediction model is constructed, in consideration of feasibility in operation, firstly select multiple factors preliminarily: height (m, x1), weight (kg, x2), chest circumference (m, x3), vital capacity (ten thousand, x4), heat rate (hundred times, x5), leg length (m, x6), 50m performance (sec, x7), long jump (m, x8), average monthly exercise time (h, x9) and 1000m running performance (y) is output.

Factor screening

During the construction of performance prediction model, the 9 factors above are selected on the basis of subjective consciousness and experience so they may be redundant or useless. Therefore, before modeling, eliminate negative impact of useless and redundant factors on prediction result. 10-fold cross validation MSE minimization principle is applied in the research to select effect factors and eliminate factors with negative impact on prediction result. MSE is defined as follows:

$$\text{MSE}=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{9}$$

Where, yi is real performance value, $\hat{y}_i$ is predicted value and n is the number of predicted samples.

Firstly, normalize factors in data set ytrain into range [-1,1] and then it serves as input variable of SVM:

$$\text{x}'=\frac{\left(\text{x}-\left(\text{A}+\text{B}\right)/2\right)}{\left(\text{A}-\text{B}\right)/2} \tag{10}$$

Where, x is initial variable value, x' is the normalized value and A and B are respectively the maximum and minimum in each line.

The factor screening process is as follows:

Step 1: firstly, normalize n factors as SVM input, select best kernel function parameters by use of 10-fold cross validation and then the training precision of all the factors is obtained and is presented as MSE0.

Step 2: remove the ith (i=1,…, n) factor and the residual factors are SVM input. Training precision MSE-i is obtained in the same way. Repeat n factors and a removed MSE set {MSE-1,MSE-2,…,MSE-n} is obtained.

Step 3: MSEmin is the minimum MSE value in {MSE-1,MSE-2,…,MSE-n}. If MSEmin<MSE0, prediction precision of model can be improved after factor corresponding to MSEmin is removed. Then the factor must be removed and change MSE0 into MSEmin.

Step 4: repeat step 1 to step 3 (the number of factors would be reduced by 1 after every circulation) until MSEmin>MSE0. At this moment, the residual factors are retention factors.

Weighting of retention factor

After factors are screened, retention factors are favorable to improvement of model prediction precision. However, contribution of different factors varies to some extent so linear regression can be used to give a dominant expression for each retention factor. As linear method cannot precisely reflect complex non-linear factor between sports performance and factor and the obtained weight value is not reliable, SVM non-linear mandatory screening adopted in the research can weight every retention factors and the process is as follows:

Step 1: m retention factors are SVM input, select the best kernel function parameter by use of 10-fold cross validation and then training precision MSEm as model background precision is obtained.

Step 2: remove the ith retention factor by force and then compute 10-fold cross validation training precise MSE-I of residual retention factors after that one is removed. Bigger MSE-i means the model prediction precision becomes worse after the retention factor is removed, which means the factor is more important to the prediction result.

Step 3: normalize each retention factor after background is deducted to obtain the weight value of factors:

$$w_j=\frac{\left(\text{MSE}_{-j}\text{-MSE}_m\right)}{\sum_{j=1}^{m}\left(\text{MSE}_{-j}\text{-MSE}_m\right)} \tag{11}$$

Sample set construction

After mom-linear factor screening and weighting above are completed, the most effective factors

of prediction and weight of factors are obtained. Then compute the sample subject to retention factors to obtain sports performance sample set based on SVM modeling.

Simulation experiment

Simulation environment

Conduct simulation experiment in the environment of software operation system: Windows 7, Pentium (R)Dual-Core E6300@ 2.80GHz, double-kernel CPU, 4GB storage to validate performance of weight-SVM.

Simulation data are 1000m test results of 235 students majoring in sports and science and engineering. See Table 1 for partial data. SVM algorithm is subject to libsvm tool (http://www.csie.ntu.edu.tw/~cjlin/libsvm/) constructed by Lin Zhiren, a Taiwanese scholar. This tool includes 4 major programs: svm-scale is used to normalize primary data; gridregression is sued to search the best kernel function parameter; svm-train is used to construct prediction model; svn-predict is used to predict test sample. Non-linear factor screening and weighting algorithms are both realized subject to self-compiled program of matlab R2012a tool kit.

Table 1 Sports Performance Data

No	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	y
1	1.80	65	0.82	0.51	0.60	0.90	6.7	2.70	75	3.18
2	1.80	80	1.00	0.55	0.65	0.88	6.9	2.55	81	3.32
3	1.80	74	0.91	0.50	0.60	0.95	6.9	2.58	70	3.10
4	1.77	72	0.88	0.48	0.62	0.86	6.7	2.70	82	3.38
5	1.70	52	0.80	0.40	0.70	0.78	6.9	2.67	81	3.24
6	1.77	71	0.82	0.52	0.64	0.87	6.9	2.80	83	3.32
7	1.85	73	0.86	0.52	0.64	0.97	6.7	2.60	80	3.24
8	1.85	75	0.85	0.60	0.65	0.95	6.8	2.50	82	3.20
9	1.72	72	0.84	0.53	0.62	0.79	6.9	2.61	46	4.13
10	1.71	70	0.79	0.49	0.68	0.80	6.8	2.56	30	4.25
11	1.68	64	0.75	0.45	0.75	0.70	7.1	2.33	25	4.56
12	1.65	65	0.85	0.52	0.64	0.69	6.7	2.42	71	3.62
13	1.62	63	0.81	0.56	0.63	0.68	6.9	2.57	35	3.92
14	1.71	71	0.79	0.49	0.67	0.75	6.8	2.48	15	4.81
…	…	…	…	…	…	…	…	…	…	…
235	1.72	68	0.80	0.59	0.71	0.80	7.0	2.63	32	3.90

Reference model and prediction evaluation index

Seven reference models are selected to test the prediction performance of weight-SVM: multiple linear regression (MLR), which has no variable screening and weighting; artificial neutral network ANN), which has no variable screening and weighting as well; stepwise linear regression (SLR), which is backward and variable α =0.2 is removed; SVM model (construct model on the basis of initial descriptor and variables are not screened and weighted); screen-SVM has the same process as weight-SVM process except that the variables are not weighted; SLR-SVM model constructs prediction model with variables screened by SLR but variables are not weighted; SLR-ANN model is the same as SLR-SVM except that model is constructed on the basis of artificial neutral network. The best kernel function parameters of 8 models are selected by 10-fold cross validation on the basis of training sample. MSE prediction performance of independent test sample is model evaluation standard. Smaller MSE means the higher model prediction precision.

Non-linear factor screening and weighting

Select 170 samples from all the 235 samples at random and include them into a training set anf the residual 65 samples are included into independent test set. Screen 9 descriptors by SVM non-linear screening on the basis of training samples and 6 retention factors are obtained, which are: long jump, weight, vital capacity, heart rate and montly exercise time. See Table 2 for retention factors and their weights.

Table 2 6 Retention Factors and Their Weights

Retention factors	Long jump	Weight	Vital capacity	Heart rate	Monthly exercise time
Weight value	0.27	0.21	0.19	0.18	0.15

It can be seen in Table 2 that 1000m running performance is closely related to 6 indexes of long

jump, weight, vital capacity, heart rate, monthly exercise time and weight of long jump is the maximum because it is closely related to strength of thigh. Therefore, it can well describe 1000m running performance; what's more, another 4 indexes have relatively the same weight. Height, chest circumference, leg length and 50m sprint performance are removed, which means 1000m running performance of student is not directly related to height and leg length. In addition, chest circumference is included into vital capacity and 500m sprint performance and long jump may be mutually redundant.

Result and analysis

See Table 3 for prediction precision of weight-SVM and independent test samples of other 7 reference models.

Table 3 Independent Test Precision of Models

Model	MSE
MLR	0.3521
SLR	0.2342
SVM	0.1791
KNN	0.2563
SLR-ANN	0.1528
SLR-SVM	0.1214
screen-SVM	0.0812
weight-SVM	0.0032

Conclusion

A sports performance prediction mode based on weight-SVM is established for complex non-linear sports performance. Factors closely related to prediction result are selected via SVM non-linear factor screening and non-linear weighting and every factor is weighted in a non-linear way. Simulation experiment result shows weight-SVM improves sports performance prediction precision and provides a new research idea for non-linear sports performance prediction.

Reference

[1] Jinyu Hu, Zhiwei Gao and Weisen Pan. Multiangle Social Network Recommendation Algorithms and Similarity Network Evaluation[J]. Journal of Applied Mathematics, 2013 (2013).

[2] Yishuang Geng, Jin Chen, Ruijun Fu, Guanqun Bao, Kaveh Pahlavan, Enlighten wearable physiological monitoring systems: On-body rf characteristics based human motion classification using a support vector machine, IEEE transactions on mobile computing, 1(1), 1-15, Apr. 2015

[3] Jinyu Hu and Zhiwei Gao. Modules identification in gene positive networks of hepatocellular carcinoma using Pearson agglomerative method and Pearson cohesion coupling modularity[J]. Journal of Applied Mathematics, 2012 (2012).

[4] Yishuang Geng, Jin Chen, Ruijun Fu, Guanqun Bao, Kaveh Pahlavan, Enlighten wearable physiological monitoring systems: On-body rf characteristics based human motion classification using a support vector machine, IEEE transactions on mobile computing, 1(1), 1-15, Apr. 2015

[5] Lv Z, Tek A, Da Silva F, et al. Game on, science-how video game technology may help biologists tackle visualization challenges[J]. PloS one, 2013, 8(3): 57990.