A study of decision tree application in the problem of accounting for non-insured periods of a pensioner

L V Naykhanova¹⁴, A A Buldaev², N N Ausheeva³, I V Naykhanova², N B Haptahaeva³

¹ Department "Computer Science Systems", East Siberia State University of Technology and Management, 40V, Klyuchevskaya St., Ulan-Ude, 670013, Russia
² Computer Center, East Siberia State University of Technology and Management, 40V, Klyuchevskaya St., Ulan-Ude, 670013, Russia
³ Department "Automation and Electrical Equipment of Industrial Enterprises", East Siberia State University of Technology and Management, 40V, Klyuchevskaya St., Ulan-Ude, 670013, Russia
⁴ E-mail: obeka_nlv@mail.ru

Abstract. The article is devoted to the study of machine learning algorithm "decision Tree" for solving the problem "Accounting for non-insured periods". In pension legislation, there are problems that required a large amount of combinatorial calculations. These include the problem of "Accounting for non-insured periods". To solve similar problems, machine learning methods should be suitable. To implement the algorithm, training and test samples are created by means of a random number sensor. However, studies have shown that the use of this method gives satisfactory results, but not acceptable for this task. The authors conclude that it is necessary to investigate other, more complex machine learning algorithms, or to find a new dependence of the output variable on the input variables. This dependence is probably to be nonlinear that will determine a choice of a new method of machine learning.

1. Introduction

Production systems are well suited for automatization of the basic processes of pension assignment. The paper [1] describes the concept of such system. However, in the actual systems it is necessary to create hybrid systems. Such systems are based on the use of different theories. In the pension legislation, there are tasks for which using the production rules is inefficient, such as the task "Recalculation of pensions taking into account non-insured periods".

At the beginning of 2015 the Law #400-FZ "On insurance pensions" came into force [2]. It changes the procedure for calculating the insurance part of the pension and adds new conditions to the existing ones. Now, the size of the pension is affected by additional parameters. One of them is now a number of children of a female pensioner. The point is that periods of taking care for children should be taken into the account. Recalculation of the pension will allow these women to receive an increase in their pensions.

Recently, the number of large families in Russia has increased. For example, in St. Petersburg the increase in 2008 compared to 2007 amounted to less than a thousand families, but since 2015 experts noted an annual increasing to 5,000 families in this category [3]. This fact shows that this problem is relevant not only for the recalculation of pensions, but also for an appointment of new ones.
To solve this problem, some cases cause difficulties. For example, for a mother with many children (more than 3 children), periods of "Child Care" would interfere with periods of "Work". There is no perfect formula for calculation of the increase. In this case, it all depends on individual parameters of a pensioner (experience, salary, number of children, etc. ...). If high wages have been taken into an account during the period of Child Care, this period should not be considered. Large computing costs require cases, where a woman has many children and various amounts of wages.

The problem can be solved by a brute force method of all variants, but there are too many of them. One of the methods, the least squares method, fits well. However, this method also requires large computational costs. Therefore, it is necessary to choose the appropriate machine learning method.

This article is focused on the study of the algorithm for building a decision tree. The method allows you to build a rating of periods (children) for each region. In Russia, there are regions with multiple numbers of large families and there are regions with small families. A suitable machine learning algorithm will allow training for each region.

The task of machine learning is to obtain a set of data samples and, subsequently, to try to predict the properties of unknown data, i.e. data of a pensioner. For the task "Account for non-insured periods" the rating of the reviewed period of child care is to be taken into consideration.

2. Sampling training

To apply the method, you must create a sampling training. The initial data of the task are: length of service, wages, data on children and data on periods of work.

The paper uses the training sampling, which consists of three input variables (attribute) and one output (response). The training sampling is obtained through random numbers by a sensor. To determine the limits of modeling of each variable, calculations were performed on real data.

Input data:
- insurance experience;
- work experience till 1991;
- entire work experience;
- average pensioner's salary for the period under review;
- average monthly salary per years in Russia.

In addition, data on children and periods of work are needed. The data are used to select periods of work that intersect with periods of child care.

Features of the training sampling (X, y):
- $x_1$ – the ratio of the average salary of a pensioner to the corresponding time average salary in the country;
- $x_2$ – insurance experience;
- $x_3$ – work experience till 1991;
- $y$ – output variable.

Introducing notations:
- $z_1$ – deduction from the length of service, calculated by the rules of the law FZ-4000;
- $z_2$ – deduction from the experience before 1991, calculated according to the rules of the law FZ-400;
- $k$ – coefficient of incomplete experience, calculated according to the rules of the law FZ-400;
- $y$ – rating coefficient of the i-th child care period.

The output variable $y$ is calculated:
- $y = \left( (x_3 - 450) \cdot k \cdot (1.1 + 0.01 \cdot (x_3 - z_2)) \right) \cdot 5.146 / 64.1,$
- $z_3 = \left( (x_2 - z_1 - 20) \cdot 0.01 + 0.55 \cdot x_1 \right) \cdot 1671.$

All samples are created by using a random number generator.

Generated sampling (table 1) responds to quality characteristics: stochasticity, independence and uniformity. The sampling length is 100 and the significance level is 0.002. The dependence of the output variable $y$ on features X meets the requirements of the least squares method.
3. Decision tree

To build the decision tree, Python language and scikit-learn library were used. Scikit-learn is an open-source machine learning library in the Python programming language. It can be used to implement various classification, regression, and clustering algorithms, including SVM, random forest, k-nearest neighbor, and DBSCAN, which are built on the interaction of the NumPy and SciPy libraries with Python [4].

Table 1. A fragment of sampling training

| N  | x₁  | x₂  | x₃  | y        |
|----|-----|-----|-----|----------|
| 1  | 0.07| 41  | 7   | 19.56    |
| 2  | 0.11| 9   | 9   | 5.90     |
| 3  | 0.51| 8   | 7   | 6.84     |
| 4  | 0.65| 7   | 5   | 5.77     |
| ...|     |     |     |          |
| 34 | 0.69| 39  | 17  | 21.24    |
| 35 | 0.73| 19  | 14  | 19.54    |
| 36 | 0.78| 19  | 8   | 17.75    |
| 37 | 0.81| 21  | 11  | 20.06    |
| 38 | 0.95| 17  | 13  | 26.84    |
| ...|     |     |     |          |
| 98 | 0.86| 30  | 10  | 19.89    |
| 99 | 1.72| 26  | 15  | 50.61    |
| 100| 1.84| 30  | 28  | 135.67   |

The regression tree is used [4,5]. The root node is the attribute $x_1$ – the ratio of the average wage of the pensioner to the corresponding time average wage in the country. $x_2$ – insurance experience; $x_3$ – experience prior to 1991. The decision tree is shown in Figure 1.

Figure 1. A fragment of the decision tree
In this binary tree (Fig. 1) left branch – branch "true"; right – "false". The tree has a depth of 19. All cases are reviewed until one example remains in the node. The figure shows that the value of \( y \) increases with the value of the characteristic \( x_1 \). This is logically correct. When you limit the depth of the tree, squared error regression loss increases significantly. Table 2 shows the test sampling.

### Table 2. Test sampling

| \( x_1 \) | \( x_2 \) | \( x_3 \) | \( y \) |
|----------|----------|----------|-------|
| 0.5      | 37       | 10       | 20.1  |
| 1.3      | 8        | 7        | 6.8   |
| 0.3      | 22       | 11       | 20.1  |
| 0.6      | 22       | 16       | 21.1  |
| 0.3      | 24       | 20       | 21.6  |
| ...     | ...      | ...      | ...   |
| 1.4      | 18       | 5        | 18.9  |
| 0.9      | 45       | 44       | 26.0  |
| 1.9      | 33       | 25       | 86.1  |
| 1.2      | 27       | 22       | 55.2  |
| 1.2      | 24       | 20       | 21.7  |

As a result of testing the tree obtained characteristics are shown in tables 3 and 4.

### Table 3. The tree characteristics (1)

| \#  | \( y \) | precision | recall | f1-score | support |
|-----|--------|-----------|--------|----------|---------|
| 1   | 4.0    | 1.00      | 1.00   | 1.00     | 1       |
| 2   | 5.0    | 0.00      | 0.00   | 0.00     | 0       |
| 3   | 10.0   | 0.00      | 0.00   | 0.00     | 1       |
| 4   | 11.0   | 1.00      | 1.00   | 1.00     | 2       |
| 5   | 12.0   | 1.00      | 1.00   | 1.00     | 1       |
| 6   | 13.0   | 1.00      | 1.00   | 1.00     | 1       |
| 7   | 19.0   | 1.00      | 1.00   | 1.00     | 2       |
| 8   | 20.0   | 0.00      | 0.00   | 0.00     | 0       |
| 9   | 21.0   | 0.00      | 0.00   | 0.00     | 0       |
| 10  | 22.0   | 0.00      | 0.00   | 0.00     | 0       |
| 11  | 23.0   | 1.00      | 0.50   | 0.67     | 2       |
| 12  | 29.0   | 1.00      | 1.00   | 1.00     | 1       |
| 13  | 35.0   | 1.00      | 1.00   | 1.00     | 1       |
| 14  | 37.0   | 0.00      | 0.00   | 0.00     | 1       |
| 15  | 54.0   | 0.00      | 0.00   | 0.00     | 0       |
| 16  | 55.0   | 0.00      | 0.00   | 0.00     | 1       |
| 17  | 61.0   | 0.00      | 0.00   | 0.00     | 1       |
| 18  | 70.0   | 0.00      | 0.00   | 0.00     | 1       |
| 19  | 72.0   | 0.00      | 0.00   | 0.00     | 0       |
| 20  | 83.0   | 0.00      | 0.00   | 0.00     | 1       |

The characteristics:
- Precision – the share of objects named positive by the classifier and at the same time really being positive.
- Recall – the proportion of objects of a positive class of all objects of a positive class found by the algorithm.
- F1-score – the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account.
- Support – number of instances of each class.
Table 4. The tree characteristics (2)

|           | micro avg | 0.62 | 0.62 | 0.62 | 16 |
|-----------|-----------|------|------|------|----|
| macro avg | 0.42      | 0.39 | 0.40 | 16   |
| weighted avg | 0.69  | 0.62 | 0.65 | 16   |

The type of averaging performed on the data:
- micro avg – the total true positives, false negatives and false positives;
- macro avg – unweighted mean;
- weighted avg – average weighted by support (the number of true instances for each label). This alters ‘macro’ to account for label imbalance; it can result in an F-score that is not between precision and recall.

The results are satisfactory, but not good. Many objects (examples) are not recognized by the decision tree. There should be at least five examples in the tree sheets to avoid overfitting. The experiments were carried out with many features of different strength. In some cases the result was better, but not much.

4. Conclusion
In the study additional predictor variables were introduced into the training sampling, but this did not provide acceptable results. The restriction on the depth of the tree caused the growth of error values.

In the pension legislation, this task is not the only one. The authors conclude that it is necessary to investigate other more complex machine learning algorithms, or to find a new dependence of the output variable on the input. This dependence is probably to be nonlinear that will determine the choice of a new method of machine learning.

References

[1] Naykhanova L and Naykhanova I 2018 Conceptual model of knowledge base system Journal of Physics: Conf. Series N 1015 International Conference Information Technologies in Business and Industry 2018 032097

[2] ConsultantPlus 2014 The Federal Law of the Russian Federation No. 400-FZ of 28 December 2013 "On Insurance Pensions" Retrieved from: http://www.consultant.ru/document/cons_doc_LAW_156525/

[3] Schmidt B 2018 Monitoring of large families and its importance in assessing the demographic situation of St. Petersburg Bulletin of science and education 4(40) volume 1 pp 116-120

[4] Grus Joel 2017 Data Science. Nauka o dannyh s nulya (St. Petersburg: BHV-Peterburg) p 336

[5] Breiman L, Friedman J, Olshen R and Stone C 1984 Classification and Regression Trees (Monterey: Wadsworth and Brooks) p 368