Is it possible to predict long-term success with k-NN? Case study of four market indices (FTSE100, DAX, HANGSENG, NASDAQ)

2014 J. Phys.: Conf. Ser. 490 012082

(http://iopscience.iop.org/1742-6596/490/1/012082)

Journal of Physics: Conference Series **490** (2014) 012082 doi:10.1088/1742-6596/490/1/012082

Is it possible to predict long-term success with k-NN? Case study of four market indices (FTSE100, DAX, HANGSENG, NASDAQ)

Y Shi, A N Gorban, TY Yang

Department of Mathematics, University of Leicester, LE1 7RH, UK

ys98@leicester.ac.uk , ag153@leicester.ac.uk, yty725@sina.com

Abstract. This case study tests the possibility of prediction for 'success' (or 'winner') components of four stock & shares market indices in a time period of three years from 02-Jul-2009 to 29-Jun-2012.We compare their performance ain two time frames: initial frame three months at the beginning (02/06/2009-30/09/2009) and the final three month frame (02/04/2012-29/06/2012).To label the components, average price ratio between two time frames in descending order is computed. The average price ratio is defined as the ratio between the mean prices of the beginning and final time period. The 'winner' components are referred to the top one third of total components in the same order as average price ratio it means the mean price of final time period is relatively higher than the beginning time period. The 'loser' components are referred to the last one third of total components in the same order as they have higher mean prices of beginning time period. We analyse, is there any information about the winner-looser separation in the initial fragments of the daily closing prices log-returns time series.The Leave-One-Out Cross-Validation with k-NN algorithm is applied on the daily log-return of components using a distance and proximity in the experiment. By looking at the error analysis, it shows that for HANGSENG and DAX index, there are clear signs of possibility to evaluate the probability of long-term success. The correlation distance matrix histograms and 2-D/3-D elastic maps generated from ViDaExpert show that the 'winner' components are closer to each other and 'winner'/'loser' components are separable on elastic maps for HANGSENG and DAX index while for the negative possibility indices, there is no sign of separation.

Keywords: possibility of prediction, long-term success, Leave-One-Out Cross-Validation, k-NN, ViDaExpert and elastic maps

1. Introduction

It is important to study the predictability of the time series separately from constructing the specific predictors because in creation of each model we assume some additional hypotheses about the model structure.

Our case study is aimed to find possibility of the prediction of 'success' within three years' time interval from 02-JUL-2009 to 29-JUN-2012 for four selected stock and share market indices. We compare their performance in two time frames: initial frame three months at the beginning

(02/07/2009-30/09/2009) and the final three month frame (02/04/2012-29/06/2012).The idea of the main experiment is based on backward analysis. The backward analysis can be defined as an analysis to determine properties of the inputs of a program from properties or contexts of outputs. This case study is aimed to construct experiments on the data to test if it is possible to predict the long-term success. The possibility of long-term 'success' of the selected indices is tested from the results of the experiment in the three years' time interval. For each stock market index, the closing prices of all components are collected from Yahoo! Finance website. After data pre-processing step, the remaining components are labeled with 'winner' companies or 'loser' companies (or simply just 'winner', 'loser') by using the '1/3 average price' approach. For this approach, we compute the average price ratio which is defined as the mean price of the end period divided by the mean price of beginning period. The companies are then sorted by the descending order of this computed ratio. We label the first 33.3% of companies as 'winner'and label the last 33.3% of the companies as 'loser'. Then the log-return prices are computed on this data. The k-NN algorithm with Leave-One-Out Cross-Validation with two distance measurements is used as the indicator to test the possibility of prediction.

We use Leave-One-Out Cross-Validation for k-NN classifier [1] to test the possibility of prediction of 'success' components for each market index. The data is collected from data and cleaned. Then we did experiment of Cross-Validation for k-NN using two different forms of distance measurements. Then we analyze the total error and separate error and use two methods to visualize our result. We investigate that there is a possibility of predictions for long-term success for HANGSENG and DAX indices in the result section. We summarize our result and conclude in the last part.

2. Results and analysis

The experiment is computed using MATLAB. It begins with data pre-processing step. The companies that do not have enough amounts of closing prices are deleted from the company list. The next step is to apply the'1/3 Average Price' method for the remaining companies, labeled them with 'winner' if the corresponding average prices are the largest 1/3 of the sequence of average prices, and 'loser' if the average prices are the last 1/3 of the sorted sequence. The result data is generated by joining the 'winner' and 'loser' companies in matrix form vertically. In this matrix each column represents each date and each row represents each company. The daily log-return matrix is computed from the joint matrix. This log-return matrix is then used for Leave-One-Out Cross-Validation of 1-NN algorithm. The 1-NN (k=1 case) is admissible as the error is up bounded by twice of Bayes error rate. We use 1-NN for our experiments.

2.1. Analysis of total error and separate error

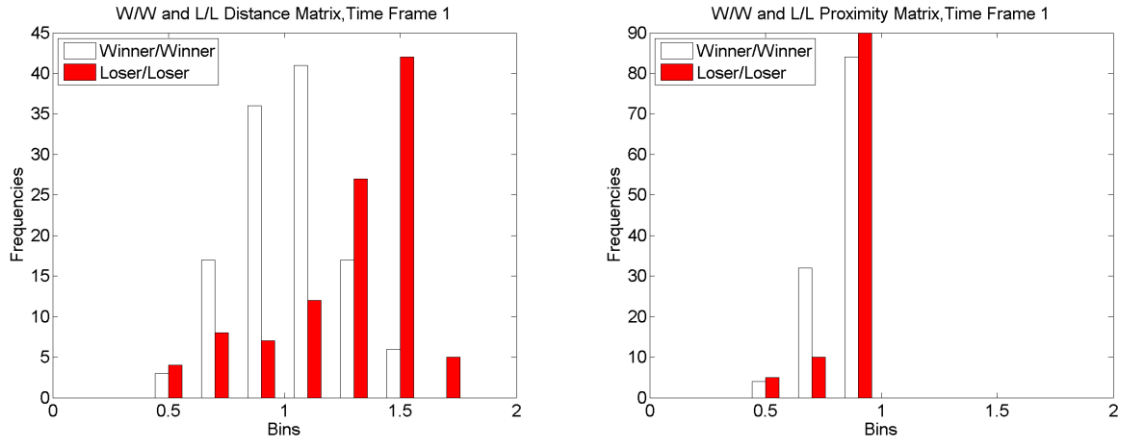
The error of Leave-One-Out Cross-Validation for 1-NN is performed for different time periods. (i.e. from 3 months to 18 months) The total error is referred to the number of misclassified points for both 'winner' and 'loser' companies. The errors generated using different functions of measurement are almost identical. The proximity can generate a bit smaller error than the distance function for several months. The total error for time period of 3 months is the minimum for three indices. For indices FTSE, DAX and NASDAQ, the errors are mostly around 50%.This means the prices are random and there is no sign of possibility of prediction. The error percentage is less than 50% for HANGSENG index. This shows this market is not completely random. To analyze the characteristics for 'winner' and 'loser' companies, the separate error analysis is applied.The sum of 'winner' error number and 'loser' error number should be the same as the total error number.

For indices FTSE and NASDAQ, the error percentages are approximately 50% for both 'winner' and 'loser' companies. Hence the separate error analysis shows that there is no sign of possibility of

prediction. For DAX index, the errors are ranged from 30% to 60%, it can be considered as a border case and the 'loser' companies in general have slightly lower errors than 'winner' companies. For HANGSENG index, both errors are below 50% and the errors of 'winner' companies are much smaller than the errors of 'loser' companies. Hence this means for HANGSENG index, there are some conclusions about predictability for the 'winner' companies.

2.2. Visualization using histograms of correlation distance matrix

From result of experiment using different time frames, the minimum error rate occurs when the time interval is within 3 months for FTSE, DAX and HANGSENG index. The sign of possibility of prediction in this time period is most dominant among all. The histograms of correlation distances can be used to study the distribution of correlation distance. For each index, a set of four histogram plots are generated. They represent the distributions of in-class (winner/winner or loser/loser) correlation distances and cross-class (winner/loser) correlation distances. The in-class histograms show that the distribution of distances between 'winner' themselves and 'loser' themselves are almost identical. The distributions of in-class distances and cross-class distances are quite similar as well. This makes 'winner' and 'loser' companies are not easily separated. Since k-NN is structure sensitive algorithm, if the points are mixed together, it generates errors. Hence if 'winner' and 'loser' companies are mixed together, this market has no sign of possibility of prediction. The histogram figures for FTSE and NASDAQ indices show no sign of possibility of prediction. For DAX index, the histogram shows some possibility of prediction but this sign is not very clear. From the in-class distributions, the 'winner' and 'loser' seems to be separated, but this separation is not clear enough since the distribution peak of winner/winner distances is slightly shifted to the left hand side of the distribution peak of loser/loser distances.



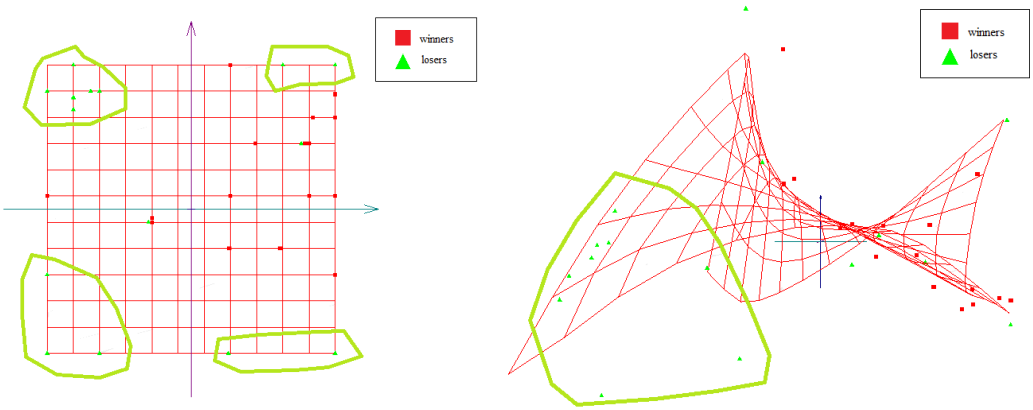
(a) (b)

Figure 2.1 Histograms of different expressions of correlation distances for HANGSENG index when using first 3 months closing prices for experiment. (a) Using Distance between winner/winner and loser/loser companies (b) Using Proximity between winner/winner and loser/loser companies

The histograms Figure 2.1 show the distributions of correlation distances for HANGSENG index. For this index, it has a clear sign of possibility of prediction. The distribution of winner/winner distances is on the left hand side of the distribution of loser/loser distances. Hence the 'winner' companies are more compactthan 'loser' companies and there is a good separation between 'winner' and 'loser' companies.

2.3. Visualization of using 2-D/3-D elastic maps

The principal graphs and manifolds can be used to visualize the companies using idea of metaphor of elastic membrane and plate to construct the principal manifold approximations of varies topologies. The software 'ViDaExpert' enables us to visualize multidimensional data with the idea of using principal object to reduce the dimension of this data [2].



(a) (b)

Figure 2.2. Visualization of components (companies) of HANGSENG index (log-returns) using elastic maps: (a) 2D - Elastic Map (b) 3D - Principal Manifold Graph The Hand-made green lines show the "clusters" of loser companies

For HANGSENG index with first 3-month time frame, the elastic maps Figure 2.2 are generated. From (a), it shows that the winners and losers are almost separated nicely. The losers generate 4 "clusters" within the internal coordinates. These "clusters" are located on each corner of the coordinate plane. From (b), it shows the winners are closer to each other than the loser companies as the red points are more compact to each other. However, the map boundary is not linear since some loser companies that are very close to the winner companies and there exists around six points that is seems to be very close to the red points. These green points can be considered as outliers. As the error of k-NN is dominated by the structure of data, this visualization supports the LOOCV error analysis.

3. Conclusions

In this case study, the 1-NN classifier combined with Leave-One-Out Cross Validation is used as an indicator to test the possibility of long-term success. The error analysis and two visualization methods show that for HANGSENG index, there is a possibility of predicting long-term success as the winner companies are closer to each other. The DAX index may be possible to predict but the sign is not clear. For FTSE and NASDAQ indices, there is no sign of possibility of prediction. Our results support some previous observations about better predictability of emerging markets [3]. For more details we refer to [4].

References

[1] Fernández-Rodríguez F, Sosvilla-Rivero S, Andrada-Félix J 1999 *International Journal of Forecasting* **15** 383
[2] Gorban A N, Zinovyev A 2010 *Int. J. Neural Syst.* **20** 219
[3] Harvey C R 1995 Review of Financial studies **8** 773
[4] Shi Y, Gorban A N and Yang T Y 2013 arXiv: 1307.8308