FULL-LENGTH REPRINT – differs from the original in layout but not in contents
Chmielewski L., Janowicz M., Kaleta J., Orłowski A.,
Pattern recognition in the Japanese candlesticks.
In A. Wiliński et al., editors, Soft Computing in Computer and Information Science vol. 342
of Advances in Intelligent Systems and Computing, pages 227-234
(Proc. Int. Conf. ACS 2014, Międzyzdroje, Poland, 22-24 Oct 2014)
The copyright owner of the paper is Springer. The final publication is available at SpringerLink
via doi:10.1007/978-3-319-15147-2_19 .

Pattern recognition in the Japanese candlesticks

Leszek Chmielewski, Maciej Janowicz, Joanna Kaleta, and Arkadiusz Orłowski

Faculty of Applied Informatics and Mathematics (WZIM),
Warsaw University of Life Sciences (SGGW),
ul. Nowoursynowska 159, 02-775 Warsaw, Poland
{leszek_chmielewski,maciej_janowicz,joanna_kaleta,arkadiusz_orlowski}@sggw.pl
http://www.wzim.sggw.pl

Abstract. Pattern recognition analysis based on k-Nearest Neighbors classifiers is applied to the representation of the stock market dynamics with the help of the Japanese candlesticks augmented by the accompanying volume of transactions. Examples from a post-emerging Warsaw stock market are given. Conditions under which the Japanese candlesticks appear to have a reasonable predictive power are provided. The dependence of the results on the number of nearest neighbors, the length of the candlestick sequence, and the forecast horizon are shown. Possible ways of the forecast improvement are discussed.

Keywords: pattern recognition, stock market forecast, Japanese candlesticks, k-Nearest Neighbors

1 Introduction

The problem of forecasting in the time series containing a stochastic component belongs to the most fascinating as well as practically important issues in the study of dynamics of natural, social and economical systems. Among the various methods of analysis of stochastic time series one should mention very traditional ones, based on the broadly understood concept of regression, as in e.g. [1], those based on linear filters, see e.g. [2], as well as the methods associated with the powerful concept of state space [3].

Methods and techniques of pattern recognition and data mining are relatively new in the field, but last decade of the previous century brought a real outburst of research in the area. Let us only mention a few papers associated with indexing [4,5,6], clustering [7,8], classification [9,10], and anomaly detection [11,12], see [13] for a more detailed bibliography. There exist three more or less standard categories into which the methods of pattern recognition and data mining usually fall: supervised learning, which includes classification, unsupervised learning (pattern detection, clustering, class discovery, characterization, change detection, and Fourier, wavelet, and principal component decomposition

© Springer International Publishing Switzerland 2015
A. Wiliński et al. (eds.): Soft Computing in Computer and Information Science,
Advances in Intelligent Systems and Computing 342.
DOI: 10.1007/978-3-319-15147-2_19

among other things) and semi-supervised learning. The survey paper [13] contains a useful exposition of the above components.

In this contribution we study the classification of patterns in the Japanese candlesticks representation of the time series which appear in the stock markets. Thus, our study is devoted to the *supervised learning* in the broad sense. More precisely, we perform the candlestick patterns classification using the k-Nearest Neighbor classifier. The emerging patterns are then analyzed from the point of view of their predictive power (or lack of it). From this point of view, our work may be viewed as a contribution which accompanies that of [14], in which the machine learning approach has been applied to develop an efficient investment strategy.

The main body of this work is organized as follows. In Section 2 we recall the definition of the Japanese candlesticks as well as what we call *augmented candlesticks* which include information about the volume of transactions. Distances between the sequences of candlesticks are also defined. In Section 3 we present our approach to the pattern recognition problem in the time series associated with stock markets. Section 4 contains the discussion of the predictive power of the sticks. Finally, Section 5 comprises some concluding remarks.

2 Japanese candlesticks as a representation of value of assets in stock market

The Japanese candlestick is a 4-tuple $(O(a,t), X(a,t), N(a,t), C(a,t))$, where O denotes the opening value of the asset a at the trading day t, X is the maximum value (*high*) reached during the trading session, N is the minimum (*low*), and C is the closing value.

In what follows below we employ five elements (O, X, N, C, V) which we call an *augmented Japanese candlestick*, where V represents the transaction volume associated with the asset and the trading day. An augmented candlestick of the asset a on the day t can be denoted as a 5-tuple

$$L(a,t) = (O(a,t), X(a,t), N(a,t), C(a,t), V(a,t)). \tag{1}$$

In the following we shall call it simply a candlestick. The time series of $n+1$ candlesticks, called otherwise a sequence, can be written down as

$$S_n(a,t) = (L(a,t), L(a,t+1), ..., L(a,t+n)). \tag{2}$$

Each sequence has its starting time t and ending time $t+n$. The sequences having the length from 1 to 5 will be investigated below.

We define the distance between two candlesticks as

$$D(a,t_1,t_2) = \sqrt{\sum_{A \in Z} \left(A(a,t_1) - A(a,t_2)\right)^2}, \tag{3}$$

where $Z = \{O, X, N, C, V\}$. In order to consider this formula meaningful, the values of the asset and the transaction volume must be comparable. To achieve

this, we normalize all time series by subtracting the closing values from the opening ones as well as from the maxima and minima, and dividing O, X, N, and C by the standard deviation of C. Similarly, the volume is also divided by its standard deviation. This way, the standard deviations of renormalized C and V are exactly 1. All time series analyzed further are normalized in the above sense.

3 Pattern recognition with the help of *k*-Nearest Neighbor classifier

The above concepts have been applied by us to the share prices of the Warsaw stock market.

A sequence as defined by Eq. (2) will be treated as a pattern. If the typical pattern recognition nomenclature is used, the feature set of this pattern is formed by $n+1$ sets, each being a 5-tuple of the elements of a candle.

For each two sequences, a distance between them can be defined in the natural (Euclidean) way:

$$D_n(a,t_1,t_2) = \sqrt{\sum_{i=0}^{n} D^2(a,t_1+i,t_2+i)}. \tag{4}$$

For each sequence $S_n(a,t)$ we can define a preceding sequence $S_n(a,t')$ where $t' < t$. The sequences preceding a given sequence can have common days with it. With the help of the distance D_n, for each sequence $S_n(a,t)$ we can identify k preceding sequences which are the nearest to it in the sense of minimizing the distance D_n. Such k preceding, nearest sequences form the set of *nearest neighbors* of S.

We have considered the set of 20 assets belonging to the so-called WIG20, the group of the largest companies of the Warsaw stock market – Polish "blue chips". For each sequence in this set we have found its k nearest neighbors among the preceding sequences for the same asset.

To facilitate the search of nearest neighbors, the elements O, X, N, C of each candlestick have been translated by the constant B chosen in such a way that

$$O(a,t') - B = O(a,t). \tag{5}$$

Thus, the compared sequences have equal first elements.

For each sequence S, for given n and k we have calculated k starting times t'_l, $l = 1,...,k$, and the corresponding distances from the considered sequence.

4 Forecasts using patterns in candlesticks: examples

Quite intuitively, the above notions and procedure can serve the purpose of an elementary forecasting. Indeed, let $S_n^{(l)}(a,t'_l)$, $l = 1,2,...,k$, be one of the k

nearest neighbors of $S_n(a,t)$ of the length $n+1$. We are interested in a prediction of the closing value $C(a,t+m)$, where $m > n$. Proceeding along a rather well-established route, we can associate with every distance $D_n^{(l)}(a,t,t'_l)$ between S and $S^{(l)}$ a weight

$$W_n^{(l)}(a,t,t'_l) = 1/D_n^{(l)}(a,t,t'_l) \tag{6}$$

and form a prediction $\bar{C}_n(a,t+m)$ as the weighted sum:

$$\bar{C}_n(a,t+m) = \sum_{l=1}^{k} W_n^{(l)}(a,t,t'_l) C(a,t'_l+n). \tag{7}$$

The prediction will be called *correct* if

$$\mathrm{sign}(\bar{C}_n(a,t+m) - C(a,t+n)) = \mathrm{sign}(C(a,t+m) - C(a,t+n)), \tag{8}$$

otherwise it will be called *incorrect*. We have also considered a modification of the above definitions, namely, to use the candles to predict the *mean* change during the following N trading sessions. More precisely, let C_N denote the following average of the closing values:

$$C_N(a,t) = \frac{1}{N} \sum_{p=0}^{N-1} C(a,t+p). \tag{9}$$

We define the prediction of the above average as $\sum_{l=1}^{k} W_n^{(l)}(a,t,t'_l) C_N(a,t'_l)$, with understanding that all t'_l are not greater than $t-N$. Again, the prediction will be correct if the sign of the predicted mean value is the same as the sign of the actual mean value.

To measure the quality of predictions given by augmented candlesticks, one can introduce the following simple function:

$$Q = \frac{P_{\mathrm{correct}} - P_{\mathrm{incorrect}}}{P_{\mathrm{correct}} + P_{\mathrm{incorrect}}}, \tag{10}$$

where P_{correct} ($P_{\mathrm{incorrect}}$) is the total number of correct (incorrect) predictions.

Tables 1-4 contain the results for two shares recorded in the Warsaw stock market: Alior Bank and KGHM Polish Copper. The results for the function Q are shown as dependent of k (the number of nearest neighbors) and N for two different lengths of the candlesticks sequences, 1 and 5 (we have also obtained results for intermediate lengths but do not present them as they are not particularly illuminating). Predictions have been made for the mean of the N values of the closing value C following the ending time of a sequence of candlesticks.

The obvious conclusion one can draw from the above tables is that the overall predictive power of the Japanese candlesticks is very well approximated by zero. In fact, of many thousands of numbers like the above we obtained for Q from our numerical procedures there have been only a few larger than 0.1. What is perhaps a little astonishing is the fact that forecasts based on a single candlestick may actually work better than those based on five sticks.

Table 1. Prediction quality function Q as dependent on the number of nearest neighbors k and the forecast depth N for the shares of Alior Bank. The sequences of candlesticks consist of one element

$N\backslash k$	1	2	3	4	5	6	7	8
1	-0.099	-0.072	-0.046	-0.059	-0.013	-0.039	-0.066	-0.052
2	0.063	0.063	0.003	0.003	-0.056	-0.036	0.003	0.029
3	0.013	0.046	0.093	0.060	0.033	0.033	0.033	0.06
4	0.023	0.036	0.130	0.117	0.043	0.083	0.050	0.030
5	0.040	0.040	0.087	0.100	0.020	0.060	0.020	0.046
6	0.070	0.070	0.097	0.104	0.037	0.030	-0.023	0.016
7	0.074	0.074	0.135	0.108	0.027	0.033	0.000	-0.020
8	0.084	0.084	0.071	0.057	-0.023	0.010	-0.010	-0.037

Table 2. Prediction quality function Q as dependent on the number of nearest neighbors k and the forecast depth N for the shares of Alior Bank. The sequences of candlesticks consist of five elements

$N\backslash k$	1	2	3	4	5	6	7	8
1	-0.067	-0.040	-0.053	0.013	-0.100	-0.026	-0.134	-0.026
2	-0.050	-0.043	0.043	0.016	-0.023	-0.010	0.003	-0.023
3	-0.094	-0.081	-0.060	-0.033	0.013	-0.013	-0.006	-0.060
4	-0.057	-0.044	0.003	0.010	-0.030	0.010	0.016	0.003
5	-0.013	-0.013	-0.020	-0.020	-0.040	0.000	-0.047	-0.006
6	-0.030	-0.030	0.030	0.030	-0.023	-0.023	0.017	0.010
7	-0.020	-0.020	0.013	0.020	-0.047	-0.027	0.020	-0.020
8	-0.044	-0.044	-0.003	0.003	-0.072	-0.051	-0.058	-0.099

Table 3. Prediction quality function Q as dependent on the number of nearest neighbors k and the forecast depth N for the shares of KGHM Polish Copper. The sequences of candlesticks consist of one element

$N\backslash k$	1	2	3	4	5	6	7	8
1	-0.101	-0.060	-0.049	-0.057	-0.059	-0.048	-0.057	-0.051
2	-0.024	-0.011	-0.015	-0.001	-0.010	-0.021	-0.020	-0.005
3	-0.003	0.002	-0.001	0.007	-0.001	-0.011	0.003	0.011
4	0.014	0.021	0.016	0.018	0.006	0.009	0.008	0.019
5	0.017	0.021	0.019	0.022	0.002	0.010	0.007	0.005
6	0.012	0.014	0.016	0.020	0.014	0.009	0.016	0.010
7	0.012	0.016	0.006	0.006	0.016	0.011	0.018	0.011
8	0.003	0.004	-0.002	0.000	0.010	0.012	0.014	0.011

Table 4. Prediction quality function Q as dependent on the number of nearest neighbors k and the forecast depth N for the shares of KGHM Polish Copper. The sequences of candlestick consist of five elements

$N\backslash k$	1	2	3	4	5	6	7	8
1	-0.107	-0.058	-0.041	-0.038	-0.039	-0.039	-0.034	-0.031
2	-0.032	-0.019	-0.004	-0.011	-0.007	-0.009	0.002	0.000
3	-0.006	0.005	0.015	0.014	0.026	0.016	0.023	0.016
4	-0.010	-0.005	0.016	0.011	0.007	-0.006	-0.001	0.006
5	0.001	0.009	0.025	0.021	0.029	0.008	0.008	0.011
6	-0.015	-0.013	0.006	-0.003	-0.011	-0.015	-0.006	-0.011
7	-0.007	-0.001	0.012	0.001	0.003	-0.001	-0.010	0.000
8	-0.004	-0.002	0.016	0.007	0.014	0.014	0.004	0.016

The situation changes somewhat, however, if we allow for more modest forecasting principle. The procedure applied above can be summarized as an attempt to predict the change of the value of the asset with respect to the close value of the last known candlestick. However, what if we want to predict the change with respect to the *first* (i.e. the earliest) candlestick in a sequence? We have observed the improvement in the forecasting quality even in the case of three sticks, but it is of course better visible in the case of five, as specified in the following tables.

Table 5. Prediction quality function Q as dependent on the number of nearest neighbors k and the forecast depth N for the shares of KGHM Polish Copper. The sequences of candlestick consist of five elements. Predictions has been calculated with respect to the first (earliest) of the candlesticks.

$N\backslash k$	1	2	3	4	5	6	7	8
1	0.171	0.194	0.185	0.197	0.184	0.190	0.208	0.218
2	0.179	0.185	0.198	0.207	0.199	0.207	0.204	0.217
3	0.161	0.162	0.180	0.190	0.191	0.194	0.192	0.201
4	0.160	0.163	0.162	0.184	0.183	0.190	0.191	0.197
5	0.143	0.143	0.153	0.167	0.161	0.167	0.184	0.179
6	0.145	0.146	0.147	0.165	0.150	0.159	0.167	0.165
7	0.151	0.151	0.136	0.160	0.151	0.159	0.153	0.162
8	0.151	0.151	0.146	0.165	0.160	0.171	0.162	0.174

The reason of the improvement of the results in Table 5 with respect to that of Tables 1-4 is that many of the sequences are parts of approximately linear trends. It is, naturally, far easier to achieve better quality of predictions if one refers to the starting or early points in a trend, instead of a point in the middle of an ascent or a descent.

5 Concluding remarks

In this work we have applied the k-Nearest Neighbors classifier to the patterns which emerge in the Japanese candlesticks being a representation of the state of share prices in the stock market. The candlesticks have been augmented to include information about the volume of transaction in the market for a given asset. The discovered patterns have been used to check the possible predictive power of the candlesticks.

Our results support, in general, the highly skeptical evaluation of the possibility to exploit correlations in the share prices for making profits although Table 5 strongly suggests that one still can succeed if proper questions are asked as regards the forecast.

We must admit here that the way the technical analysts use the Japanese candlesticks in their studies of share prices dynamics is quite different from that adopted here. Indeed, they use only some very special combinations of the sticks, the most significant ones, for their trading tactics. Also, the similarity between the patterns strongly depends on the context (i.e. the general situation on the market). We plan to investigate these matters in the further work.

References

1. Anderson, T.W.: The Statistical Analysis of Time Series. Wiley, New York (1971) zbmath.org/0225.62108
2. Haykin, S.: Adaptive Filter Theory. Prentice-Hall, Englewood Cliffs (1986)
3. Durbin, J., Koopman, S.J.: Time Series Analysis by State Space Methods. Oxford University Press, Oxford (2001) zbmath.org/0995.62504
4. Agrawal, R., Psaila, G., Wimmers, E.L., Zait, M.: Querying shapes of histories. In: Proceedings of the 21st International Conference on Very Large Databases. Zurich, 11-15 September 1995
5. Camerra, A., Palpanas, T., Shieh, J., Keogh, E.: iSAX 2.0: Indexing and mining one billion time series. In: 2010 IEEE International Conference on Data Mining, ICDM, pp. 58-67 (2010)
6. Keogh, E., Chakrabarti, K., Pazzani, M.: Locally adaptive dimensionality reduction for indexing large time series databases. In: Proceedings of ACM SIGMOD Conference on Management of Data, Santa Barbara, 21-24 May 2001
7. Keogh, E., Pazzani, M.: An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback. In: Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining. New York, 27-31 August 1998
8. Liao, T.W.: Clustering of time series data—a survey. Pattern Recognit. 38, 1857 (2005) doi:10.1016/j.patcog.2005.01.025
9. Geurts, P.: Pattern extraction for time series classification. In: Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery, Freiburg, 2001
10. Radovanovic, M., Nanopoulos, A., Ivanovic, M.: Time-series classification in many intrinsic dimensions. In: Proceedings of the SIAM International Conference on Data Mining, SDM, Columbus, Ohio, 29 April-1 May 2010

11. Keogh, E., Lin, J., Fu, A.W.: HOT SAX: Efficiently finding the most unusual time series subsequence. In: Proceedings of the 5th IEEE International Conference on Data Mining, Houston, 27-30 November 2005
12. Preston, D., Protopapas, P., Brodley, C.: Event discovery in time series. In: Proceedings of the SIAM International Conference on Data Mining, SDM, Sparks, Nevada, 30 April-2 May 2009
13. Lin, J., Williamson, S., Borne, K., DeBarr, D.: Pattern recognition in time series. In: Way, M., Scargle, J.D., Kamal, M.A., Srivastava, A.N. (eds.) Advances in Machine Learning and Data Mining in Astronomy. J. Chapman and Hall/CRC Press, Boca Raton (2012)
14. Wiliński, A., Zabłocki, M.: The investment strategy based on the difference of moving averages with parameters adapted by machine learning. In: Wiliński, A., El Fray, I., Pejaś, J. (eds.) Soft Computing in Computer and Information Science, Advances in Intelligent Systems and Computing Series, Vol. 342, Springer, pp. 207-225 (2015) 10.1007/978-3-319-15147-2_18