Two-stage incidents prediction in heat supply systems using gradient boosting

A A Akhvaev¹, V F Shurshev¹ and N A Nosikovsky¹
¹Astrakhan State Technical University, Tatisheva-16, Astrakhan, Russian Federation

E-mail: astu.arm@yandex.ru

Abstract. Information technologies usage in heat supply sphere has an important role to solve accidents problems in heat networks. Heat networks are accidents (incidents) sources. Accident situation predicting in heat supply systems will make it possible to rationally finance incidents prevention funds. AI is the tool for predicting accidents. Modern AI methods and algorithms allow to solve many technical systems problems. Research area has a data bank that allows to forecast heat networks accidents. Gradient boosting and production rules were chosen as basic methods.

1. Introduction
Heat supply systems state is a determining factor in the social and economic sufficiency of any state, region and town. The acceptable state of heat supply systems is determined by this infrastructure accident problem solution. Due to their nature, heat supply systems are subject to components wear and tear, and therefore, incidents problem remains relevant yet.

Accident rate problem solution in heat supply systems, specifically in heat networks, is prediction using artificial intelligence tools.

2. Heating networks accidents prediction model description
Nowadays artificial intelligence is one of the major science trends. One of the main artificial intelligence systems trends is expert system. Expert system is similar to the expert decisions in specific area. Unlike other artificial intelligence systems (neural networks, genetic algorithms), they are applicable for solving problems in a narrowly focused specific area.

Expert systems usage is complicated by transferring expert knowledge problem. In addition, correct expert group formation plays an important role in system adequacy determining. This means that knowledge from expert sufficiency and reliability are directly determined by expert system result reliability [7-8]. Another solution may be another artificial intelligence tool usage, specifically machine learning. In classification and recovery context, those regressions are suitable:
- linear regression;
- logistic regression;
- random forest;
- gradient boosting.

Accident prediction problem solution is considered in classifier problem solution context; so algorithm work result will be two classes: «accident» and «non-accident». However, for a full-fledged
prediction, we still need to know how correct this prediction is. To do this, you need to consider the following metrics: accuracy, completeness, and f-measure. In this way, we can get our prediction estimate numerical value. Among the proposed algorithms, gradient boosting can provide more accurate predictions, but it cannot provide information about parameters (input data) which affected the prediction. This problem can be solved by implementing another model that should identify incidents based on the production rules. Production rules are formed on expert opinion base and extracting attributes importance from a machine learning algorithm. The first three of the algorithms above (linear regression, logistic regression, random forest) can give an attributes importance idea.

Therefore, the system consists of two models: prediction model (1) and prediction analysis model (2).

\[ M = \{MP, OV, SL, VI\} \]  \hspace{1cm} (1)

where MP – input parameters set; SL – Basic algorithm used in gradient boosting (random forest); OV-training sample; VI-output information.

\[ CM = \{MP, BP, MK, VI\} \]  \hspace{1cm} (2)

where MP – input parameters set; BP – Production rules base; MK – coefficients set; VI-output information (PA – failure causes, PM – countermeasures).

3. Model input parameters analysis
It is necessary to build pipe failure predict chance model based on available attributes. The attributes list is presented in Figure 1.

| Number | Attribute Description                                      |
|--------|------------------------------------------------------------|
| 0      | Sys                                                        |
| 1      | source number                                              |
| 2      | name of the beginning of the section                        |
| 3      | name of the end of the section                              |
| 4      | balance sheet holder                                        |
| 5      | plot length                                                |
| 6      | Internal diameter of the supply pipeline, m                 |
| 7      | Internal diameter of the return pipe, m                     |
| 8      | Nominal diameter, PT, mm                                    |
| 9      | Nominal diameter, OT, mm                                    |
| 10     | Roughness of the supply pipeline, mm                        |
| 11     | Roughness of the return pipeline, mm                        |
| 12     | Year of commissioning                                       |
| 13     | Type of heat network laying                                 |
| 14     | Local resistance coefficient of the supply pipeline         |
| 15     | Coefficient of local resistance of the return pipeline      |
| 16     | Overgrowth of the supply pipeline, mm                       |
| 17     | Overgrowth of the return pipework, mm                       |
| 18     | Water flow rate in the supply pipeline, t/h                 |
| 19     | Water flow rate in the return pipeline, t/h                 |
| 20     | Pressure loss in the supply pipeline, m                     |
| 21     | Pressure loss in the return pipe, m                         |
| 22     | Specific linear pressure loss in the supply pipeline, mm/m  |
| 23     | Specific linear pressure loss in the return pipeline mm/m   |
| 24     | Accident rate                                               |

**Figure 1. Attributes list**

It is necessary to get rid of unnecessary attributes before data analysis. Removing all string type attributes, leaving and coding only "Type of heating network laying", we get following attributes list.
Figure 2.

Figure 3. Numeric attributes

Table 1 shows a matrix containing correlation coefficients between all attributes. The correlation matrix shows that some attributes have a weak effect on the failure rate. They will not be included in the final attributes list.

While researching accident rates based on the “Year of commissioning” attribute, large number of accidents were detected in heat network sections installed in the 1980s. At first, it is logical to assume that the pipes are subject to accidents, because they are out of service. However, pipes installed in
1968 have a much lower accident rate. Moreover, number of pipes installed in 1968 and 1978 in sample is approximately the same. Such anomalies with a clear difference in values require removing elements from the selection.

In correlation matrix, attribute "Roughness of the supply pipeline, mm" is clearly highlighted. In Figure 4, it can be seen that almost all pipes with values from 0.5 to 2 mm had an accident.

![Figure 4. Violin plot «Roughness of the supply pipeline, mm»](image)

In Figure 5 a significant spike can be seen of values 0.5 and 1.5 mm.

![Figure 5. Summary table «Roughness of the supply pipeline, mm»](image)

During the research of each input parameter in our model, the following list of parameters (Figure 6) was determined. These attributes are taken into account when forming the training sample and training the algorithm.
4. Analysis of gradient boosting parameters

The first model version uses gradient boosting. Machine learning algorithms require not only correct training sample formation, but also algorithm internal parameters values selection. In gradient boosting, following internal parameters are considered:

- max_features - count of attributes, which are taken into account when searching for best split;
- max_depth - max. trees depth;
- learning_rate - reduces the impact of an individual tree;
- n_estimators - trees count [1-6].

Based on the training sample formed by us with the previously selected attributes, gradient boosting internal parameters analysis was carried out (Figure 7).

The graphs show prediction estimate depending on gradient boosting internal parameters values. Red line indicates dependence on test sample data, and blue line indicates prediction score. Analysis
showed that increasing random forest trees number, which is used as algorithm base, to over 200, leads to certain linear correlation. As for trees depth, if trees number is 200, optimal depth value is 9.

Thus, following parameter values were selected for our training sample and prediction score is high.

```
gbrt = GradientBoostingClassifier(random_state=0, learning_rate=1, max_depth=9, max_features=30, n_estimators=200)
gbrt.fit(X, y)
print(gbrt.score(X_train, y_train))
print(gbrt.score(X_test, y_test))
0.9927777777777778
0.9926666666666667
```

**Figure 8. Prediction score**

5. Conclusion

During the research, heat networks parameters were studied to form a further machine learning algorithm sample. Analysis results are useful not only for training sample, but also for production rules base further formation [10-13].

Prediction using machine learning requires not only data analysis for the training sample, but also a competent algorithm parameters selection itself. Following parameters were investigated for gradient boosting: max_features, max_depth, learning_rate, n_estimators. Prediction score equals to 0.9, which is considered a high prediction score. Next researching task is to make sure that this is not over-training, but really an adequate prognosis assessment.

References

[1] Radchenko S A, Sergeev A N and Radchenko S S 2016 Heat supply and heating systems accidents: causes, damage and opportunities to reduce it *Monograph*
[2] Shurshev V F and Abzalov A V 2008 Identification system for pre-emergency situations in an ammonia refrigeration unit Caspian Journal: Management and High technologies 1 56-59
[3] Abzalov A V and Zhedunov R R 2013 Methodology for analyzing pre-emergency situations at technological control facilities Caspian Journal: Management and High technologies 4 50-59
[4] Shurshev V F, Kochkin G A and Kochkina V R 2013 Decision support system model on base precedent reasoning *Bulletin of the Astrakhan State Technical University. Series: Management, Computer Engineering and Computer Science* 2 175-183
[5] Shurshev V F and Buj L V 2015 Scanning receivers and transceivers rational selection model and algorithm *Bulletin of the Saratov State Technical University* 3(1) 166-175
[6] Kvyatkovskaya I Yu, Chertina E V and Belov S V 2017 Using additional professional education system in the IT specialists advanced training process: Astrakhan State Technical University experience *Collection of Information and communication technologies in science, production and education ICIT-2017 Collection of articles of the International Scientific and Practical Conference* 411-419
[7] Shurshev V F and Umerov A N 2005 Decision-making process modeling in refrigerating agents mixtures flow modes identification *Bulletin of the Kuzbass State Technical University* 5 27-29 (in Russian).
[8] Axvaev A A and Shurshev V F 2018 In the ground biological objects recognition with using radiophysical properties *Collection of articles of the II International Scientific and Practical Conference* 69-71
[9] Byaleczkaya E M, Kvyatkovskaya I Yu and Shurshev V F 2011 Indicators set formation for assessing residential buildings management quality *Bulletin of the Astrakhan State Technical University. Series: Management, Computer Engineering and Computer Science* 2 143-149
[10] Shurshev V F and Byaleczkaya E M 2010 Residential object control and measurement system indicators expert evaluation algorithm Bulletin of the Astrakhan State Technical University. Series: Management, Computer Engineering and Computer Science 2 117-121
[11] Shurshev V F and Demich N V 2006 Complex evolutionary method for finding solutions to the problem of refrigerating agents mixtures synthesis algorithm Yuzhno-rossijskij vestnik geologii, geografii i global`noj e`nergi 11 65-68
[12] Chertina E V and Kvyatkovskaya I Yu 2016 Information system for innovative IT projects evaluation, analysis, selection and monitoring Fundamental`ny`e issledovaniya 5-3 526-530
[13] Kvyatkovskaya I Yu and Shurshev V F 2019 Information systems for quality management in automated and automated production facilities