Fuzzy models in forecasting time series of project activity metrics

I A Timina¹, E N Egov¹ and A A Romanov¹

¹Ulyanovsk State Technical University, Severny Venec str. 32, Ulyanovsk, Russia, 432027

e-mail: i.timina@ulstu.ru, e.egov@ulstu.ru, romanov73@gmail.com

Abstract. The aim of this study is to show application of entropy metric to time series of software projects prediction based on anomaly patterns. The analysis of project activity metrics is described. The proposed forecasting algorithm based on the fuzzy trends of indicators time series. Experiments that shows the analysis of a time series with software project indicators is demonstrated.

1. Introduction
One of the factors of cost-effective activity of design organizations is the regular analysis of large sets of software projects throughout the life cycle. The task of measuring the characteristics of the project activity should be considered as dependent on the creation of a tools 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 problem is that today projects may be developed by modern methodologies. In this case structure of life cycle and project codebase may be heterogeneous or not deterministic. We propose to use entropy measures of a time series. Time series for the analysis of the state of the software project metrics is used.

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. We extract software project metrics from the version control system, such as git, svn and others.

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

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

where \(C_t\) – commits time series, \(R_t\) – release time series, \(B_t\) – bugs time series, \(I_t\) – improvement time series, \(F_t\) – new feature time series, \(R^{BI}\) – 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).

We define these types of time series for detailed structurisation of software production processes and their management. We propose use our techniques for extracting dependencies between represented types of time series.

We use discrete time series to represent the project data. The algorithm for constructing a time series model for solving this problem consists of the following stages:
Two pairs of parameters:

- The discrete time series \( Y = \{ ts_i \}, i \in [1, n] \), where \( ts_i = [ t_i, x_i ] \) is a time series element in time moment \( t_i \) with level \( x_i \), transformed into an fuzzy time series \( \tilde{Y} = \{ \tilde{ts}_i \}, i \in [1, n] \), \( \tilde{ts}_i = [ t_i, \tilde{x}_i ], \tilde{x}_i \in \tilde{X} \), where \( \tilde{x}_i \) – fuzzy label [2]. Fuzzy label in most cases has values: {dangerously low, low, normal, high, dangerously high}.

- The fuzzy elementary trend modeling method was used to predict the numerical values and structural model of fuzzy trend an fuzzy time series [3] [4]:
  \[
  \tilde{x}_i \in \tilde{a}, \tilde{v}, \tilde{\tau} \subset \tilde{A}, \tilde{V}, \tilde{I}, \tilde{T}, \tilde{\Delta} \]

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 FTS. 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:
  \[
  H^\mu_i = \mu(x_i)\ln(\mu(x_i)),
  \]
  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".

\[ \hat{H}^\mu_i = \begin{cases} \text{Authentically}, & H^\mu_i \to 0, \\ \text{Uncertain}, & H^\mu_i \to \max, \\ \text{Probably}, & \text{otherwise} \end{cases} \]

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:

\[ \Delta \tau_{i-1}^{\text{act}} = \tau_{i-2}^{\text{act}} - \tau_{i-1}^{\text{act}} \]

- 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:

\[ p_{i-1} = \text{CalcCodePoint}(\tau_{i-1}^{\text{act}}, \Delta \tau_{i-1}^{\text{act}}). \]

- Three sets of points of the phase plane are determined:

(i) The most probable are points (usually one point), which are most often followed after the point \( p_{i-1} \):

\[ \omega_{\text{mostexpect}} = \text{Max}(\text{Probability}(p_{i-1})). \]

(ii) Probable points are points to which they also follow after point \( p_{i-1} \), but they are not included in the first set:

\[ \omega_{\text{probability}} = \text{Probability}(p_{i-1}) \notin \omega_{\text{mostexpect}}. \]

(iii) Anomalous points are all points not included in the first two sets (transition to them is not expected in normal operating conditions):

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

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

\[ \Delta \tau_i^{\text{act}} = \tau_i^{\text{act}} - \tau_i^{\text{act}}. \]

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

\[ p_i = \text{CalcCodePoint}(\tau_i^{\text{act}}, \Delta \tau_i^{\text{act}}). \]

- The resulting point is determined to which of the three sets: \( \omega_{\text{mostexpect}} \), \( \omega_{\text{probability}} \) or \( \omega_{\text{anomaly}} \) it refers to the formula \( p_i \).

(i) If the point belongs to the set \( \omega_{\text{mostexpect}} \), 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.

\[ H^T_i = 0, \quad p_i \in \omega_{\text{mostexpect}}. \]
(ii) If the point refers to the set of anomaly, 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.

\[ H^{\tau}_{i} = 0.5, \quad p_{i} \in \omega_{\text{probability}}. \]

(iii) If the point belongs to the set 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:

\[ H^{\tau}_{i} = 1, \quad p_{i} \in \omega_{\text{anomaly}}. \]

- The linguistic interpretation of the obtained numerical value of the entropy measure according to the fuzzy trend is determined according to the formula:

\[ \tilde{H}^{\tau}_{i} = \begin{cases} 
\text{Stability}, & H^{\tau}_{i} = 0, \\
\text{Change}, & H^{\tau}_{i} = 0.5, \\
\text{Anomaly}, & H^{\tau}_{i} = 1 
\end{cases} \]

The algorithm for detecting 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:

\[ \text{if}(\text{Template}(S_{y})[j]! = S_{i}) \text{ then } \text{deleteFromSelect}(\text{Template}(S_{y})), \]

\[ \text{if}(\text{Template}(S_{y}).\text{count} - 1 == j) \text{ then } \text{nextAnomaly}, \]

\[ \text{else } j = j + 1, \]

\[ y \in Y, j \in [1, \text{Template}(S_{y}).\text{count}], \]

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:

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

where $x \in X$ 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:

$$ P(S_i) = \frac{\text{count}(S_i)}{\text{countPointsOfSeries}} \times 100\% $$

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:

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

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

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:

$$ \text{Template}(S_i) =< S_n, S_{n-1}, ..., S_{i-1} \rightarrow S_i >. $$

4. Fuzzy sets of type 2

In most cases for fuzzy systems implementation, fuzzy sets of the first order are used. In 1975, Lotfi Zadeh presented fuzzy sets of the second order (type 2) and fuzzy sets of higher orders, to eliminate the disadvantages of fuzzy sets of type 1. These disadvantages can be attributed to the problem that membership functions are mapped to exact real numbers. This is not a serious problem for many applications, but in cases where it is known that these systems are uncertain.

The solution to the above problem can be the use of fuzzy sets of type 2, in which the boundaries of the membership areas themselves are fuzzy, which in themselves are fuzzy [9].

It can be concluded that this function represents a fuzzy set of type 2, which is three-dimensional, and the third dimension itself adds a new degree of freedom to handle uncertainties. In [9] Mendel defines and differentiates two types of uncertainties, random and linguistic. The first type is characteristic, for example, for the processing of statistical signals, and the
characteristic of linguistic uncertainties is contained in systems with inaccuracies based on data determined, for example, through expert statements.

To illustrate, we note the main differences between fuzzy sets of type 1 and fuzzy sets of type 2. Let us turn to 1, which illustrates a simple triangular membership function. Figure 1 (a) shows a clear assignment of the degree of membership. In this case, to any value of $x$ there corresponds only one point value of the membership function. If you use a fuzzy membership function of the second type, you can graphically generate its designation as an area called the footprints of uncertainty (FOU). In contrast to the use of the membership function with clear boundaries, the values of the membership function of type 2 are themselves fuzzy functions.

In our approach we use fuzzy sets type 2 to solve next problem. Indicators of processes, also in software development, may contains different levels. But common rules for time series can be described in fuzzy labels identically. We use fuzzy set type 1 for modeling time series, and fuzzy sets of type 2 to reflect the uncertainty of the expert in describing fuzzy labels of different time series with the same nature.

This proposition affects on first step of algorithm of time series modeling (see section 2). Let’s describe fuzzification step.

- Define fuzzy labels for time series. These fuzzy labels are common for different time series with the same nature. Fuzzy labels are calculated with sets of type 1.
- Define fuzzy set of type 2 for each fuzzy label from previous step.
- Calculate fuzzy time series using max between membership functions.

5. Experiment

The metrics of the events of the open project "FreeNAS9" were taken for research. The time series of "closed", "Bug" and "Feature" were taken [10]. The results of the similarity of these metrics for determining the dependencies are presented in Table 1.

| 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 1. Analysis of project data.
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 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 2. Forecasting in the project "FreeNAS9"]/](image)

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

Table 3. Analysis of project data.

| 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                           |
6. 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.

7. References
[1] Moshkin V S, Pirogov A N, Timina I A, Shishkin V V and Yarushkina N G 2016 Intelligent data analysis and ontological approach in project management Automation of management processes 4(46) 84-92 (in Russian)
[2] Herbst G and Bocklish S F 2009 Online Recognition of fuzzy time series patterns International Fuzzy Systems Association World Congress and European Society for Fuzzy Logic
[3] Kacprzyk J and Wilbik A 2009 Using Fuzzy Linguistic summaries for the comparison of time series International Fuzzy Systems Association World Congress and European Society for Fuzzy Logic
[4] Pedrycz W and Chen S M 2013 Time Series Analysis, Modeling and Applications: A Computational Intelligence Perspective Intelligent Systems Reference Library 47 404
[5] Krol T Ya and Kharin M A 2011 Methods for solving the problem of clustering and forecasting in an electronic archive Young Scientist 1(6) 135137 (in Russian)
[6] Yarushkina N G, Afanasyeva T V, Romanov A A and Timina I A 2014 Extraction of knowledge about the dependencies of time series for forecasting problems Radiotekhnika 7 141-146 (in Russian)
[7] Yarushkina N G, Voronina V V and Egov E N 2015 Application of the entropy measure in the diagnosis of technical time series Automation of Control Processes 2 55-63 (in Russian)
[8] Egov E N, Yarushkina N G and Yashin D V 2016 Fuzzy modeling and genetic optimization of time series in the intellectual system of technical diagnostics Radiotekhnika 9 64-71 (in Russian)
[9] Mendel J M and John R I B 2002 Type-2 Fuzzy Sets Made Simple IEEE Transactions on Fuzzy Systems 10(2) 117-127
[10] Free NAS 9 Web Site (Access mode: https://bugs.pcbsd.org/projects/freenas) (30.09.2017)

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