Application of the anomaly pattern in forecasting time series of project activity metrics

I.A.Timina[1], E.N.Egov[1], A.A. Romanov[1]

[1]Ulyanovsk State Technical University, SevernyVenecstreet32, Ulyanovsk, Russia, 432027

Abstract. This article describes the method of application of an anomaly template of time series of project metrics based on entropy. The analysis of project activity metrics is described. The forecasting algorithm based on the fuzzy trends of time series of indicators was developed and implemented. The formula for the entropy measure for the fuzzy time series is determined. The algorithm uses the dependence of the forecast on the measures of entropy. The hypothesis of trend stability is used for forecasting. Experiments based on this approach are presented.

Keywords: fuzzy time series, fuzzy trend, entropy, hypotheses, forecasting.

1. Introduction

At present, one of the factors of cost-effective activity of design organizations is the constant analysis of many projects of a large design organization throughout the life cycle. The task of measuring the characteristics of the project activity should be considered as dependent on the creation of a means for project management [1]. This tool automates the clustering processes by the similarity of all available enterprise project events for the subsequent forecast of values. The article is devoted to solving the problem of project monitoring. The solution consists in applying tools to analyze the state of the software project metrics using entropy measures. Project metrics are unloaded from the version control system.

2. Analyzing Project Metrics

The model of analysis and management of a set of projects in the process of project activity is developed.

$\{C_t, R_t, B_t, I_t, F_t, R^{BI}, R^{IF}\}$,

whereC_t – time series commits,

R_t – time series release,

B_t – time series bugs,

I_t – time series improvement,

F_t – time series NewFeature,

R^{BI} – dependence of the number of bugs on improvements,

R^{IF} – the dependence of new functional properties on the number of improvements (New Features from improvements).

Discrete series represent the project data.

The algorithm for constructing a time series model for solving this problem consists of the following stages:

- The discretetime series $Y = \{t_i, x_i\},\ i \in [1, n]$, where t_i – some point in time, x_i – level time series, transformed into an fuzzy time series $\widetilde{Y} = \{t_i, \tilde{x}_i\},\ i \in [1, n]$, $X = \{x_i\}, \tilde{x}_i \in \widetilde{X}$, where $\tilde{x}_i$ – fuzzy label [2].
- The fuzzy time series is transformed into a time series of fuzzy elementary trends. It is indicated by $\tau_i = \mathrm{ETend}(\tilde{x}_i, \tilde{x}_{i+1})$, where *ETend* – functional, which is implemented on the basis of operations: *Ttend* – determination of the type of trend ($\tilde{v}_t$) and *Rtend* – determination of the intensity of the trend ($\tilde{a}_t$) on a special linguistic scale constructed for the initial time series [3].
- Structural model of fuzzy trend $\tau \in \Im$an fuzzy time series [3,4]: $\tau = \langle \tilde{v}, \tilde{a}, \Delta t, \mu \rangle$, where τ – name of a fuzzy tendency from the set $\Im$, $\tau \in \Im$; $\tilde{v}$ – type of fuzzy trend, $\tilde{v} \in \widetilde{V}$; $\tilde{a}$ – intensity of fuzzy trend, $\tilde{a} \in \widetilde{A}$; Δt – duration of fuzzy trend, $\Delta t \in \Delta T$; μ – accessory function of a fuzzy time series, bounded by an interval Δt, of fuzzy trend τ.[4]
- Time series of fuzzy elementary trend: $\tilde{v}_t = \mathrm{TTend}(\tilde{x}_t, \tilde{x}_{t+1})$, $\tilde{a}_t = \mathrm{RTend}(\tilde{x}_t, \tilde{x}_{t+1})$, $\mu_t = \min\big(\mu(\tilde{x}_t), \mu(\tilde{x}_{t+1})\big)$. [4]
- The fuzzy elementary trend modeling method was used to predict the numerical values and fuzzy tendencies of the state of the organization's project [5]. The forecast uses hypothesis testing:
  - *Hypothesis 1. The hypothesis of conservation / change of trend:* $\tau_{t+1} = \tau_t + \tau_p$, where τ_{t+1} – forecast for the next period of time; τ_t – real value at time t; τ_p – the value of the trend over the previous period of time.
  - *Hypothesis 2. The hypothesis of stability / instability of the trend:* $\tau_{t+1} = \tau_t + G\tau_p$, where $G\tau_p$ – mportance of a dominant fuzzy trend [6].
  - *Hypothesis 3. Forecasting for a given period based on a fuzzy elementary trend* [7].

An entropy time series is additionally introduced to select the best hypothesis [8].

3. Algorithm for calculating anomalies

Two pairs of parameters:

- The first pair is a fuzzy label and a fuzzy trend.
- The second pair is the measure of entropy by function and the measure of entropy by the fuzzy trend.

There are many methods, so there will not be an emphasis on this to determine the first pair. In addition, there may be cases when an input is supplied to the HBP.

The algorithms of the expected state of the second series are given below.

The measure of entropy in terms of functions is defined in 2 steps [8]:

- The value of entropy is calculated by the membership function according to formula:

$H_i^{\mu} = \mu(x_i)\ln\big(\mu(x_i)\big)$,

where $\mu(x_i)$ – value of the membership function of the point$x_i$to the fuzzy interval.

- The linguistic interpretation of the measure of entropy is determined on the basis of the value obtained. The value of the measure of entropy close to 0 corresponds to the state "Authentically". The value of the measure of entropy close to the maximum corresponds to the state "Uncertain." In other cases, the value of the entropy measure corresponds to the state "Probably".

if$\big(H_i^{\mu} \to 0\big)$ then $\widetilde{H_i^{\mu}} = \mathrm{Authentically}$,

else if $\big(H_i^{\mu} \to \max\big)$ then $\widetilde{H_i^{\mu}} = \mathrm{Uncertain}$,

else $\widetilde{H_i^{\mu}} = \mathrm{Probably}$.

The measure of entropy is obtained on the basis of the membership function. Then it is not able to clearly record the change of fuzzy timestamp marks. This measure of entropy only shows how likely the point will be to the label. In this case, if the entropy is close to the maximum value, then this

indicates that the point is in the "boundary" position and can relate with equal probability to two different fuzzy marks.

Measure of entropy by a fuzzy trend.

- Dynamics of the trend at the previous point is determined on the basis of the formula:

$$\Delta\tau_{i-1}^{fact} = \tau_{i-2}^{fact} - \tau_{i-1}^{fact}.$$

- The position of the fuzzy trend in the phase plane is calculated on the basis of the weight of this fuzzy trend and the value of the dynamics of the fuzzy trend at the previous point by the formula:

  $p_{i-1} = CalcCodePoint(\tau_{i-1}^{fact}, \Delta\tau_{i-1}^{fact})$.
- Three sets of points of the phase plane are determined:
  1. The most probable are points (usually one point), which are most often followed after the point p_{i-1}:

     $\omega_{mostexpect} = Max(Probability(p_{i-1}))$.
  2. Probable points are points to which they also follow after point p_{i-1}, but they are not included in the first set:

     $\omega_{probability} = Probability(p_{i-1}) \notin \omega_{mostexpect}$
  3. Anomalous points are all points not included in the first two sets (transition to them is not expected in normal operating conditions):

$$\omega_{anomaly} = AllPoint \notin Probability(p_{i-1}).$$

- The dynamics of the trend at the current point is determined by the formula:

$\Delta\tau_i^{fact} = \tau_{i-1}^{fact} - \tau_i^{fact}$.

- The point of the phase plane for the fuzzy trend and dynamics at the current point of the series is calculated according to formula:

$p_i = CalcCodePoint(\tau_i^{fact}, \Delta\tau_i^{fact})$.

- The resulting point is determined to which of the three sets:$\omega_{mostexpect}$, $\omega_{probability}$ or$\omega_{anomaly}$ it refers to the formulap_i.
  1. If the point belongs to the set $\omega_{mostexpect}$, then the value of the entropy measure is set to 0. Since the received point was just expected, then we did not learn anything new.

     if$(p_i \in \omega_{mostexpect})$ then $H_i^\tau = 0$.
  2. If the point refers to the set of $\omega_{probability}$, then the value of the entropy measure is set to 0.5. Since the resulting point, although not expected, but also was not something completely new.

     else if $(p_i \in \omega_{probability})$ then $H_i^\tau = 0.5$.
  3. If the point belongs to the set $\omega_{anomaly}$, then the value of the entropy measure is set to 1. Since the resulting point was not expected to be seen, then at the moment the system being analyzed is in an unknown state:

     else if $(p_i \in \omega_{anomaly})$ then $H_i^\tau = 1$.
- The linguistic interpretation of the obtained numerical value of the entropy measure according to the fuzzy trend $H_i^\tau$is determined according to the formula:

if$(H_i^\tau = 0)$ then $\widetilde{H_i^\tau} = Stability$,
else if$(H_i^\tau = 0.5)$ then $\widetilde{H_i^\tau} = Change$,
else $\widetilde{H_i^\tau} = Anomaly$.

The algorithm for finding anomalies in the time series begins with the expert entering patterns of anomalies. Anomaly pattern is a sequence of numbers of situations that precede anomalies. The last number in the sequence is the number of the abnormal situation. In the work, the algorithm uses one pair of parameters: either a fuzzy label - a fuzzy trend, or measures of entropy by membership function and by a fuzzy trend. Regardless of the choice of a pair of parameters, the anomaly search algorithm will be identical, except for the first step.

The algorithm for finding known anomalies from given patterns consists of steps:

Step 1. For a new point, determine the value of a pair of parameters (say, fuzzy label f and trend t): $f_i \in F, t_i \in T$.

Step 2. Find out the situation number. The values of the pair of parameters for the previous point and the current one are known:

$S_i = (f_{i-1}, t_{i-1}) \rightarrow (f_i, t_i)$.

Step 3. For each of the selected patterns of anomalies, obtain a number of the following expected situations. If the next situation coincides with the current situation, then check how many more situations remain in the template. If the pattern is already complete, the next point will lead to an anomaly. If the following situation of the template does not coincide with the situation that appeared, then exclude the template from the selected templates of the anomalies:

$$\text{if}\left(\text{Template}(S_y)[j]! = S_i\right) \text{ then deleteFromSelect}\left(\text{Template}(S_y)\right), \text{else if } \left(\text{Template}(S_y).\text{Count} - 1 == j\right)\text{then nextAnomaly}, \text{else } j = j + 1,$$

$y \in Y, j \in [1, \text{Template}(S_y).\text{Count}]$,

where Y – number of selected anomaly patterns, $\text{Template}(S_y)[j]$ – expected situation for the template.

Step 4. Check the first situation for all patterns of anomalies. If it coincides with the situation of S_i, then the template is included in the selected templates:

$\text{if}(\text{Template}(S_x)[1] == S_i) \text{ then addFromSelect}(\text{Template}(S_x))$,

where$x \in X$,X – the number of all anomaly patterns for the series.

Templates of anomalies can not be specified in some cases when analyzing time series. For example, if the system under investigation is new and its behavior is a black box. In such cases, time series analysis is possible. The coming situations will be revealed less and less with such an analysis. An expert can select situations after the analysis is completed. These situations will be anomalous. Also, these situations may not be abnormal, but simply seldom occurring. The exclusion of a template is possible from the list of anomalies in the course of the analysis if the frequency of its meetings changes and becomes higher than the threshold value.

The algorithm for detecting new anomalies in time series:

Step 1. The value of the pair of parameters (say, the value of the fuzzy label f and the trend t) is determined for the new point according to the formula:

$f_i \in F, t_i \in T$.

Step 2. The situation number is determined for the previous point and the current one, knowing the values of the pair of parameters according to the formula

$S_i = (f_{i-1}, t_{i-1}) \rightarrow (f_i, t_i)$.

Step 3. The probability of this situation is determined by the formula:

$$P(S_i) = \frac{\text{count}(S_i)}{\text{countPointOfSeries}} * 100\%.$$

Step 4: If the probability is less than 0.01, then this may be an anomaly of the series. Otherwise, if the template is already present in the list of abnormal, then exclude it from the list:

$\text{if}(P(S_i) < 0.01) \text{then } S_i \text{ is anomaly},$

$\text{else if}(S_i \in \text{Templates}) \text{then removeFromTemplates}(S_i)$.

Step 5. *N* recent situations are determined for anomalies. These situations precede the onset of an anomaly. The anomaly pattern is obtained according to formula:

$\text{Template}(S_i) = \langle S_n, S_{n-1}, \dots, S_{i-1} \rightarrow S_i \rangle$.

4. Experiments

The metrics of the events of the open project "FreeNAS9" were taken for research. The time series of "closed" "Bug" and "Feature" were taken [9]. The results of the similarity of these metrics for determining the dependencies are presented in Table 1. The results of the analysis and forecasting of project activity metrics are presented in Tables 1 and 2. Forecasting the appearance of errors in the FreeNAS9 project, taking into account the effect of adding new functionality to the project (the hypothesis of trend stability) is shown in Figure 1.

Table 1. Analysis of project data.

Label type	The general trend	The dominant trend	A measure of similarity	Correlation	Interpretation of correlation
Bug	Growth	Stability	0.75	0.9387	Strong
New Feature	Growth	Stability			

Table 2. The results of forecasting taking into account the influence of the predictor time series.

Dependent time series	Time series predictor	Hypothesis 1	Hypothesis 2	Hypothesis 3
Bug	New Feature	Falling average	Growth strong	Falling average

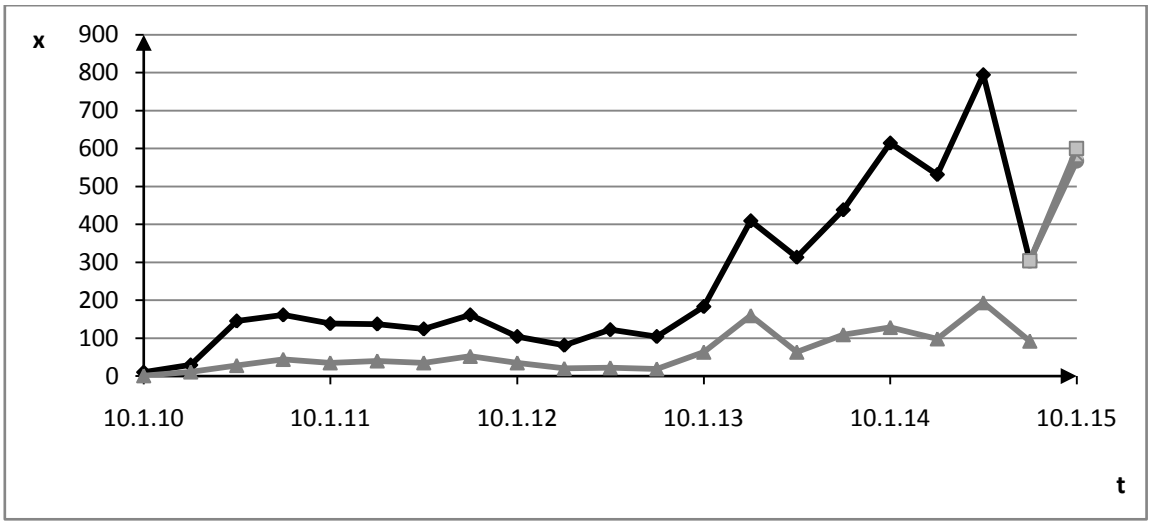

Figure 1. Forecasting in the project "FreeNAS9".

The entropy time series according to the metrics of the FreeNAS9 project is presented in Table 3.

Table 3. BUG-FreeNAS9 (closed).

Point number	Measure of entropy by the membership function	Measure of entropy by fuzzy trend
1	Reliably	Stability
2	Reliably	Stability
3	Reliably	Stability
4	Reliably	Stability
5	Reliably	Stability
6	Reliably	Stability
7	Reliably	Stability
8	Reliably	Stability
9	Probably	Stability
10	Reliably	Stability
11	Probably	Change
12	Reliably	Stability
13	Probably	Stability
14	Probably	Change
15	Reliably	Change

5. Conclusion

The following conclusion can be made on the basis of the obtained data: applying the hypothesis of retaining the trend will give an incorrect result, since the measure of entropy by the fuzzy trend at the last point is in the "Change" state. This indicates a change in the trend in the time series. The absence of repeated situations during shifts also indicates that the use of the hypothesis for a given period will give an incorrect result. Therefore, it is worth choosing the hypothesis of stability of the trend as the most probable in such a situation.

6. Acknowledgments

The article was supported by the Russian Foundation for Basic Research (grant No. 16-47-732070).

7. References

[1] Moshkin, V.S. Intelligent data analysis and ontological approach in project management / V.S. Moshkin, A.N. Pirogov, I.A. Timina, V.V. Shishkin, N.G. Yarushkina // Automation of management processes. – 2016. – Vol. 4(46). – P. 84-92. (in Russian).

[2] Herbst, G. Online Recognition of fuzzy time series patterns / G. Herbst, S.F. Bocklish // International Fuzzy Systems Association World Congress and European Society for Fuzzy, 2009.

[3] Kacprzyk, J. Using Fuzzy Linguistic summaries for the comparison of time series / J. Kacprzyk, A. Wilbik // International Fuzzy Systems Association World Congress and European Society for Fuzzy Logic, 2009.

[4] Pedrycz, W. Time Series Analysis, Modeling and Applications: A Computational Intelligence Perspective (e-book Google) / W. Pedrycz, S.M. Chen // Intelligent Systems Reference Library. – 2013. – Vol. 47. – P. 404.

[5] Krol, T.Ya. Methods for solving the problem of clustering and forecasting in an electronic archive / T.Ya. Krol, M.A. Kharin // Young Scientist. – 2011. – Vol. 1(6). – P. 135-137. (in Russian).

[6] Yarushkina, N.G. Extraction of knowledge about the dependencies of time series for forecasting problems / N.G. Yarushkina, T.V. Afanasyeva, A.A. Romanov, I.A. Timina // Radiotekhnika. – 2014. –№ 7. – P.141-146. (in Russian).

[7] Yarushkina, N.G. Application of the entropy measure in the diagnosis of technical time series / N.G. Yarushkina, V.V. Voronina, E.N. Egov// Automation of control processes. – 2015. – № 2. – P. 55-63. (in Russian).

[8] Egov, E.N. Fuzzy modeling and genetic optimization of time series in the intellectual system of technical diagnostics / E.N. Egov, N.G. Yarushkina, D.V. Yashin // Radiotekhnika. – 2016. – №. 9. – P. 64-71. (in Russian).

[9] FreeNAS 9Web Site [Electronic resource]. – Access mode: https://bugs.pcbsd.org/projects/freenas (30.09.2017).