Optimization of the algorithm for identifying digital traces of schoolchildren in the Altai Territory

V V Zhuravleva¹, A S Manicheva¹, A V Feshchenko² and A V Berestov¹

¹Institute of Mathematics and Information Technologies, Altai State University, 61 Lenin ave., Barnaul, 656049, Russia
²Educational and Scientific Laboratory of Computer Training Tools Of the Institute of Distance Education, Tomsk State University, 36 Lenin ave., Tomsk, 634050, Russia

E-mail: vvzhuravleva@mail.ru

Abstract. The use of digital traces of social networks users has gained great popularity in various studies with the development of methods for analyzing big data. When processing data from social networks users, the problem of the incompleteness of the provided information arises (age, educational institution, year of admission/graduation are not specified). Users with such gaps do not fall into the field of view of the university as a result the number of potential applicants is significantly reduced. The aim of the project is to develop an algorithm for restoring information in a digital trace of a social networks user and its application to identify the group affiliation of schoolchildren whose digital trace contains incomplete information on the grade and place of study. The study was carried out on the data of schoolchildren from "The VKontakte" social network, corresponding to the ninth and eleventh grades, which were divided into four groups. The analysis of unique community subscriptions revealed significant differences in the age groups of schoolchildren. Then, based on gradient boosting, the algorithm that allows restoring missing information in the digital trace of schoolchildren was built. The optimization of the parameters of this algorithm based on numerical experiments allowed obtaining the precision of the order 0.6. The algorithm was used to identify groups of schoolchildren in 9th and 11th grades, in the digital trace of which there was incomplete information. The practical significance of the project is to expand the target audience of future university applicants and, as a result, the opportunity for the university to help them using career guidance measures in social networks to choose an educational program more consciously.

1. Introduction

Many regional universities face the challenge of increasing competitiveness. It can be solved by working with graduates of secondary schools and organizing early career counseling among schoolchildren of the 7th–10th grades. Traditionally, universities hold events related to recruiting and preparing for the successful passing of entrance exams. Non-graduating schoolchildren are involved in the educational and career-oriented activities of the university, highly motivated ones are identified among them and their further support takes place.

Over the past few years, schoolchildren have been massively using social networks for communication, games, and for receiving a wide variety of information through groups of social networks and public (in fact, pages of interest communities).
At the same time, the use of digital traces of social networks users has gained great popularity in various studies in connection with the development of methods for analyzing big data. For example, approaches that allow analyzing user data and interpret them to organize effective information impact are used in political, marketing and other studies [1–4]. In works [5–8], the experience of analyzing of data of applicants on "The VKontakte" social network was presented for, firstly, modeling the signs of schoolchildren’s giftedness by digital traces on the social network, and secondly, predicting their future degree program and invitation to enter recommended direction.

When processing data from social networks users, the problem of the incompleteness of the provided information arises (lack of indication of place of residence, age, educational institution, year of admission/graduation, etc.). Users with such gaps do not fall into the field of view of the university; as a result the number of potential applicants is significantly reduced.

In this regard, the goal of the project was to develop an algorithm for restoring information in the digital traces of a social networks user and use it to identify adolescents whose digital trace contains incomplete information on the grade and place of study.

The main objectives of the project were: 1) collecting user data from "The VKontakte" social network corresponding to the ninth and eleventh grades; 2) the identification of records in the data array with complete and incomplete information on the grade and place of study; 3) the division of records with complete information into age groups; 4) analysis of community subscriptions and determination of the degree of similarity/dissimilarity of age groups of adolescents; 5) the construction of an algorithm for identifying digital traces of schoolchildren (which allows restoring missing information about grade and place of study in a digital trace); 6) optimization of the algorithm by parameters based on numerical experiments; 7) identification of groups that include adolescents with incomplete information on the grade and place of study in the digital trace (automatic assignment to certain groups).

2. Analysis of digital traces in the age groups of adolescents in the Altai Territory according to "The VKontakte" social network

For the study, data on users of the social network "VKontakte" in February 2020, living in the Altai Territory was collected. The total number of users at the age 14-18 years was about 33 thousand people. This is approximately 27% of the population of the region of the corresponding age. In the obtained sample, the adolescents living in urban areas are most widely represented, which, obviously, is explained by the lower availability of Internet communications in rural areas.

Among all users, about 3% had a closed profile and about 52% had information in their profile that allowed them to be identified, relating to one of the groups: 9th grade, 11th grade, HEI, SVE.

The last two groups are considered by us because first-year students of higher educational institutions (HEI) and students of secondary vocational institutions such as colleges, technical schools, etc. (SVE), age overlap with schoolchildren in grades 11 and 9. For this reason, in the absence of complete information on the place of study and years of admission/graduation, identification of the age group of schoolchildren becomes problematic.

In total, about 35% of the data set was made up of users whose profile had partial information about the educational institution and grade, or there was no such information at all. About 10% of the sample could be classified as “garbage”, since the profile did not contain data indicating that they were adolescents aged 14–18, or they were schoolchildren in 10th grades.

Table 1 shows statistics on the number of studied groups of adolescents: total number found on "The VKontakte" social network, of which active users (adolescents who are members of social network groups and have public subscriptions) and the share of active users.

The largest number of users of the studied sample belongs to the 9th grade group, which is partially consistent with the distribution of all adolescents at the age of 14–18 in the Altai Territory to these groups. Indeed, after the end of 9th grade, schoolchildren are determined with the place of further education.
Table 1. Number of users of "The VKontakte" from the Altai Territory whose relationship to groups was established according to the profile.

|                  | 9th grade | 11th grade | HEI | SVE | Total |
|------------------|-----------|------------|-----|-----|-------|
| number found on "The VKontakte" | 13738     | 1794       | 1006| 603 | 17141 |
| active users      | 10213     | 1656       | 960 | 587 | 13707 |
| share of active users | 0.743     | 0.923      | 0.954| 0.973| 0.800 |

Over the past three years, no more than half of adolescents with a certificate of basic general education continue to study in 10th–11th grades. The remaining schoolchildren continue their education in secondary vocational education institutions. It is worth noting that, in general, the number of children born in 2004 and 2003 (most often in 9th–10th grades) is noticeably higher than those born in 2002 (most often these are schoolchildren in 11th grades) and 2001 (as a rule, these are graduates from last year) [9].

Here we note that a group of schoolchildren of the 10th grade did not stand out by us for the reason that the vast majority of these adolescents would be "similar" in interest to 9th grade or 11th grade. If such a schoolchild "resembles" a graduate (we can assume that he has already developed interests in a certain profile of further education), then we can conduct recruiting activities for him. If a 10th grader "resembles" a 9th grader (who still haven't yet decided in which direction of higher education they would like to continue their studies), then it will be possible to carry out career guidance work with him.

For each of the four groups of adolescents, their community subscriptions were uploaded. For this purpose, the tools of the "Portal for working with data from social networks" of the University Consortium of Big Data Researchers were used [10]. The total number of communities amounted to about 200 thousand. Among them, those whose number of subscriptions was at least five were highlighted. Approximately 85% of the communities are widely represented in all selected groups of adolescents. Many of them have content of a regional nature, others relate to topics for a wide range of users.

Table 2 presents the results of pairwise comparison of the studied groups (unique communities in a certain group when compared with another group of adolescents).

Table 2. Table of the number of unique communities in pairwise comparison.

|         | 9th grade | 11th grade | SVE | HEI |
|---------|-----------|------------|-----|-----|
| 9th grade | –         | 6553       | 6901| 7547|
| 11th grade | 96        | –          | 829 | 1449|
| SVE     | 6         | 32         | –   | 1151|
| HEI     | 408       | 695        | 956 | –   |

As a result, the total number of pairwise differences between the 9th grade and 11th grade groups was 6649 communities; between groups of 11th grade and SVE – 861; between groups of 9th grade and SVE – 6907; between groups of 9th grade and HEI – 7955; between groups of 11th grade and HEI – 2144; between the groups of the HEI and SVE – 2107.

As a percentage of the total number of communities, these numbers are small, however, the hypothesis that this amount of uniqueness is enough to identify a user with incomplete information by grade and place of study in the profile, using a comparison of community subscriptions was put forward. It is worth noting that a rather significant part of unique communities is associated with the specifics of learning and can explain the distinguishability of the studied groups.

3. Research methods

Information recovery in the digital trace of a social networks user occurs during the classification of records with missing values. The classification process is based on the use of the CatBoost machine
learning algorithm developed by Yandex for the Python language [11]. This algorithm uses gradient boosting of decision trees. Decision trees contain the vertices where the conditions to be checked are recorded, and the leaves where the final values are written, for example, one of the classes when solving the classification problem [12].

Boosting is a common method for increasing the performance of any machine learning algorithm. The essence of boosting is to train each subsequent model and further reduce errors, using data on errors from the previous model. This method is used to increase the performance of the machine learning algorithm and reduce learning errors [13]. Gradient boosting of decision trees allows using an iterative approach to constructing an additive function in the form of a sum of decision trees in the same way as in the gradient descent method [14].

The CatBoost algorithm works well with categorical variables [15] and is used in various studies [16, 17].

In the case of the problem described above, the algorithm was applied to binarized data. The problem of assessing the quality of data classification with binary features, and especially with unequal classes, is relevant [18]. In the simplest approach, the quality of the model resulting from the use of machine learning algorithms (including CatBoost) is estimated by learning errors, or by the values of precision. The precision of the model can be interpreted as the fraction of objects called positive by the classifier and, at the same time, really positive, and the completeness shows what proportion of objects of the positive class from all objects of the positive class the algorithm found.

4. Digital traces identification algorithm

Figure 1 shows the general scheme of the algorithm for identifying missing information by grade and place of study in the digital traces of "The VKontakte" social network users.

Steps 1 and 2 of the algorithm (data preprocessing and subscription uploading) are discussed in Section 2 of the article. The result is a table that contains the following information: user identifiers, group identifiers, and community identifier lists.

As a result of step 3 (data binarization), a complete list of communities for all users was determined. Next, it was established that each user had a subscription to each of the communities (community id are signs, a value of 1 corresponds to the presence of a subscription, 0 to absence).

At step 4 (reduction of dimension), communities were excluded from the dataset, the total number of subscriptions for which is less than 5.

The separation into the training and test samples in step 5 was carried out in a 3:1 ratio.

In the following steps of the algorithm, a targeted selection of parameters for CatBoost was carried out:

• max_depth – depth of decision trees;
• learning_rate – learning rate in gradient boosting;
• iterations – number of iterations;
• class_weights – set of weights for classes.

5. Optimization of the algorithm by parameters

All steps of the described algorithm, starting with data binarization, are performed in the Python 3.0 programming language.

The values of the parameter sets, as well as the precision of the model for several training options for the model are shown in table 3. Common parameters for all computational experiments were max_depth = 2 and class_weights = [1, 1, 1, 1].
Table 3. Table comparing the precision of the model with various combinations of CatBoost parameters.

| Parameter     | Set 1 | Set 2 | Set 3 | Set 4 | Set 5 | Set 6 | Set 7 | Set 8 |
|---------------|-------|-------|-------|-------|-------|-------|-------|-------|
| learning_rate | 0.03  | 0.03  | 0.05  | 0.05  | 0.1   | 0.1   | 0.3   | 0.3   |
| iterations    | 2000  | 5000  | 2000  | 5000  | 2000  | 5000  | 2000  | 5000  |
| Precision     | 0.51  | 0.53  | 0.53  | 0.53  | 0.55  | 0.56  | 0.55  | 0.55  |

Figure 1. Diagram of the algorithm for identifying missing information in the digital traces of the social network "VKontakte" users.
Figure 2 demonstrate the dependence of the precision of the model on the main parameters.

The optimal values of the algorithm parameters: max_depth = 2; learning_rate = 0.1; iterations = 5000; class_weights = [1, 1, 1, 1]. The best prediction precision for the trained model was 0.6. If we take into account that in the initial aggregate the proportion of "defined"adolescents was 0.52, and the share of "uncertain" 0.32, then with the indicated precision of the prediction, the proportion of adolescents who can be more likely to be attributed to one of the four classes under consideration will increase by 43%.

It is worth noting that with binary classification (division into groups of 9th grade and 11th grade), the result was significantly better. Model precision = 0.78 and completeness = 0.92 with an optimal set of parameters: max_depth = 2; learning_rate = 0.05; iterations = 570; class_weights = [1, 1].

![Figure 2. The dependence of the precision of the model on the number of iterations with different learning rate values (lr).](image)

6. Conclusion

The authors analyzed digital traces in four age groups of adolescents living in the Altai Territory, according to the profile and subscriptions to "The VKontakte" social network communities. The distinguishability of the studied groups (9th grade, 11th grade, higher educational institutions and secondary vocational education) was substantiated based on comparisons of community subscriptions. It has been hypothesized that the number of unique subscriptions is sufficient to determine the groups that include users with incomplete information on the grade and place of study in the profile.

At the next stage, an algorithm for identifying missing information by grade and place of study in the digital traces of adolescents on "The VKontakte" social network was developed. Optimization for CatBoost parameters was done. At the optimal parameter values (max_depth = 2; learning_rate = 0.1; iterations = 5000; class_weights = [1, 1, 1, 1]), the precision of the prediction according to the trained model was 0.6.

Based on the trained model, groups were identified (grade 9, grade 11, higher educational institutions or secondary vocational education) for adolescents with incomplete information on the grade and place of study in the digital trace.

The practical significance of the project is to expand the target audience of future university applicants and, as a result, the opportunity for the university to help them using career guidance measures in social networks to choose an educational program more consciously. Especially relevant this approach is now in the context of the spread of the threat of COVID-19 disease. In Russia, the delivery of the exam has already been postponed to a later date and universities have much less time to conduct recruiting companies. In addition, the traditional offline channels of communication with
applicants (subject Olympiads, open days, forums and others) are now unavailable due to the self-isolation regime introduced in Russia. In these conditions, the transfer of career guidance and recruiting communications to social networks may be the most effective solution.

References
[1] Schwartz H A et al 2013 PloS one 8(9) 73791
[2] Kosinski M et al 2014 Machine learning 95(3) 357–380
[3] Markovikj D et al 2013 7th Int. AAAI Conf. on Weblogs and Social Media 23–26
[4] Mangal N, Niyogi R and Milani A 2016 Computational Science and Its Applications – ICCSA 2016: 16th Int. Conf. V 12–23
[5] Feshchenko A, Goiko V and Stepanenko A 2017 EDULEARN17 Proc. 9th Int. Conf. on Education and New Learning Technologies 6077–6082
[6] Feshchenko A et al 2018 INTED2018 Proc. 12th Int. Technology, Education and Development Conf. 640-646
[7] Kiselev P et al 2019 Postproc. of the 10th Annual Int. Conf. on Biologically Inspired Cognitive Architectures 158–163
[8] Matsuta V V, Kiselev P B, Feshchenko A V and Goyko V L 2017 Psihologiyai Psihotekhnika 4 104–121 [in Russian]
[9] ALTAJKRAJSTAT Available at: https://akstat.gks.ru (Accessed 30 April 2020)
[10] University consortium of big data researchers Available at: https://opendata.university (Accessed 30 April 2020)
[11] CatBoostGitHub Available at: https://github.com/catboost (Accessed 30 April 2020)
[12] Kaftannikov I L and Parasich A V 2015 Vestnik YUUrGU. Komp'yuternye tehnologii, upravlenie, radioelektronika 3(15) 26–32 [in Russian]
[13] Freund Y and Schapire R 1996 Proc. 13th Int. Conf. on Machine Learning 148–156
[14] Druzhkov P N, Zolotyh N Yu and Polovinkin A N 2011 Vestnik YUUrGU. Matematicheskoe modelirovanie i programmirovanie 10 82–89 [in Russian]
[15] Prokhorenkova L, Gusev G, Vorobev A, Dorogush A V and Gulin A 2018 Advances in Neural Information Processing Systems 31 6638–6648
[16] Salahutdinova K I, Lebedev I S and Krivcova I E 2018 Nauchno-tekhnicheskij vestnik informacionnyh tehnologij, mekhaniki i optiki 18(6) 1016–1022 [in Russian]
[17] Gorshenin A K and Martynov O P 2019 Informatics and Applications 13(3) 34–40 [in Russian]
[18] Zhuravleva V V and Mar'in E M 2019 Vserossijskaya konferenciya po matematike s mezhdunarodnym uchastiem MAK: "Matematiki – Altajskomu krayu" 256–257 [in Russian]