Interval forecasting of cyberattack intensity on informatization objects of industry using probability cluster model

To cite this article: Y M Krakovsky et al 2018 J. Phys.: Conf. Ser. 1015 032088

View the article online for updates and enhancements.
Interval forecasting of cyberattack intensity on informatization objects of industry using probability cluster model

Y M Krakovsky\(^1\), A N Luzgin\(^2\), E A Mikhailova\(^3\)

\(^1\) Irkutsk State University of Railway Transport, 15, Chernyshevsky St., Irkutsk, 664074, Russia
\(^2\) Administration of Irkutsk City, 14, Lenina St., Irkutsk, 660025, Russia
\(^3\) Chita Institute of Baikal State University, 56, Anokin St., Chita, 672000, Russia

E-mail: alexln@mail.ru

Abstract. At present, cyber-security issues associated with the informatization objects of industry occupy one of the key niches in the state management system. As a result of functional disruption of these systems via cyberattacks, an emergency may arise related to loss of life, environmental disasters, major financial and economic damage, or disrupted activities of cities and settlements. When cyberattacks occur with high intensity, in these conditions there is the need to develop protection against them, based on machine learning methods. This paper examines interval forecasting and presents results with a pre-set intensity level. The interval forecasting is carried out based on a probabilistic cluster model. This method involves forecasting of one of the two predetermined intervals in which a future value of the indicator will be located; probability estimates are used for this purpose. A dividing bound of these intervals is determined by a calculation method based on statistical characteristics of the indicator. Source data are used that includes a number of hourly cyberattacks using a honeypot from March to September 2013.

1. Introduction
Currently, cyber-security issues of informatization objects of industry (IOI) are categorized by the details of their development. Most IOI are built on a base of modern technologies for information exchange and process management. As a result of functional disruption of IOI, an emergency may arise related to loss of life, environmental disasters, major financial and economic damage, or disrupted activities of cities and settlements. In this regard, research is needed on the development of new protection methods against cyberattacks [1,2]. Here, the most promising research direction is cyberattack forecasting based on machine learning [3-6]. Thus, in the direction related to protection against cyberattacks, in addition to assessing different risks and using traditional protection systems, attention should be paid to cyberattack forecasting [3].

It should be noted that in the past few years, researchers have become interested in probabilistic forecasting [4,6,7]. This can be explained by the fact that probabilistic forecasts make it possible to obtain not only forecasts of future events, but also probabilistic estimates of these events. One type of probabilistic forecasting is interval forecasting: this involves forecasting of one of the two predetermined intervals in which a future value of an indicator will be located. Probability estimates are used for this purpose. A dividing bound of these intervals is determined by a calculation method based on statistical characteristics of the indicator.
In this paper, source data are used that include a number of hourly cyberattacks received using a honeypot from March to September 2013 for the group ‘zeppo-norcal’ [8]. For analyzing ‘fronts’ of a beginning and end of cyberattacks, the author of [8] recommends applying a transformation by the simple moving average to the source data. In addition, this approach reduces the influence of outliers and high-frequency noise in the source data. Taking into account these recommendations, the simple moving average was applied to the source data with a window width of 4 hours, and a corresponding indicator (ZNI) has been obtained. To obtain statistical characteristics of the indicator, its class was determined by the method described in [9]. Subsequently, a study of cyberattack interval forecasting using an algorithm based on a probabilistic cluster model (PCM) was carried out. The obtained results demonstrate high accuracy of the interval forecasting based on applying the algorithm.

2. Formalization of a cyberattack indicator

Each indicator was formalized as the time series:

$q = \{q_t; t \in t\}$.  \hspace{1cm} (1)

Here, $q_t$ is the value of the indicator at the discrete moment of time $t$; $t \in t$; $t = \{1, ..., n\}$; $n$ is the number of values. For ZNI $n = 4374$.

Figure 1 shows a graph of the final 500 values of ZNI.

![Figure 1. A graph of the final 500 values of ZNI](image)

Calculations show that ZNI is a first class indicator [9]. This category of the indicator is the most difficult to use for interval forecasting implementation (in the statistical sense) compared with indicators of other classes.

3. Interval forecasting formalization

Let us define the range $(q_{\text{min}}; q_{\text{max}})$ of possible values $q$ (1) and an inner point (a dividing bound of intervals) $q_{\text{int}} (q_{\text{min}} < q_{\text{int}} < q_{\text{max}})$ and construct two intervals:

$I^- = (q_{\text{min}}; q_{\text{int}}), I^+ = (q_{\text{int}}; q_{\text{max}})$.  \hspace{1cm} (2)

Let us calculate $q_{\text{int}}$ in (2) as follows:

$q_{\text{int}} = \text{med}(q) + \beta \cdot \text{med}(|q - \text{med}(q)|) = \text{med}(q) + \beta \cdot \text{MAD}(q)$.  \hspace{1cm} (3)

Here $\beta \in [-1; 1]$ is a coefficient (the value is set beforehand); med(·) is median; and MAD(·) is median absolute deviation.

At time $t = n - 1$, it is necessary to identify the interval (2) in which future (unknown) value $q_{t+p}$ will be located. The following estimates of probabilities are required: $p_{t^-+p}$ and $p_{t^++p}$ where $p = 1, ..., r$ is the look-ahead period, $p_{t^++p}$ is the probability that indicator future value $q_{t+p} \in I^+$ and $p_{t^-+p}$ is the probability that the indicator future value is $q_{t+p} \in I^-; p_{t^++p} + p_{t^-+p} = 1$. 
Let \( \hat{\rho}_{t:p} \) and \( \hat{\rho}_{t:p}^- \) be probability estimates of \( \rho_{t:p}^+ \) and \( \rho_{t:p}^- \). The interval forecasting is carried out according to the following rules: future value \( q_{t+p} \in I^+ \) if \( \hat{\rho}_{t:p}^+ > \hat{\rho}_{t:p}^- \) and future value \( q_{t+p} \in I^- \) if \( \hat{\rho}_{t:p}^+ \leq \hat{\rho}_{t:p}^- \).

It is necessary to consider expression (3) in more detail and provide more detailed explanations.

Let us note that indicator (1) is considered as a random variable with some unknown probability distribution function, whose robust characteristics are the location parameter and the scale parameter. These parameters characterize the center of grouping of random variable values and the degree of deviation of random variable values relative to this center [9].

In expression (3), the first term is the location parameter estimate and the second term is the scale parameter estimate that is multiplied by \( \beta \). Using \( \beta \), the value of \( q_{\text{int}} \) can be specified in the range from \( \text{med}(q) - \text{MAD}(q) \) to \( \text{med}(q) + \text{MAD}(q) \). For \( \beta = -1.0 \), interval \( I^+ \) contains approximately 75% all values of \( q \) (1) and for \( \beta = 1.0 \) - about 25%. It is assumed that the value variation of \( q_{\text{int}} \) in this range is sufficient for further research.

It should be noted that cyberattack intensity is the total number of these attacks per unit of time (in this case, this is an hour). With this in mind, \( q_{\text{int}} \) (3) is defined as the pre-set level of cyberattack intensity. The higher the value of \( q_{\text{int}} \) (3), the more intense the cyberattack must be in order to get into interval \( I^+ \) (2). It is necessary to set the value of \( \beta \) for which interval forecasts of \( q_{t+p} \in I^+ \) will occur, to indicate the necessity for additional protective measures. Then, forecasts of \( q_{t+p} \in I^- \) will be considered as a usual situation and will not require additional attention. The lower the value of \( \beta \), the more often the forecasts of \( q_{t+p} \in I^+ \) will be made, and vice versa. Of course, in each specific case and in each particular organization, the values of \( \beta \) can be chosen by experts and, consequently, these values can be different.

4. Specificity of constructing a probabilistic cluster model
Let us suppose that \( t = n \) and there is a sequence of values \( q_{t-f+1}, \ldots, q_t \) in a number of \( f \). Let us create vector \( z = (q_{t-f+1}, \ldots, q_t) \). Let \( y_{t:p} \) be a dependent variable (or a response) whose true value is unknown and that can take only two possible values: \( y_{t+p} = 1 \) if \( q_{t+p} \in I^+ \) and \( y_{t+p} = -1 \) if \( q_{t+p} \in I^- \).

Performing interval forecasting using \( z \) requires making a forecast of \( y_{t+p} \) based on probability estimates that \( q_{t+p} \in I^+ \) and \( q_{t+p} \in I^- \). Let us recall that if \( \hat{\rho}_{t:p}^+ > \hat{\rho}_{t:p}^- \), then \( y_{t+p} = 1 \), as well as \( y_{t+p} = -1 \).

Next, let us create a training set based on the values of \( q \) (1) for \( t = 1, \ldots, m \), where \( m = n - f - p + 1 \) (this value is chosen so that the responses’ values can be calculated based on pre-history values of the indicator):

\[
\begin{pmatrix}
q_1 & \cdots & q_{1+f-1} \\
q_m & \cdots & q_{m+f-1}
\end{pmatrix},
\quad y = (y_1, \ldots, y_m).
\]  

(4)

Here \( x \) is the matrix of dimensions \( m \times f \) (a training set); \( y \) is a vector of responses of size \( m \) (these responses are calculated based on the pre-history values of the indicator); and \( m \) is the number of training samples.

Each row \( i \) of matrix \( x \) corresponds to the response of \( y \) (4): \( x_i \rightarrow y_i \).

Finally, a PCM is trained using training set (4) and the interval forecasting is implemented using \( z \).

Let us consider a specific example of creating a PCM:
Let us define two variables \( n^+ = 0 \) and \( n^- = 0 \) and choose value \( i = 1 \). For \( x_i \) and \( z \), let us calculate the value of the non-parametric correlation coefficient of Tarsitano-Lombardo [10] \( r_i \). Let us approximate linearly the values of \( z \) by values of \( x_i \): \( \hat{z} = a \cdot x_i \). The value of \( a \) is estimated by the Thail-Sen method [11]. If \( a \cdot (q_{i+f-1+p} - q_{i+f-1} + q_n > q_{\text{int}} \), this is the clustering rule for the raw data for two clusters), then \( n^+ = n^+ + \max(0, r_i) \); otherwise, \( n^- = n^- + \max(0, r_i) \). Using \( n^+ \) and \( n^- \) for all values \( i \), let us find probability estimates as follows:
Thus, the algorithm for the interval forecasting of the cyberattacks based on a probabilistic cluster model has only three parameters: $f$, $\beta$ and $p$.

5. Results and discussion
For analyzing the interval forecasting results of the cyberattacks, two scores are used:

$$ps = \frac{l}{u}, \quad bs = u^{-1} \sum_{t \in G} (\hat{\rho}_{t+p}^+ - v_{t+p})^2.$$  

Here $l$ is the number of true forecasts; $u$ is the total number of forecasts; $\hat{\rho}_{t+p}^+$ is the probability estimate of $\rho_{t+p}^+$ based on a selected model; $G$ is the set of values $t$ at which the forecast was made; $v_{t+p}$ is the outcome of the event (equal to 1 if $q_{t+p} \in I^+$ and equal to 0 if $q_{t+p} \in I^-$); $ps$ is the frequency of true forecasts [9]; $bs$ is the Brier score (the measure of the forecasting accuracy) [12,13]; and $0 \leq ps \leq 1$; $0 \leq bs \leq 1$. The higher the value of $ps$ and the lower the value of $bs$, the better the interval forecasting accuracy.

Scores (6) were estimated by the leave-one-out cross-validation using the training set (4). That is, rows with the number of $t = 1, \ldots, m$ were sequentially excluded from matrix (4) and considered as vectors ($z$) for the interval forecasting.

For the implementation of all algorithms, R was used [14]. For the acceleration of some functions and procedures, C ++ was used. For integration of R and C ++, the Rcpp-package was used [15].

Table 1 shows the interval forecasting results of the cyberattacks one day ahead ($p = 1$) for different values of $\beta$ and values of $f$ for which the interval forecasting accuracy was maximum. The value of $f$ varied from 1 to 10. In addition, the frequency ($sa$) of the indicator values (1) is in $I^+$ for a given value of $\beta$.

| $\beta$ | $f$ | $ps$ | $bs$ | $sa$ |
|---------|-----|------|------|------|
| -1.0    | 5   | 0.83 | 0.13 | 0.75 |
| -0.8    | 5   | 0.81 | 0.14 | 0.69 |
| -0.6    | 5   | 0.79 | 0.15 | 0.64 |
| -0.4    | 5   | 0.78 | 0.16 | 0.58 |
| -0.2    | 5   | 0.77 | 0.16 | 0.52 |
| 0.0     | 5   | 0.78 | 0.16 | 0.47 |
| 0.2     | 7   | 0.79 | 0.15 | 0.41 |
| 0.4     | 7   | 0.80 | 0.15 | 0.37 |
| 0.6     | 7   | 0.80 | 0.14 | 0.33 |
| 0.8     | 7   | 0.81 | 0.13 | 0.29 |
| 1.0     | 7   | 0.82 | 0.12 | 0.25 |

According to Table 1, it is clear that the interval forecasting accuracy of the cyberattacks for selected indicator is quite high. The minimum value of $ps$ is 0.77, and the maximum value of $bs$ is 0.16. With a linear increase of value $\beta$, the frequency of the cyberattacks, which are in $I^+$, nonlinearly decreases from 0.75 to 0.25. Thus, cyberattack interval forecasting based on a probabilistic cluster model handles the task successfully for all values of $\beta$.

Taking into account the obtained results, the following scheme for protection against the cyberattacks using interval forecasting can be recommended:

1) Determine the value of $\beta$ and, accordingly, the value of $q_{int}$ (3). For example, this can be done empirically or by expert judgment. Here, an additional criterion for choosing a value of $\beta$ can be the frequency ($sa$) of the values of $q$ (1) that are in $I^+$. This value is easily calculated based on the values of $q$ (1);
2) Estimate the value of $f$ by leave-one-out cross-validation at which the value of $bs$ (6) is minimum; 

3) Make an interval forecast for the hour ahead ($p = 1$). If $q_{t+p} \in I^+$, then implement the additional measures (they can be automatic) against the cyberattacks. If $q_{t+p} \in I^-$, then work in a normal mode;

4) After one hour, add a new value to the end of $q$ (1) and return to step 2.

It should be noted that the pre-set level of the cyberattack intensity (3) can be periodically reviewed. For example, the growing number of cyberattacks on industrial control systems can be taken into account.

6. Conclusion

The analysis of contemporary scientific studies has shown that for several years, the scientific community has been conducting research on cyberattack forecasting by different methods with the objective of creating adequate instruments for protection against cyberattacks. This paper has studied the interval forecasting of cyberattacks based on a probabilistic cluster model. The authors’ approach allows carrying out cyberattack interval forecasting with a pre-set intensity level of cyberattacks. This approach demonstrates the high accuracy of cyberattack interval forecasts for selected data. At the same time, the necessary practical recommendations for an application of the interval forecasting results in the protection against cyberattacks in industrial control systems were formulated. It should be noted that similar studies about cyberattack interval forecasting are not known by the authors, so this work can be considered as a novel effort. The authors hope that the obtained results will find their practical application and theoretical development in the field of research investigating protection against cyberattacks.

References

[1] Mezzour G, Carley L R and Carley K M 2014 Global Mapping of Cyber Attacks (USA: Carnegie Mellon University)
[2] Han M L, Han H C, Kang A R, Kwak D I, Mohaisen A and Kim H K 2016 WHAP: Web-hacking profiling using Case-Based Reasoning IEEE Communication and Network Security (Philadelphia: IEEE)
[3] Gandotra E, Bansal D and Sofat S 2015 Computational Techniques for Predicting Cyber Threats Proc. Intelligent Computing, Communication and Devices (AISC, vol 308) ed J Kacprzyk (New Delhi: Springer )
[4] Kim Y, Lee T, In H, Chung Y J, Kim I and Baik D K 2005 A probabilistic approach to estimate the damage propagation of cyber attacks Proc. Information Security and Cryptology (ICISC, vol 3935) ed D Won, S Kim (Seoul: Springer)
[5] Wu J 2012 Cyber Attacks Prediction Model Based on Bayesian Network IEEE Parallel and Distributed Systems (Singapore:IEEE)
[6] Du H 2017 Probabilistic Modeling and Inference for Obfuscated Network Attack Sequences (New York: Rochester Institute of Technology)
[7] Kargapol'tsev S K, Krakovsky Y M, Lukyanov A V and Luzhin A N 2017 A dynamic updating algorithm of smoothing parameter values of probabilistic neural networks Far East Journal of Electronics and Communications 17(4) 909–914
[8] Data driven Security, http://datadrivensecurity.info/blog/posts/2014/Jan/blander-part1
[9] Kargapol’tsev S K, Krakovsky Y M and Luzhin A N 2016 Nonparametric classification of technical condition parameters based on shift and scale tests IEEE Industrial Engineering, Applications and Manufacturing ed A Radionov (Saint-Petersburg: IEEE) pp. 1–5
[10] Tarsitano A and Lombardo R A 2013 Coefficient of correlation based on ratios of ranks and anti-ranks Journal of Economics and Statistics 233 (2) 206–224
[11] Shah SH, Rehman A, Rashid T, Karim J and Shah M A comparative study of ordinary least squares regression and Theil-Sen regression through simulation in the presence of outliers Journal of Science Technology 5 137–142
[12] Scolnik R 2016 Predictive Accuracy Measures for Binary Outcomes: Impact of Incidence Rate and Optimization Techniques (USA: Florida State University)
[13] Rufibach K 2010 Use of Brier score to assess binary predictions Journal Clinical of Epidemiology 63(8) 938–939
[14] The R project of statistical computing, http://www.r-project.org
[15] Seamless R and C++ Integration, https://cran.r-project.org/web/packages/Rcpp/index.html