Testing Behavior Rate Models on data from Vk.com Social Network

Aleksandra Toropova[a], Tatiana Tulupyeva[b,c]

[a] *Saint-Petersburg State University, Address, St. Petersburg, Index, Russian Federation*
[b] *St. Petersburg Institute for Informatics and Automation of RAS, Address, St. Petersburg, Index, Russia*
[c] *The North-West Institute of management RANEPA, St. Petersburg, Index, Russian Federation*

Abstract

In human science, we often face problem of estimating human behavior parameters. Behavior frequency is one of the most used behavior indicators. Knowing the behavior frequency, we can draw conclusions about significant aspects of behavior both in the present and in the future. It is often impossible to obtain data about behavior frequency directly. Therefore, there is a problem to estimate behavior frequency on limited data such as data about some last episodes of behavior. We propose two models based on Bayesian belief networks to estimate behavior frequency. We use last three behavior episodes, maximum and minimum intervals between the episodes as initial data and test the models on the data set collected from the social network Vk.com.

Keywords 1

Bayesian belief network, hidden variables, behavior rate, behavior episodes, posting behavior, behavior frequency

1. Introduction

In human science, we often face problem of estimating human behavior parameters. One of the most significant indicators of behavior is its frequency. Knowing the behavior frequency (or behavior rate), it is possible to draw conclusions about significant aspects of behavior both in the present and in the future. In [1] the number of training sessions for six months and a year after the survey is predicted, based on respondents ' responses about the frequency of training and the effort involved. In [2, 3], authors propose methods for calculating the probability of spreading multi-pass socioengineering attacks, taking into account data on the frequency of interaction between users of social networks.

Direct observation is the most reliable way to collect information about the behavior rate, but it is not always available [4, 5]. This makes it necessary to develop tools to assess the frequency of behavior based on the data provided by respondents. Using self-reports is a commonly used method [6], but the disadvantage of this method is that it can last quite a long time.

In [7–11] models were presented that assess the behavior rate using information about the last three episodes and about the minimum and maximum intervals between episodes. However, these models include data about the moment of the interview. Since the interview is not an episode of behavior, information about it can distort the assessment of the behavior rate. In this paper, we propose two models that evaluate the behavior rate without using data about the moment of the interview. The difference between proposed models is that one of the models has nodes that characterize real values about the last, minimum, and maximum intervals between episodes, rather than those obtained from respondents.

Russian Advances in Fuzzy Systems and Soft Computing: selected contributions to the 8-th International Conference on Fuzzy Systems, Soft Computing and Intelligent Technologies (FSSCIT-2020), June 29 – July 1, 2020, Smolensk, Russia
EMAIL: alexandra.toropova@gmail.com; tvt@dcsc.pro

Bayesian belief networks are used as the main modeling tool. Bayesian belief networks allow combining different types of information, working with inaccurate data, and having many other useful properties and they find use in many fields, including sociology, economics etc. [12].

To test the models, we use data from the social network Vk.com, as well as data synthesized based on "inaccurate" responses from respondents.

Models description

The main difference between the proposed models from those proposed in [7–11] is that instead of the interval between the last episode of behavior and the interview episode, we consider the interval between the last episode of behavior during the study period and the first episode of behavior at the end of the study period. As an example, we can analyze this situation: the study period is 2019, the last episode of behavior in 2019 occurred on December 26, and the next episode, that is, the first in 2020, occurred on January 8. Thus, we consider the interval from December 26 to January 8, i.e. 13 days.

The figure 1 shows a behavior rate model as a Bayesian belief network [13]. Vertex λ characterizes the behavior rate, *t_12* is the interval between the last and penultimate episodes of behavior, *t_23* is the interval between the penultimate and third from the end of episodes of behavior for the study period, t_*min* and t_*max* are the minimum and maximum intervals between episodes for the study period, *t_next* is the interval between the last episode for the study period and the first episode after the end of the study period, *n* is the number of episodes for the study period.

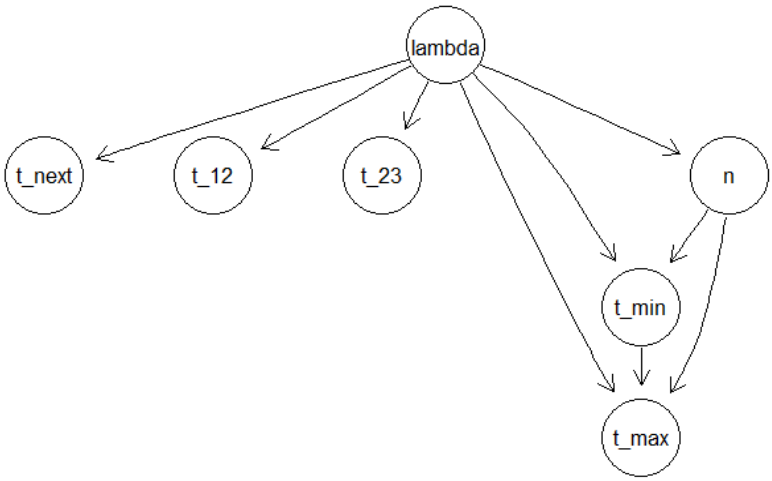

Figure 1: Behavior rate model

In [11], a model of socially significant behavior with hidden variables was presented. This model took into account that the information received from respondents may be incorrect. This may be because in some cases respondents, in order to get social approval, may intentionally distort the actual values, as well as the fact that by answering from memory, respondents may inadvertently make a mistake.

Figure 2 shows a behavior rate model with hidden variables. Vertexes with a zero (t^0_12, t^0_23, min^0, max^0) are information about the corresponding intervals provided by the respondents; the other vertexes are described as for the model above. In this case, we do not include the t^0_next vertex, since a respondent cannot provide information about this episode if it has not already occurred. Thus, *t_next*, *t_12*, *t_23*, *min*, *max*, and *n* are hidden variables that characterize real data about the frequency of behavior.

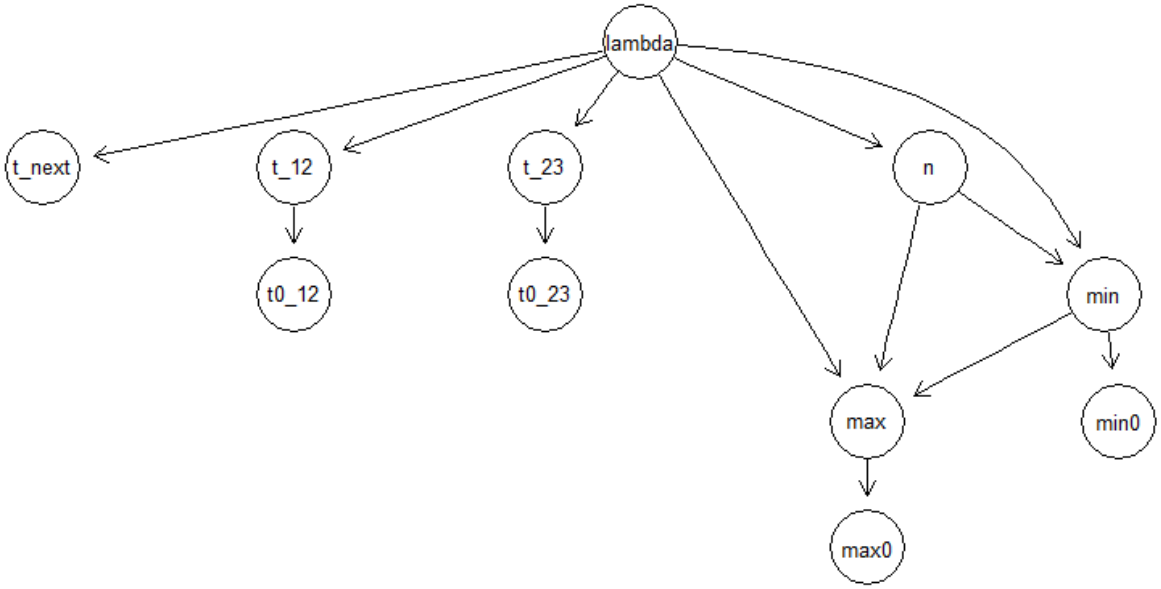

Figure 2: Behavior rate model with hidden variables

2. Data description

To test these models, we used data collected in the social network Vk.com [14], as well as synthesized data based on them as "inaccurate" responses from respondents.

To collect data from Vk.com, a program was written in C# programming language. We also used wall.get method for this purpose. The VK API [15] provides this method. It allows to get information about the last 100 records on the user's wall. This is sufficient if we consider a one month as the study period. In addition, this method has a limit of 5,000 requests per day.

The accounts of users who provided the appropriate permission were processed. The program extracts the time of the last three posts for the study period, the time of the first post made after the study period, the minimum and maximum intervals between the publication of posts for the study period, as well as the number of posts for the study period.

We considered March 2020 as the study period. The selection of users was made randomly. Data about users with closed profiles and those users who did not have enough posts was dropped. In this way, a dataset containing 5338 records was collected.

Data on "real" respondents' answers were synthesized automatically by adding noise as follows: the distance between the last and penultimate episodes in days was calculated and a random value was added so that this distance changed no more than one and a half times, to the distance between the penultimate and the third from the end of the episode, a random value was added so that the distance changed no more than twice. Normal noise was added to the minimum and maximum intervals.

3. Learning the models

All calculations in this and the following sections were performed in R [16] using the bnlearn package [17], which provides work with Bayesian belief networks.

To work with a Bayesian belief network, all continuous data must be sampled. Therefore, the values of variables related to time (we use the day as the unit of measurement), i.e. *t_next*, *t_12*, *t_23*, t^0*_12*, t^0*_23*, *min*, *max* min^0, max^0 were divided into the intervals $t_1=(0;0.1)$, $t_2=[0.1;0.5)$, $t_3=[0.5;1)$, $t_4=[1;7)$, $t_5=[7;10)$, $t_6=[10;20)$, $t_7=[20;\infty)$; the values of the variable λ (behavior rate measured as the number of posts divided by the number of days in the month) were divided into the intervals $\lambda_1=(0;0.1)$, $\lambda_2=[0.1;0.2)$, $\lambda_3=[0.2;0.3)$, $\lambda_4=[0.3;0.5)$, $\lambda_5=[0.5;1)$, $\lambda_6=[2;\infty)$.

3338 records were used for machine learning of models parameters. In other words, tables of conditional probabilities were constructed for all pairs of network vertices connected by an arc. Below is a table of conditional probabilities for the pair λ – t_23 (table 1) for the behavior rate model.

Table 1
Conditional probabilities for the pair λ – t_23

	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
t_1	0,096	0,098	0,131	0,12	0,267	0,379
t_2	0,029	0,063	0,072	0,102	0,135	0,154
t_3	0,052	0,077	0,114	0,145	0,168	0,225
t_4	0,4	0,53	0,542	0,499	0,411	0,231
t_5	0,143	0,117	0,076	0,033	0,017	0,012
t_6	0,24	0,104	0,061	0,022	0,002	0
t_7	0,041	0,01	0,004	0	0	0

4. Predictions

We used 2000 records for testing the models. Synthesized responses from respondents were passed to the models as input data.

After getting the behavior rates predicted by the models, we can compare them with the known post publication frequencies of users. Table 2 is a confusion matrix for the behavior rate model, and table 3 is a confusion matrix for the behavior rate model with hidden variables. The rows represent real frequencies, and the columns represent the frequencies predicted by the models.

Table 2
Confusion matrix for the behavior rate model

	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
λ_1	112	173	5	16	18	0
λ_2	59	347	38	78	37	2
λ_3	1	141	49	94	33	2
λ_4	2	67	37	120	86	9
λ_5	1	10	12	70	161	30
λ_6	0	1	0	16	118	50

Table 3
Confusion matrix for the behavior rate model with hidden variables

	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
λ_1	114	152	18	20	12	3
λ_2	88	310	49	71	37	5
λ_3	16	128	62	69	35	10
λ_4	6	64	53	95	93	7
λ_5	0	16	18	62	153	35
λ_6	0	1	3	28	117	39

In this case, the problem is a classification problem for six classes, so it is worth considering such a characteristic as average accuracy (0.807 and 0.796). The confusion matrices show that most of the values are located on the diagonal or in adjacent cells. This means that even if there is a classification error, the resulting values are most likely located in neighboring classes.

Table 4 compares accuracy, average accuracy, precision, and recall, the main quality metrics.

Table 4
Quality metrics

	Accuracy	Avg. Accuracy	Precision	Recall
behavior rate model	0.42	0.807	0.42	0.387
behavior rate model with hidden variables	0.389	0.796	0.389	0.358

As you can see, the difference in the results is quite small, but the model with hidden variables showed slightly worse results, perhaps this is due to the complexity of the model (4 new vertexes and the arcs between them were added).

5. Conclusion

Two models for evaluating the behavior rate were presented. The main difference between the proposed models is that instead of the interval between the last episode of behavior and the interview episode, we consider the interval between the last episode of behavior during the study period and the first episode of behavior at the end of the study period. One of the models is based on the fact that respondents may provide incorrect information in some cases.

To learn and test the models, we used data on posting from the social network Vk.com, as well as synthesized data on "inaccurate" responses from respondents.

The results obtained can be used in tasks that require data on the behavior rate, when having a limited initial data. Taking into account a small amount of initial data, the models showed a fairly high quality of behavior rate classification.

6. Acknowledgements

The research was carried out in the framework of the project on state assignment SPIIRAS No. 0073-2019-0003, with financial support from the Russian Foundation for Basic Research, projects No. 19-37-90120, No. 18-01-00626 and No. 20-07-00839.

7. References

[1] D. Garcia, T. Daniele, T. Archer, A brief measure to predict exercise behavior: the Archer-Garcia ratio, Heliyon, 2017. doi:10.1016/j.heliyon.2017.e00314.

[2] A. O. Khlobystova, M. V. Abramov, A. L. Tulupyev, Soft Estimates for Social Engineering Attack Propagation Probabilities Depending on Interaction Rates Among Instagram Users, in: International Symposium on Intelligent and Distributed Computing. Springer, Cham, 2019. 272-277.

[3] A. O. Khlobystova, M. V. Abramov, A. L. Tulupyev, A. A. Zolotin, Search for the shortest trajectory of a social engineering attack between a pair of users in a graph with transition probabilities, Information and Control Systems, 6 (2018) 74-81.

[4] G. R. Mayer, B. Sulzer-Azaroff, M. Wallace, Behavior analysis for lasting change, Cornwall-on-Hudson, NY: Sloan Publishing, 2018.

[5] R. A. Rehfeldt, Clarifying the nature and purpose of behavioral assessment: A response to Newsome et al., Journal of Contextual Behavioral Science, 4 (2019) 37-39.

[6] D. Newsome, K. Newsome, T. C. Fuller, S. Meyer, How contextual behavioral scientists measure and report about behavior: A review of JCBS, Journal of Contextual Behavioral Science, 12 (2019) 347-354.

[7] A. V. Suvorova, Models and Algorithms for analysis of super-short granular time series on the base of Bayesian belief networks, St.Petersburg, 2013.

[8] A. V. Suvorova, Socially significant behavior modeling on the base of super-short incomplete set of observations, Information-measuring and Control Systems, 9, 11 (2013) 34-38.
[9] A. V. Suvorova, A. L. Tulupyev, A. V. Sirotkin, Bayesian belief networks in problems of estimating the intensity of risk behavior, Journal of Russian Association for fuzzy systems and soft computing, 9, 2 (2014) 115-129.
[10] A. V. Suvorova, Models for respondents' behavior rate estimate: bayesian network structure synthesis, in: Proceedings of 2017 XX IEEE International Conference on Soft Computing And Measurements (SCM), 2017, 87–89.
[11] A. V. Toropova, T. V. Tulupyeva, Synthesis and learning of socially significant behavior model with hidden variables, Advances in Intelligent Systems and Computing, 875 (2019) 76-84.
[12] A. V. Toropova, Approaches to the data coherence diagnosis in bayesian belief network models, SPIIRAS Proceedings, 6 (2015) 156-178.
[13] A. L. Tulupyev, S. I. Nikolenko, A. V. Sirotkin, Fundamentals of Bayesian Network Theory: A Textbook, St. Petersburg, SPbSU Publ, 2019.
[14] Vk.com, 2020. URL: http://www.vk.com/.
[15] Vk.com for developers, 2020. URL: https://vk.com/dev/methods.
[16] R Core Team, R: A language and environment for statistical computing, 2020. URL: https://www.R-project.org/.
[17] M. Scutari, Learning Bayesian Networks with the Bnlearn R Package, arXiv preprint. arXiv:0908.3817, 2009.