Proceedings of the International Multiconference on Computer Science and Information Technology pp. 159–167
ISSN 1896-7094

General Structure of T-Lymphocyte Applied to Immune-Based Event Detection in Financial Time Series

Tomasz Pelech-Pilichowski[1] and Jan T. Duda[2]

AGH-University of Science and Technology

Abstract. The paper is focused on the T-lymphocyte construction applied to immune-inspired event detection in financial time series. The goal is to recognize symptoms of abrupt changes of long-time mean value of many processed series. The task of the T-lymphocyte is to distinguish between 'healthy' and 'illness' states through examining individual series, with algorithms based on weak and rigorous statistical tests (detailed operation of detection is showed). General structure of the T-lymphocyte algorithm is illustrated. Comparison of the number of detected symptoms is presented.

1 Introduction

Information on early symptoms of significant changes in company environment may be very helpful in decision making process. In the paper [14] we proposed to gain such information by analysis of short-term prediction efficiency for selected financial series (e.g. stock quotations relevant to the managed company operation). For typical financial series autocorrelation of daily increments is statistically insignificant [8]. Hence, per average, the minimum variance error prediction is achievable simply by extrapolation of an averaged returns, i.e. by the Zero Order Prediction (ZOP) model [4]. To handle local changes in dynamics of a series volatility GARCH approach is usually recommended [2], which exploits AR type models describing both, the series expected value and its variance. Also neural network based algorithms are proposed [9]. However such techniques need a long time data for identification or learning. Moreover, from our point of view, a reduction of the prediction error is less important than the fast detection of 'an unusual' behavior of the series. Hence, quickly adaptable predictors – like Holt model [12] – seem to be more useful. In our earlier papers [14, 16, 13, 15] we have incorporated a set of Holt predictors of different adaptive properties into the immune based approach [11, 19]. Parameters of the predictors were adjusted genetically in such a way to make each of them the most suitable for a particular type of changes in the series shape. We showed that application of Artificial Intelligence (AI) techniques makes possible to improve prediction efficiency (error dispersion), when compared to classical applications of the Holt model [16].

One of a novel nature-inspired AI approach ([17]) is based on behavior of immunological system (Artificial Immune Systems) [10, 5, 11, 19, 20]. Practical application of this paradigm to a particular problem consists in rough algorithmic projection of relevant immunological mechanisms, with respect to the problem properties.

The paper refers to an original idea of implementation of the immune based approach to early detection of significant events in financial time series. The concept is to follow only the main rules of natural immune systems, where system protection activities are decomposed into two stages: first – fast detection of aliens ('nonselfs') in the system (performed by T-lymphocytes), and then – more time-consuming recognition of the alien type (illness) and the system recovery (B-lymphocyte). Such a decomposition may be an inspiration to improve statistical algorithms aimed at classification and segmentation of time series.

In this paper the attention is focused on construction of the first stage detection procedures, being viewed as the T-lymphocyte, dedicated to operation on financial time series. One exploits the fact mentioned above, that during usual ('healthy') behavior of a consider series no autocorrelation of daily returns may be expected [4]. Hence, a sequence of larger deviations from a long term average value may be treated as an early symptom of an 'illness'. Simple statistical criteria and rules are proposed to detect such an illness and start further recognition of its symptoms, i.e. activate a B-lymphocyte. More advanced rules embedded in the B-lymphocyte procedure are beyond the scope of this paper.

An example of the proposed algorithm application to event detection and short-range prediction of WIG20 index and quotations of three companies from Warsaw Stock Exchange is shown.

2 Immune-Based Approach to Time-Series Prediction

Considering financial time-series monitoring, one may take that more frequent typical situations mean 'healthy' state of the series, making possible application of forecasting techniques based on long-term observation of the process. For daily returns prediction, minimum variance error in such situations is produced by ZOP method [4], provided that the healthy series parameters (expected value and variance) were estimated with incidental nonstationary subseries data being omitted. It needs quick detection of local (in time) autocorrelation of subsequent data, and employing additional elimination criteria. On the other hand a sequence of such temporary events may be a symptom of significant, long term changes in dynamic behavior of the series (mainly observed as a rapid change in long term trends). Usually this is the case when such events are also observed in other series. Such a situation may be viewed as an illness, which needs dedicated reaction aimed at evaluation of the new parameters enabling for reliable long term trend prediction.

In natural immune systems, T-lymphocytes recognize pathogens ('nonselfs') among cells called 'selfs' and then activate B-lymphocytes, which in turn destroy the antigen through the first or the second immune response. To imitate such abilities we define a system cell as a subseries of presumed length parameterized with long-term mean value y_{HS} and standard deviation σ_{HS}. The lymphocyte may be represented by a classifier of the cells (subseries). The cell is treated as the 'self' if y_{HS} and σ_{HS} are adequate for long-term prediction of the series. This property has to be confirmed by consecutive analysis of current deviations of the samples from y_{HS}. Appropriate tests used for this purpose need statistics and arbitrary parameters to be adjusted (adapted) on line, depending on reached efficiency (e.g. measured by false alarm frequency, and number of missed events).

To make possible such adaptation we consider three 'healthy' situations:

1. A sequence of latest deviations is typical, i.e. it fits the long-term healthy probability distribution, hence the samples may be directly used to update long-term average (no event occurs).
2. Some of the latest deviations are of low probability but statistically acceptable. Such a cell behavior, referred to as weak event ('W'), should be under control (checking of occurrence frequency), but the sequence samples may be used for updating y_{HS}.
3. In a short time interval the sequence does not fulfill the series statistics (it is nontypical). Hence it should not be used for updating y_{HS}, but likely it does not announce long-term changes. Hence it is treated as a significant, but still admissible event ('S'-type cell).

If the subseries doesn't fit one of the above classes it is recognized as nonself.

Series processing, aimed at classification of its behavior into self or nonself, and detailed criteria for distinguishing between (1-3) 'self' subclasses, are explained in fig. 1. The processing splits into two branches: for healthy (left hand side) and ill (right hand side) state of the series, switched in the block 1 based on a binary attribute (a flag labeling the series state).

If the healthy state was retained in the loop run before (the flag 'ill' is switched off), the Student's test is applied for consecutive samples treated as random independent data (block 2.1). The goal is to detect a given number of significant deviations (in the t-statistics meaning) in series of the same sign or a bit longer sequence of significant deviations of any sign. To get the critical value t_{cr} for t-test we fixed the significance level $\alpha = 0.025$. Each relative deviation exceeding t_{cr} is recorded to check overall 'healthy' statistics of the considered series. In the example discussed in sequel we have assumed the sequence length 3 and 5, which is a compromise between detection reliability and delay.

Critical value for an individual deviation in the sequence of L_{sk} length is calculated as:

$$\alpha = 1 - \prod_{k=1}^{L_{sk}} p_k \ , \quad t_{thr} = \arg\operatorname{erf}\left((1-\alpha)^{\frac{1}{L_{sk}}}\right) \tag{1}$$

Samples are treated as stationary of type (1) if no significant deviation was found.

When the sequence of significant changes is recognized, one uses the Holt predictor (block 4) to evaluate a current mean value y_{av} and its one-step ahead prediction. The following formulae are employed [12, 16, 15]:

$$y_{pred_{t+1}} = y_{av_t} + T_t + 1/2R_t^2 \tag{2}$$

$$y_{av_t} \stackrel{\Delta}{=} \alpha \cdot y_t + (1-\alpha)\left(y_{av_{t-1}} + T_{t-1} + 1/2 \cdot R_{t-1}\right) \tag{3}$$

$$T_t \stackrel{\Delta}{=} \beta \cdot \Delta y_{av_t} + (1-\beta)T_{t-1} \ , \quad R_t \stackrel{\Delta}{=} \gamma \cdot \Delta^2 y_{av_t} + (1-\gamma)R_{t-1} \tag{4}$$

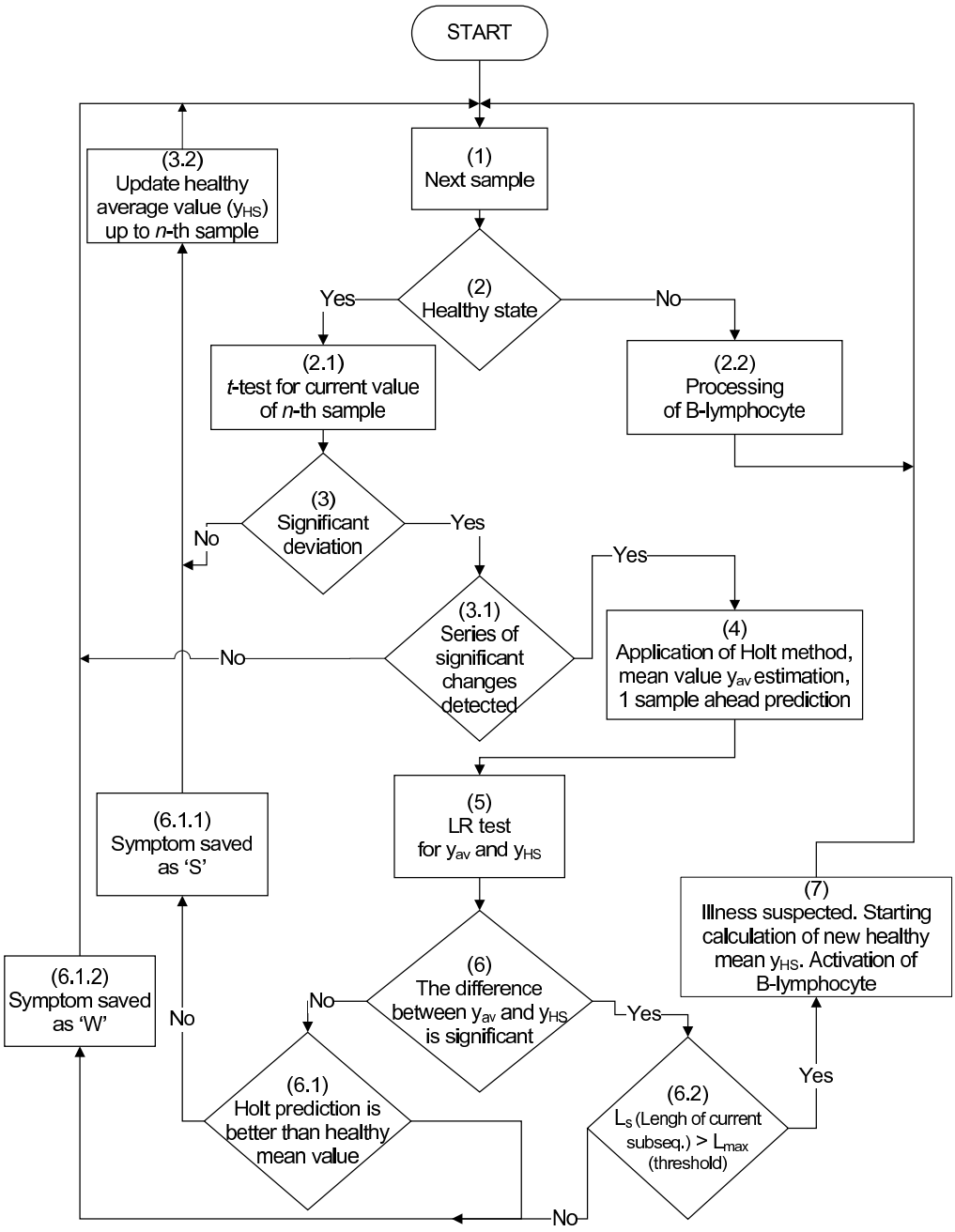

Fig. 1. Block-diagram of immune-like processing of time series

$$\Delta y_{av_t} = y_{av_t} - y_{av_{t-1}} \ , \quad \Delta^2 y_{av_t} = \Delta y_{av_t} - \Delta y_{av_{t-1}} \tag{5}$$

They involve three smoothing constants α, β, γ [12, 1], which are to be adjusted by a genetic way [15] in a range from 0 to 1, to achieve more accurate prediction. As the starting value y_{av_0} the healthy average y_{HS} is used.

The next step (blocks 5 and 6) is to answer the question, does the local expected value $E\{y_n\}$ is closer to y_{av} than to the healthy average y_{HS}. To this aim we apply the one-side likelihood ratio (LR test), i.e. the Page-Hinkley method [3, 7, 15]. It consists in checking of the H_0 hypothesis assuming, that expected current mean value is the healthy average ($E\{y_n\} = y_{HS}$) against H_1 hypothesis: $E(\{y_n\}) = y_{av}$ (abrupt change in a series).

The H_0 hypothesis is rejected when the condition (6) is fulfilled :

$$\Lambda_{rn} = \prod_{a}^{b} \frac{\mathrm{p}\,(y_k|H_1)}{\mathrm{p}\,(y_k|H_0)} \ , \quad S_r^n = \log \Lambda_{rn} \tag{6}$$

It leads to the following criterion:

$$\max_{n-L<r\leq n} \sup_{\nu} S_r^n(y_{HS}, \nu) > \lambda \tag{7}$$

where $\sup_\nu S_r^n(y_{HS}, \nu)$ denotes cumulative sum of data in the subseries Y_{rn} containing samples from r to n, λ is an adjustable parameter (proportional to σ_{HS}^2), and ν denotes jump size ($\nu = y_{av} - y_{HS}$).

$$\frac{\sup_\nu S_r^n(y_{HS}, \nu)}{\sigma_{HS}^2} = \frac{\nu}{\sigma_{HS}^2} \cdot \sum_{k=r}^{n} \left(y_k - y_{HS} - \frac{\nu}{2} \right) = \log\left(\Lambda_{rn}\right) \tag{8}$$

The maximum likelihood function value $\sup_\nu S_r^n(y_{HS}, \nu)$ is calculated for $r = n - L + 1, \ldots, n$, where L is the number of samples from the beginning of the analyzed sequence $L = 3, 4, 5, \ldots L_{max}$, L_{max} is the minimum life-time (threshold) of the possible ill state. Having current average y_{av} evaluated with the Holt formula (eq. 3) we may apply the following criteria:
For positive change:

$$Up_j = \sum_{k=n-L+1}^{j} \left(y_k - \left(y_{HS} + \frac{\nu}{2} \right) \right) \ , \qquad \text{for} \quad j = n - L + 1, \ldots n, \tag{9}$$

$$H_1 : \text{when} \max_{n-L<j<n} Up_j > (\lambda/\nu), \quad r = \arg\left(\max_{n-L<j<n} Up_j \right) \ ,$$

and for negative change:

$$Un_j = - \sum_{k=n-L+1}^{j} \left(y_k - \left(y_{HS} - \frac{\nu}{2} \right) \right) \ , \qquad \text{for} \quad j = n - L + 1, \ldots n, \tag{10}$$

$$H_1 : \text{if} \max_{n-L<j<n} Un_j > (\lambda/\nu), \quad r = \arg\left(\max_{n-L<j<n} Un_j \right) \ .$$

where r is an estimate of the jump time instant.

The adjustable parameter λ decides on a false alarm probability, while L_{max} should be selected in such a way to reach acceptable probability of a change omitting [6]. Within the T-lymphocyte moderately high values for these probabilities must be taken, because of relatively short L values, as L_{max} determines the maximum admissible delay for detection of an illness (we have taken $L_{max} = 7$).

The Page-Hinkley procedure is also employed as B-lymphocyte activation (block 7) with more restricted probabilities of false alarm and a change omitting. In this case, illness state is detected (length of sequence of significant deviations exceeds assumed threshold – L_{max}). Calculation started in this step are aimed at doing estimation of new healthy mean y_{HS}.

3 Input Time Series and Calculation Results

Four time series from Warsaw Stock Exchange (WIG, Poland) has been used as example data, to analyze the immune-based algorithm properties. We have chosen individual quotations of KGHM, PEKAO,

TPSA, and WIG20 index (i.e. averaged quotations of the 20 biggest and the most liquid companies) from 1994-04-20 to 2006-04-13 (1822 samples).

To perform short-time prediction with Holt method and directly compare results to zero-order-prediction, the forecasting period τ was fixed to 1 day. Holt model parameters (α, β, γ) has been adapted in a genetic way.

Instants of detection of 3 and 5 deviations are compared to original time series in fig. 2. Notice that only long term changes are confirmed by Page-Hinkley (LR) test. Configuration of the series of 5 significant deviations is very interesting as a signal of an anxiety at the stock exchange. It may be seen in figure 2 that before rapid changes in long term trends, five deviations were often detected. No such relation is observed when considering series of 3 deviations.

Detailed immune-based processing of time series at selected situations is depicted in figures 3 and 4. They show different, subsequent stages of the immune-like symptom detection procedure. Updating 'healthy' average value is illustrated in figure 3, while new mean value after 'nonself' subseries detection is visible in fig. 4. In the both cases, the Holt method is useful for fast computing of temporary mean value, being the input parameter of Page-Hinkley test. It is also compared to ZOP prediction to distinguish between 'S' and 'W' symptoms (fig. 3).

Efficiency of symptom detecting with T-lymphocyte is illustrated in table 1. In each analyzed series the number of detected symptoms is similar (it depends on randomness inherent in genetic operators being applied). However, the number of confirmed short-range changes of mean value with Page-Hinkley test and B-lymphocyte activation is significantly different (see the 7th and 8th row). A coincidence of symptoms of long-term mean-value changes is seen (approximated number of continued series of deviations) which indicates a need of processing of multiple series to specify more reliable symptoms.

Table 1. Characterization of T-lymphocyte symptoms detection efficiency

Time series (total number of samples: 1822)	WIG20	KGHM	PEKAO	TPSA
Number of detected significant deviations				
1 deviation	55	70	56	66
2 deviations	28	30	29	24
3 deviations	17	17	14	16
5 deviations	22	21	28	22
Number of healthy data	1682	1667	1683	1674
Continued series of deviations (number of iterations)	49	50	62	42
Number of confirmed symptom with LR test	11	13	21	6
Number of B-lymphocyte activation	0	2	2	0
Number of detected symptoms 'S'	35	33	33	32
Number of detected symptoms 'W'	3	4	8	4

It may be seen in figure 2 and table 1 that applying only weak (fast) statistical test (confirmation with Page-Hinkley method omitted) to construction of T-lymphocyte is an adequate way to reduce computation time. In fact, only about 30% of detected large deviation sequences are confirmed with LR test.

Notice, that the proposed algorithm is able to detect correctly not only evident events, but also hardly visible ones (see fig. 2).

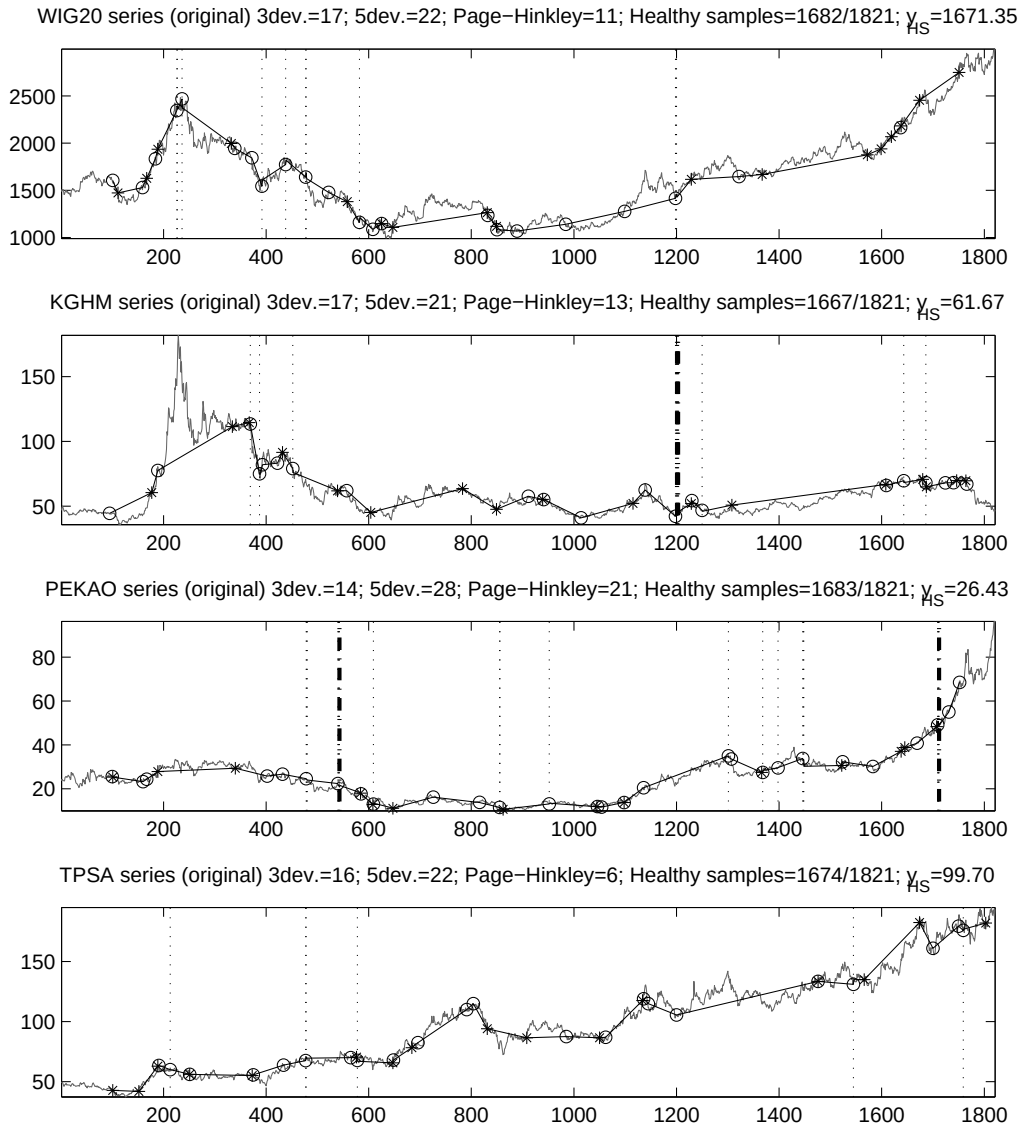

Fig. 2. Recognized series of 3 ('*') and 5 ('o') deviations (original time-series). Solid lines show final 'healthy' mean. Page-Hinkley confirmation (see block 6 in fig. 1) – vertical dotted line; B-lymphocyte activation (block 7) – vertical bolded lines

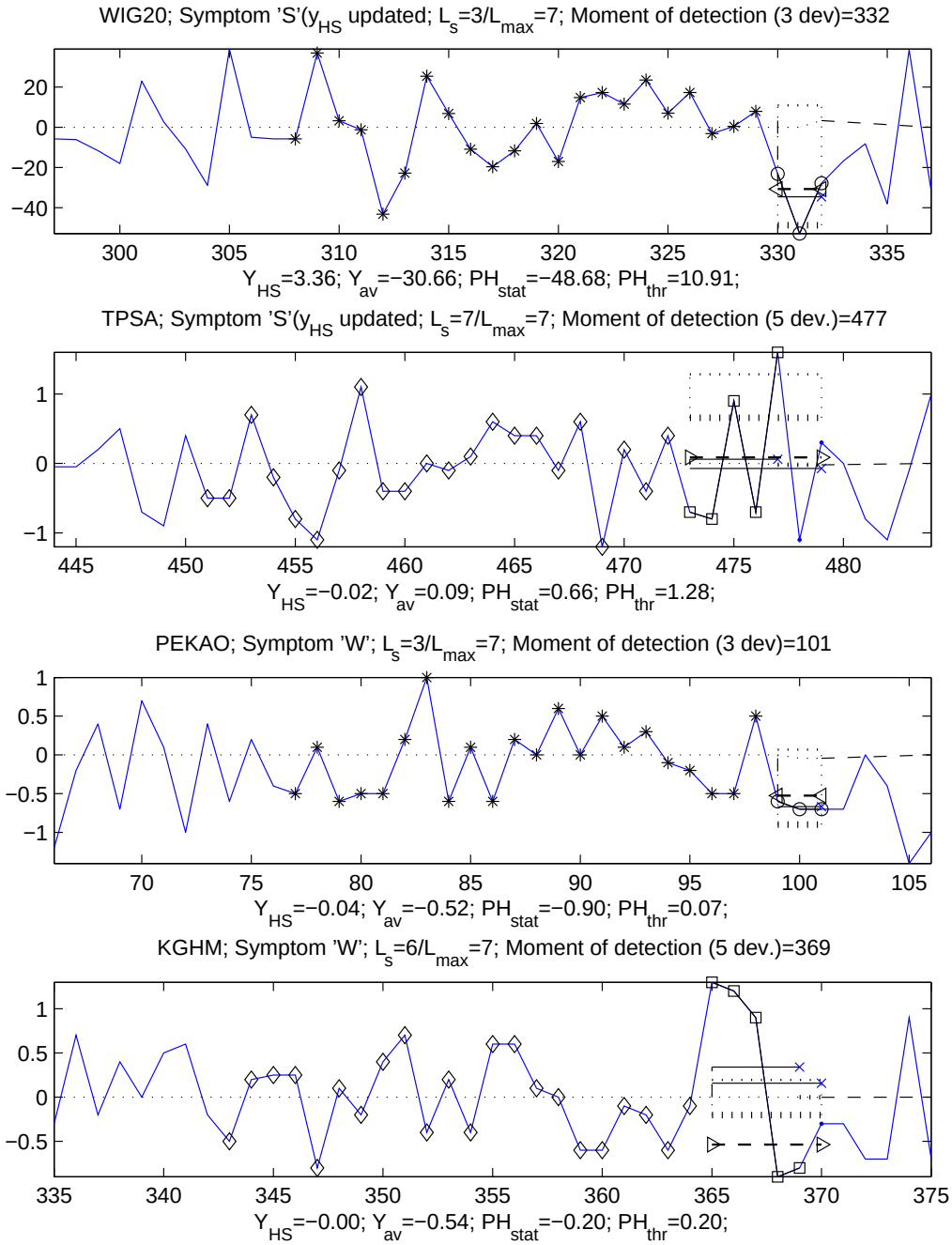

Fig. 3. Detailed operation of recognizing symptom 'S' (upper sub-figure; see block 6.1.1 in fig.1) and recognizing symptom 'W' (lower sub-figure; block 6.1.2). '*'/'diamond' sign – data taken to calculate t-statistics (for 3/5 dev. respectively); 'o'/'square' – operation after exceeding the t-test threshold; '.' – samples of continued series ($L > 3/L > 5$); y_{av} – broken line started and ended with 'triangle'

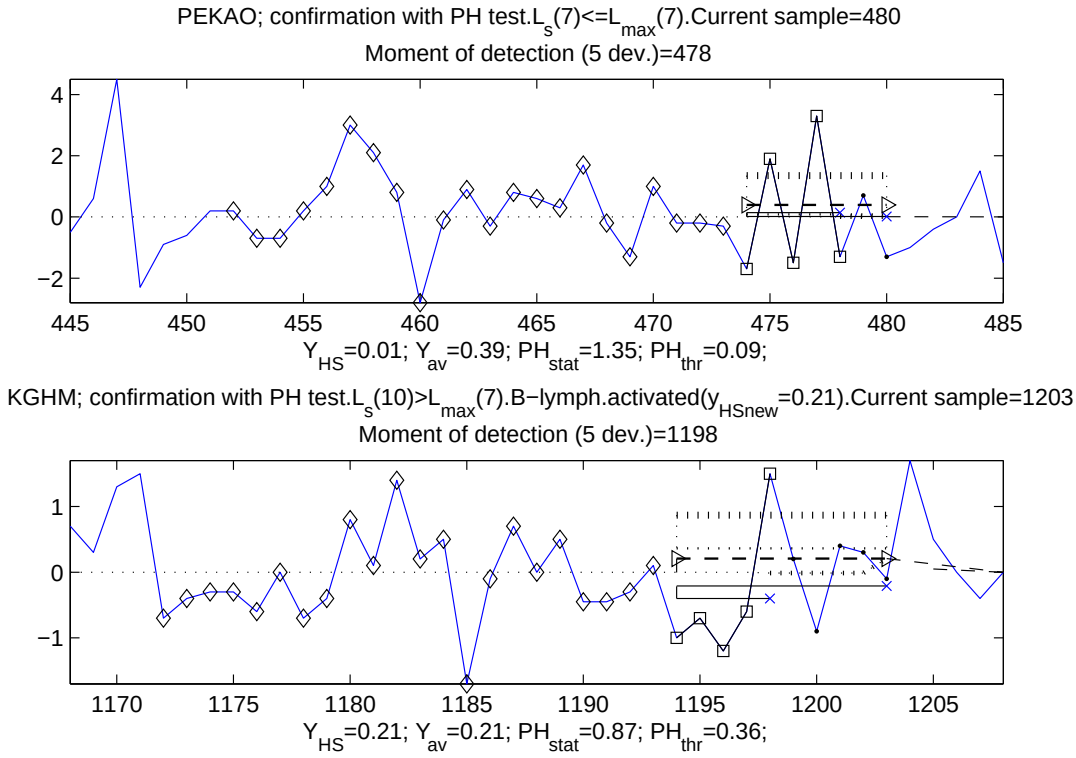

Fig. 4. Detailed operation of confirmation of the series with LR test (upper sub-figure) and recognizing pathogen (lower sub-figure). Bolded and dotted line – PH test value; thin, broken line – the threshold value (see top and bottom edge of depicted windows); window width – the length of analyzed series of significant short-time change of mean value; y_{HS} – dotted (up to symptom detection time) and broken line (after detected symptom); broken line – new healthy mean value after B-lymphocyte activation (started at the moment of calculation of the new value); solid lines ended with the sign 'x' – short-range mean values of samples (recognized as significant deviations, for the first 3/5 and consecutive deviations)

4 Conclusions

Natural immune systems may be viewed as an inspiration for construction of adaptive algorithms aimed at early detection of significant changes in mean value of time series. The classification of the series into 'healthy' and 'ill' subseries by current analysis of short-range changes of mean value (series of significant deviations) was found as suitable to detect significant symptoms.

We have found that the first stage of the event detection procedure (T-lymphocyte) may be based on simple statistical tests instead of likelihood ratio criteria. It makes possible to recognize early symptoms of large trend changes in a less time consuming way.

Further improvements of the recognition reliability need specification of rules for recording the early symptoms configurations, detected in a number of examined series. This will result in large number of T-lymphocytes capable for recognizing different cells, like in natural systems.

Application of statistics involving adjustable parameters (significance level, threshold values, window widths, etc.) seems to be a promising way for enhancement of the presented method.

Further research will be focused on the structure of another immune-like mechanism and rules (B-lymphocyte, immune memory). The main target is to reduce time delay of detection (compared to standard statistical tests) and to achieve effective event detection and elimination of 'false alarms'.

References

1. Augustynek A., Duda J.T., Klemiatio M.: *Matematyczne prognozowanie cen miedzi na gieldzie LME,* XV. Problems of Mechanical Engineering and Robotics **12** (2003) WIMiR AGH Krakow.
2. Asokan M. V., Chenouri S., Mahmoodabadi A. K.: *ARCH and GARCH models,* Department of Statistics and Actuarial Sciences (2001) University of Waterloo.

3. Basseville M.: *Detecting Changes in Signals and Systems – Survey,* Automatica, vol. **24,** No. 3, (1988) 309-326.
4. Box G. E. P., Jenkins G. M.: *Analysis of Time Series,* PWN (1983) Warszawa.
5. Dasgupta D. (Edt.): *Artificial Immune Systems and Their Applications,* (1999) Springer-Verlag.
6. Duda J. T.: *Dobor parametrow algorytmu Page'a-Hinkleya przy ustalonych prawdopodobieństwach I i II Rodzaju,* Unpublished paper (2005) AGH-UST.
7. Duda J. T.: *Modele matematyczne, struktury i algorytmy nadrzednego sterowania komputerowego,* (2003) AGH-UST University Press.
8. Duda J. T., Augustynek A.: *On possibilities of improvement of short-term predictions of stock indices with regression models. Company Management in Conditions of European Integration – Part 2,* Economy, Informatics and Numerical Techniques. Ed. M. Czyz and Z. Cieciwa (2004) AGH-UST University Press.
9. Dunis Ch.L: *Forecasting Financial Markets. Exchange Rates, Interest Rates and Asset Management,* (1996) John Wiley & Sons.
10. Farmer J. D., Packard N., Perelson A.: *The immune system, adaptation and machine learning Physica D,* vol. **22** (1986) pp. 187–204.
11. Hofmeyr S. A., Forrest S.: *Immunity by Design: An Artificial Immune System,* Proceedings of the Genetic and Evolutionary Computation Conference (1999) San Francisco.
12. Holt C. C.: *Forecasting seasonals and trends by exponentially weighted moving averages,* Carnegie Institute of Technology (1957) Pittsburgh, Pennsylvania.
13. Pelech T.: *Adaptive Holt's Forecasting Model Based on Immune Paradigm. Problemy osvoenia poleznyh iskopaemyh,* Zapiski Gornogo Instituta (2006) Sankt Petersburg State Mining Institute.
14. Pelech T, Duda J. T.: *Application of immune paradigm to monitoring of stock indices. Problems of Mechanical Engineering and Robotics,* No. **3** (2005) AGH-UST University Press.
15. Pelech T., Duda J. T.: *Event detection in financial time series by immune-based approach,* Intelligent Information Processing and Web Mining. Advances in Soft Computing (2006) Springer-Verlag.
16. Pelech T, Duda J. T.: *Immune Algorithm of Stock Rates Parallel Monitoring,* Information Systems and Computational Methods in Management. Ed. J. T.Duda (2005) AGH-UST University Press.
17. Rutkowski L., Siekmann J. H., Tadeusiewicz R, Zadeh L. A.: *Artificial Intelligence and Soft Computing,* ICAISC (2004) Springer.
18. Somayaji A., Hofmeyr S., Forrest S.: *Principles of a Computer Immune System. New Security Paradigms Workshop,* (1998) Langdale.
19. Timmis J., Knight T., De Castro L. N., Hart E.: *An Overview of Artificial Immune Systems. Computation in Cells and Tissues: Perspectives and Tools for Thought,* (2004) Springer.
20. Wierzchon S. T.: *Sztuczne systemy immunologiczne. Teoria i zastosowania,* (2001) Wyd. Exit.