Frontiers in Business, Economics and Management
ISSN: 2766-824X | Vol. 4, No. 3, 2022

The Implementation of Support Vector Machine into Pairs Trading Strategy

Zihao Yu

The HongKong Polytechnic University, Guangdong 528000, China

Abstract: Academia has considerable delves into the investment strategy of the stock market. Typically, pairs trading is one of the familiar strategies. To optimize the performance of pairs trading strategy, accurate methods for price prediction should be employed and thesupport vector machine (SVM) is a typical one. This passage focuses on China stock market, demonstrating how the SVMclassifiers assist pairs trading strategy. The time period of data is from 2020-01-01 to 2022-07-01. All of the quantitative tasks are completed in Python language in Jupyter Notebook.

Keywords: SVM, Pairs Trading Strategy.

1. Introduction

1.1. Pairs Trading Strategy (PT)

The price spread of two stocks with similar trends is a must for PT. The price spread is computed as below and the general principles of the PT are established in figure 1 and figure 2.

$$spread = r_1 - r_2[1]$$

r_1 is the accumulative return rate of stock1.

r_2 is the accumulative return rate of stock2.

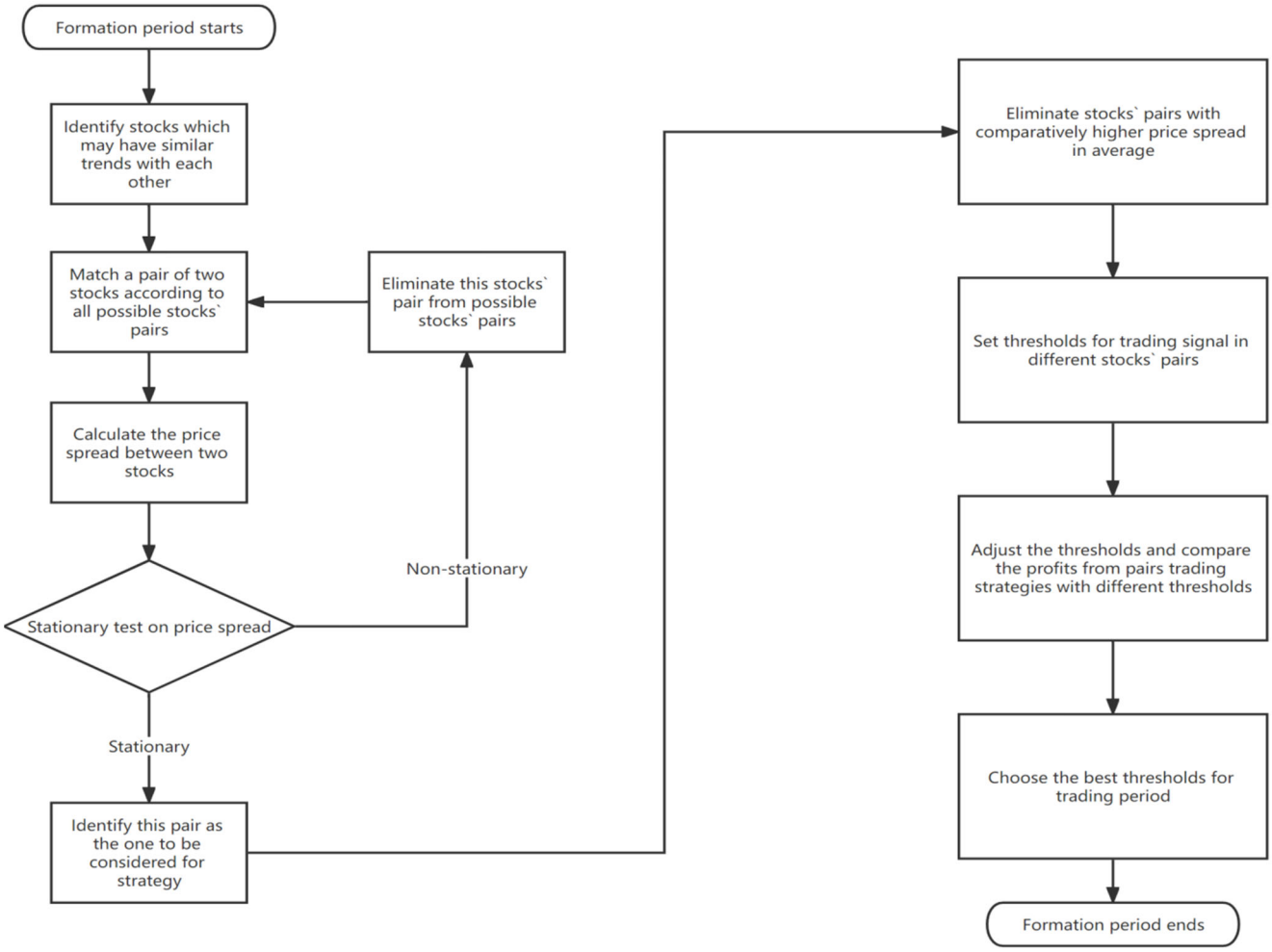

Figure 1. The flow chart of the PT in the formation period

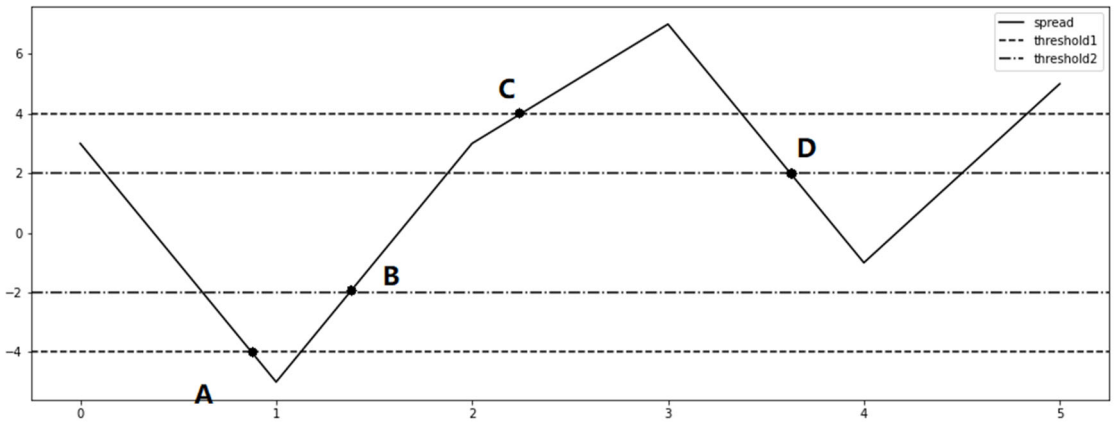

Figure 2. Sample price spread with thresholds and trading signals

Figure1 focuses on the formation period, showing the basic process of pairs` selection, spreads` selection, and thresholds` setup. According to the outcome from the formation period, some trades may be triggered by trading signals during the trading period. Four spots (A, B, C, and D) indicate the trading signals while the thresholds are represented by horizontal lines intercepting the spots. In the scenario between C and D, the trading strategy is as follows:

When the curve in figure 2 intersects C:

Long stock2 and short stock1.

When the curve in figure2 intersects D:

Closeout.

However, there is also an arbitrage opportunity between A and B. And the strategy is just inversed, e.g., long stock1 and short stock2 when such curve intersects A. In the following analysis, strategies similar to that between A and B are represented by S2 while S1 indicates another kind of strategy such as that from C to D.

For the sake of simplicity of profit calculation, it is assumed that each stock in the pair is equally weighted.Profit (from C to D)is calculated as below:

$$profit = (p_1 - p_1') + (p_2` - p_2)$$

p_1 andp_2 are the prices of stock1 and stock2 when the curve intersectsC.

$p_1`$ and $p_2`$are the prices of stock1 and stock2 when the curve intersects D.

As spot C is above spot D, it is clear that $[(p_1 - p_2) - (p_1` - p_2`)]$ is greater than 0. According to the identical transformation, $[(p_1 - p_2) - (p_1` - p_2`)]$is equal to $[(p_1 - p_1') + (p_2` - p_2)]$. Hence, the profit of the strategy is greater than 0.Meanwhile,based on such calculations above, the profit of strategy between A and B is also proved to be greater than 0.

1.2. SVM

1.2.1. Principle

The SVM is a machine learning algorithm, classifying data by curve. As presented in figure 3, two sorts of data are divided by curves. However, the classifier1 comes first as it has the averagely longestdistancefrom bilateral values.[2]

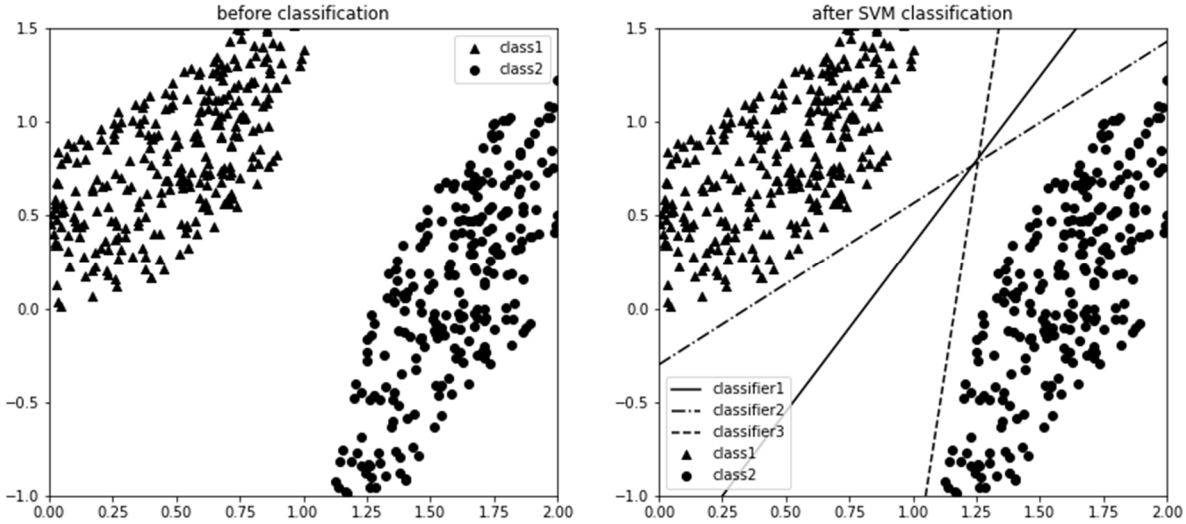

Figure 3. Classification by linear SVM

Especially, the SVM also executes non-linear classification in two ways. One is the polynomial way which classifies data by added features. As the example below in figure 4, variables with one feature can be divided by adding a feature (squared).

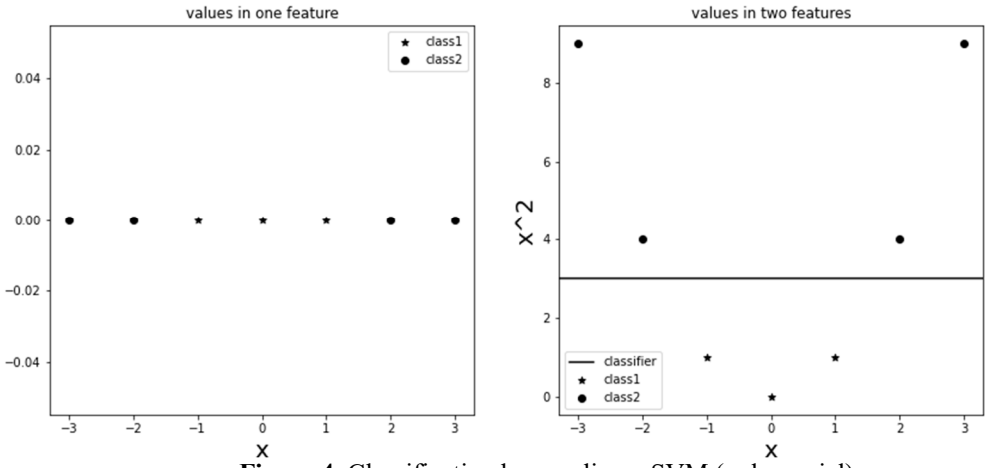

Figure 4. Classification by non-linear SVM (polynomial)

Another non-linear classifier is based on the similarity function which is usually the radial basis function (RBF, formulated as below). The $x_i - l_j$indicates the distance from the i-th value to the j-th mark while γ is the learning rate.

$$\emptyset_\gamma = \exp(-\gamma(x_i - l_j)^2)$$

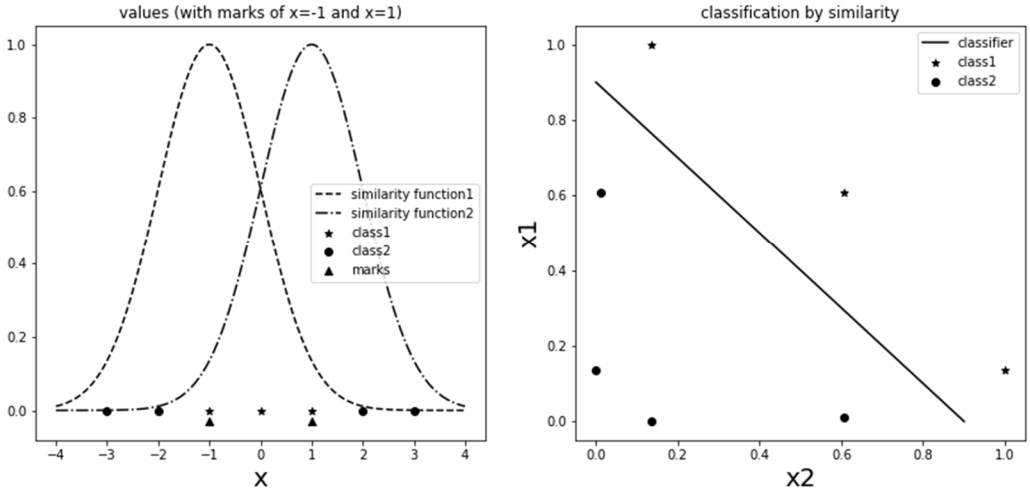

Figure 5. Classification by non-linear SVM (RBF)

Figure 5 shows that the RBF makes a new distribution of values, facilitating a linear classification.

1.2.2. Parameter

As shown in table 1, the linear, poly, and rbf represent linear classifier, non-linear classifier by polynomial features, and non-linear classifier by RBF respectively while the C is an indispensable parameter of each SVM classifier. The C reflects the extent the model tolerates misclassification in the train set and as the C decreases, more misclassifications will be accepted and the model gets more versatile. However, the model may also get over-fitted if the C is too large. Specifically, the coef0, a parameter reflecting the extent the model is affected by polynomial features in higher degrees, only works when the kernel is poly and rbf while the preset of degree is also a must for poly kernel. Lastly, gamma, the γin RBF, may make the similarity curve gets more narrowed and accordingly cause over-fitting if it is overestimated.

Table 1. Parameters of SVM

Parameters	Feasible settings
kernel	linear poly rbf
C	R^+
gamma	R^+
degree	Z^+
coef0	R

1.3. How SVM assists PT

As demonstrated in figure 6, trading signals can be put off for a day(s) if it is predicted that the price spread will maintain its trend after the original trading signals. For instance, Spot C` and spot D` represent new trading signals and there is an enlarged vertical distance between C` and D`. Hence, the profit can be improved[3].

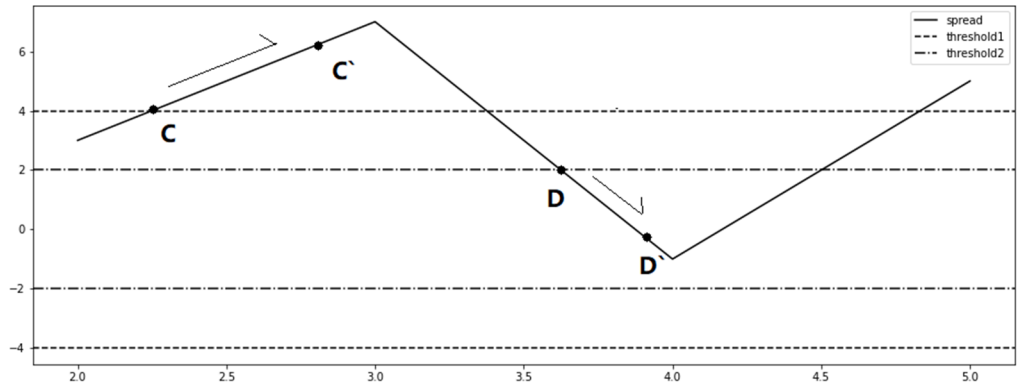

Figure 6. Modification of trading signals

2. Methodology

2.1. SVM

2.1.1. Data Pre-processing

In this passage, data is from Tushare (a free Python financial data interface package), containing basic information on stocks in China stock market. Tushare can only provide stocks data with basic variables as below:

Table 2. Variables

Name	Definition
date	trading date
open	opening price
close	closing price
high	the daily highest price
low	the daily lowest price
volume	trading volume per day

Table 3. Derivative variables

Formulas	Remarks
$closeopen_i = \frac{p_c - p_o}{p_o}$	p_c and p_o are daily closing price and opening price
$highlow_i = \frac{p_h - p_l}{p_l}$	p_h and p_l are daily highest price and lowest price
$ma_i = \frac{1}{n}\sum_{i-n+1}^{i} p_i$ (moving average)	$(i \gg n)$
$mom_{i,n} = p_i - p_{i-n+1}$ (momentum)	$(i \gg n)$
$ema_{i,n} = \alpha p_i + (1-\alpha) ema_{i-1,n}$ (exponential moving average)	$(i > n, ema_{n,n} = \frac{1}{n}\sum_{i=1}^{n} p_i)$
$rsi_{i,n} = \frac{avg_{rise}}{avg_{fall} + avg_{fall}}$ (relative strength index)	avg_{up} and avg_{fall} are average rising price and average falling price in a period of n days
$dif_i = emaX_i - emaY_i$ (difference)	$(i > Y > X)$
$dea_{i,n} = \alpha dif_i + (1-\alpha) dea_{i-1,n}$ (difference exponential average)	$(i > n, dea_{n,n} = \frac{1}{n}\sum_{i=1}^{n} dif_i)$
$macd_i = 2(dif_i - dea_{i,9})$ (moving average convergence divergence)	

Before data analysis, some derivative variables are also needed to be computed. The formulas of these variables are shown in table 3 and please be noted that $i$and $n$indicate the i-th trading day and the period of derivative variables respectively while the smoothness index in some variables is presented by α. (assumed as $2/(\mathrm{n}+1)$)

Lastly, the normalization of the values should be also completed in data pre-preprocessing. In this passage, the Min-Max normalization is employed and its definition is presented below ($value_{max}$ and $value_{min}$are the maximum and the minimum of each variable):

$$value` = \frac{value - value_{min}}{value_{max} - value_{min}}$$

Variables` selection

Because parts of values are directly calculated from other values on the same date, some values should be dropped for the sake of effect from multicollinearity. Meanwhile, some derivative values are also calculated by previous values and that makes some derivatives values in early date null which may affect the calculation in the upcoming analysis. Hence, if there is a null value, all values on the same date will be eliminated. The pre-processed data sample with selected values and the sets of different values` parameters are indicated below:

Table 4. Sample pre-processed data with selected variables

date	close	volume	close-open	ma5	ma10	high-low	rsi12	mom5	ema12	macd
26/2/2020	0.48	0.16	0.40	0.59	0.70	0.26	0.29	0.27	0.73	0.10
27/2/2020	0.52	0.11	0.34	0.58	0.71	0.10	0.36	0.27	0.73	0.13
28/2/2020	0.40	0.20	0.16	0.54	0.70	0.23	0.22	0.19	0.70	0.12
2/3/2020	0.48	0.13	0.42	0.53	0.69	0.15	0.34	0.29	0.70	0.14
3/3/2020	0.49	0.13	0.25	0.54	0.68	0.06	0.37	0.34	0.70	0.16

2.1.2. Parameters Adjustment and Metric

Parameters should keep being switcheduntil the model summits performance at the preset metric. The metric includes accuracy, precision, etc. However, extremely high values of such metrics may cause over-fitting so overall metrics should be employed. Thus, the AUC (areas under the curve) and ROC (receiver operating characteristic)are worth being the metric for the following analysis.[4]

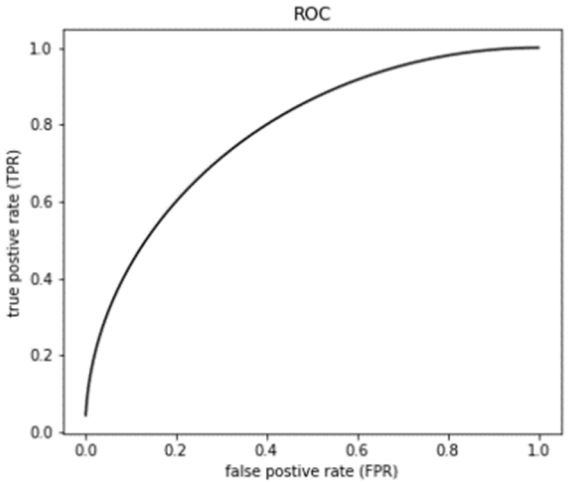

Figure 7. ROC curve

If the AUC of ROC is nearly one, the model will be in a high TPR at a lower FPR (as presented in figure 7), indicating that the model is highly accurate but not over-fitted.

2.2. Pairs Trading Strategy

2.2.1. Trade Indicated by SVM

Based on introduction 1.3, profit from strategy can be enlarged when the distance between two trading signals is enlarged. Accordingly, the trade indicated by SVM can be executed as follow:

If the curve upward crosses the trading signal:

Execute the trade if SVM indicates a downward trend

Put off the trade for one day if SVM indicates an upward trend

If the curve downward crosses the trading signal:

Execute the trade if SVM indicates an upward trend

Put off the trade for one day if SVM indicates a downward trend

2.2.2. Adjustment and Combination of Thresholds

The threshold is flexible and it should be altered in the formation period for higher profit. Typically, the threshold is formulated by the mean (μ) and standard deviation (σ) of the spread. Hence, changing the coefficient of σ (the coefficient of μ is always 1) adjusts the thresholds.

Additionally, a different set of parameters can be combined. As the thresholds of the PT may be not suitable enough to trigger considerable trades, combinations of thresholds can be employed. In other words, such a combination is to execute many PTs with different presets of thresholds for more frequent trading signals.

2.2.3. Stationary Test

As indicated in figure 1, each price spread should be stationary. Mathematically the Augmented-Dickey-Fuller (ADF) Test is a typical method for the stationary test. The null hypothesis of the ADF test is that the series has a unit root which means that the series is not stationary. The ADF test may release different outcomes at different significance levels, so in the upcoming ADF test, the preset significance level is 5%.[4]

3. Match Stocks` Pairs

Twenty stocks are chosen respectively from five industries, namely IT, finance, agriculture, retail,and transportation, for pairs`match. And the pairs (as table 5 shows) with the lowest price spread will be chosen for strategy.[5]

Table 5. Selected stocks pairs

Industry	Pair
IT	600797.SH~603421.SH
Finance	002807.SZ~601288.SH
Retail	000025.SZ~000715.SZ
Transportation	600029.SH~600115.SH
Agriculture	002696.SZ~600975.SH

4. Thresholds` Setup

The thresholds in the trading period are set up in the formation period. According to 2.2.2, some sets of thresholds can be employedsimultaneously to ensure the considerable trade volume generated by the PT in the trading period. The thresholds for each pair and the times the spread crosses thresholds in the formation period are shown below. The threshold is formulated by the mean (μ) and standard deviation (σ) of the spread in the formation period.

Table 6. Threshold`s setup

Pairs	Threshold set 1	Threshold set 2	Threshold set 3
600797.SH~ 603421.SH	$threshold1 = \mu \pm 0.3\sigma$ $threshold2 = \mu \pm 1.5\sigma$ (47 times)	$threshold1 = \mu \pm 0.5\sigma$ $threshold2 = \mu \pm 1.2\sigma$ (56 times)	$threshold1 = \mu \pm 0.8\sigma$ $threshold2 = \mu \pm 1.8\sigma$ (49 times)
002807.SZ~ 601288.SH	$threshold1 = \mu \pm 0.2\sigma$ $threshold2 = \mu \pm 1.5\sigma$ (52 times)	$threshold1 = \mu \pm 0.3\sigma$ $threshold2 = \mu \pm 1.2\sigma$ (46 times)	$threshold1 = \mu \pm 0.5\sigma$ $threshold2 = \mu \pm 1.2\sigma$ (48 times)
000025.SZ~ 000715.SZ	$threshold1 = \mu \pm 0.3\sigma$ $threshold2 = \mu \pm 1.0\sigma$ (65 times)	$threshold1 = \mu \pm 0.4\sigma$ $threshold2 = \mu \pm 1.2\sigma$ (49 times)	$threshold1 = \mu \pm 0.2\sigma$ $threshold2 = \mu \pm 1.1\sigma$ (56 times)
600029.SH~ 600115.SH	$threshold1 = \mu \pm 0.4\sigma$ $threshold2 = \mu \pm 1.1\sigma$ (49 times)	$threshold1 = \mu \pm 0.2\sigma$ $threshold2 = \mu \pm 1.0\sigma$ (47 times)	$threshold1 = \mu \pm 0.5\sigma$ $threshold2 = \mu \pm 1.2\sigma$ (44 times)
002696.SZ~ 600975.SH	$threshold1 = \mu \pm 0.3\sigma$ $threshold2 = \mu \pm 1.3\sigma$ (49 times)	$threshold1 = \mu \pm 0.5\sigma$ $threshold2 = \mu \pm 1.2\sigma$ (50 times)	$threshold1 = \mu \pm 0.4\sigma$ $threshold2 = \mu \pm 1.2\sigma$ (50 times)

5. Fit SVM Classifier

In the following analysis, the stocks` data from 2020-01-01 to 2022-01-01 is used forfitting and the model`s goal is to predict the trend of spread on the next day. The time frame of the train set is between 2020-01-01 to 2021-09-01 while the rest of the data is defined as the test set. Based on the process above, the results of the models` fitting are demonstrated in table 6.

Table 7. Model performance

Pair	Accuracy	AUC of ROC	Parameters
600797.SH~603421.SH	64%	0.63	kernel=`poly`, C=0.1, degree=3, coef0=2
002807.SZ~601288.SH	57%	0.57	kernel=`poly`, C=1, degree=3, coef0=0.5
000025.SZ~000715.SZ	56%	0.57	kernel=`rbf`, C=0.1, gamma=0.01, coef0=0
600029.SH~600115.SH	58%	0.59	kernel=`poly`, C=4, degree=3, coef0=5
002696.SZ~600975.SH	66%	0.67	Kernel=`poly`, C=2, degree=2, coef0=2

6. Evaluation of the Implementation of SVM into PT

6.1. How to Evaluate

The evaluation is based on the trading period from 2022-01-01 to 2022-07-01. As pointed out in 2.2.1, the model will indicate if it is necessary to put off the trade once the spread touched trading signals. Hence, the evaluation focuses on the delayed trade, calculating cashflow`s differencebetween the original trade (PT) and that put off (SVM-PT).To identify such differences, the trading signals graph can be utilized. For instance, the figure below illustrates a part of the original trading signals (original) and the delayed (delayed) of the PT between 600797.SH and 603421.SH.

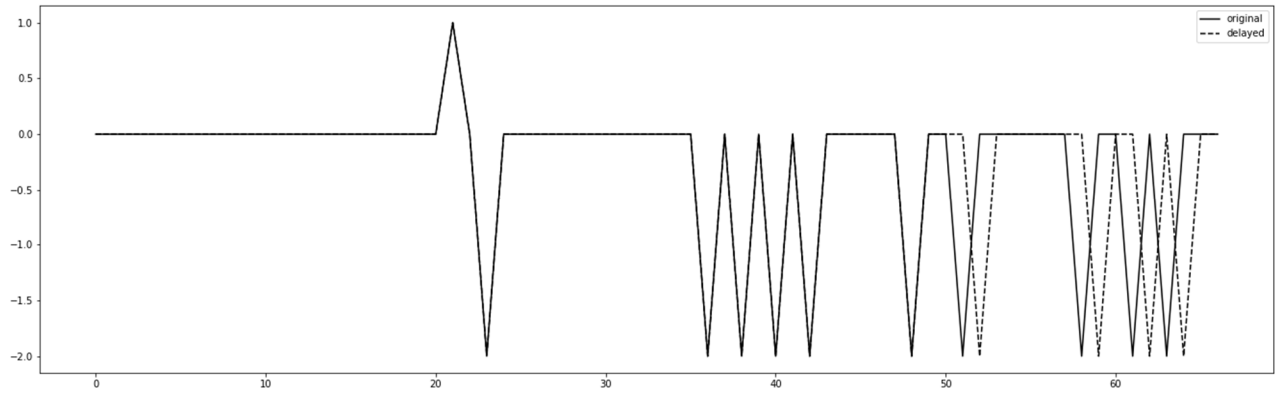

Figure 8. Trading signals

Respectively, there 5 kinds of values in both signals, namely 0, 1, 2, -1, and -2, and the indications from such values are as follow:

Table 8. Indications from signals

Signal	Indication
2	Start the S1
1	End the S1 (by closing out)
0	No indication (no trade)
-1	End the S2 (by closing out)
-2	Start the S2

Lastly, based on the trading signals and indications from them, the cashflow`s difference can be computed. As the cashflow is unchanged if the signal is not delayed, we just focus on the delayed signal and the origin of such signals. The signal generation and spread calculation are based on the close price of both stocks, so once the signal triggers a trade, the cashflow should be counted by the stock price one day after the date when signal is generated. Such prices are various and they are assumed to be represented by open prices.

6.2. Evaluation Outcome

The outcome can be seen below. Besides, another metric called accuracy is added, presenting how accurate when the model is transferred from the formation period to the trading period.

Table 9. Evaluation outcome

Pair	Threshold settings		Cashflow difference (SVM-PT - PT)	Accuracy
600797.SH~603421.SH	$\mu \pm 0.3\sigma$,	$\mu \pm 1.5\sigma$	-0.31	53%
	$\mu \pm 0.5\sigma$,	$\mu \pm 1.2\sigma$	-0.06	
	$\mu \pm 0.8\sigma$,	$\mu \pm 1.8\sigma$	0.049	
002807.SZ~601288.SH	$\mu \pm 0.2\sigma$,	$\mu \pm 1.5\sigma$	0.83	53%
	$\mu \pm 0.3\sigma$,	$\mu \pm 1.2\sigma$	1.08	
	$\mu \pm 0.5\sigma$,	$\mu \pm 1.0\sigma$	0.91	
000025.SZ~000715.SZ	$\mu \pm 0.3\sigma$,	$\mu \pm 1.0\sigma$	0.03	56%
	$\mu \pm 0.4\sigma$,	$\mu \pm 1.2\sigma$	0.33	
	$\mu \pm 0.2\sigma$,	$\mu \pm 1.1\sigma$	0.32	
600029.SH~600115.SH	$\mu \pm 0.4\sigma$,	$\mu \pm 1.1\sigma$	-0.39	58%
	$\mu \pm 0.2\sigma$,	$\mu \pm 1.0\sigma$	0.06	
	$\mu \pm 0.5\sigma$,	$\mu \pm 1.2\sigma$	-0.16	
002696.SZ~600975.SH	$\mu \pm 0.3\sigma$,	$\mu \pm 1.3\sigma$	0.83	50%
	$\mu \pm 0.5\sigma$,	$\mu \pm 1.2\sigma$	0.62	
	$\mu \pm 0.4\sigma$,	$\mu \pm 1.2\sigma$	0.62	

7. Conclusion

Table 6 tells that in most scenarios above, the return of pairs trading strategy can be improved by SVM and SVMcan also maintain its accuracy when it is transferred from the formation period to the trading period. Please note that the index of cashflow difference only indicates if SVM promotes the PT and they cannot be compared as the stocks` prices are different. However, because it is assumed that the stocks in the PT is equally weighted, it is also worth delving into the stock weights` settings in the PT for a higher and less volatile return.

References

[1] Jiang, Y. (2022). Application and Comparison of Multiple Machine Learning Models in Finance. Scientific Programming, 1–9.

[2] Jing, M.(2017). Optimal Portfolio Research with Gaussian Kernel Support Vector Machine and Genetic Algorithm, Economic Mathematics,2017,34(1):11-17.

[3] Krauss C, Stübinger J. Non-linear dependence modelling with bivariate copulas: statistical arbitrage pairs trading on the S&P 100. Applied Economics. 2017;49(52):5352-5369. doi:10.1080/00036846.2017.1305097

[4] Chang, V. et al. (2021) 'Pairs trading on different portfolios based on machine learning', Expert Systems, 38(3), pp. 1–25. doi:10.1111/exsy.12649.

[5] Zhenyu, L. (2019). Pair trading strategy design between constituent stocks of the CSI 300 Index. Degree Dissertation.