Learning Behavior Rate Models on Social Network Data

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

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

Abstract

Intensity is one of the main characteristics of human behavior, using data about behavior intensity we can make high enough quality predictions about future human behavior. But it is often impossible to get a direct behavior rate, because of high cost, time consumption or restrictions for monitoring private lives, so we need tools to estimate it indirectly. We offer two models for behavior rate evaluation with expert-defined and learned structures. These models are Bayesian belief networks. They include information about the intervals in days between the last three behavior episodes of the study period, the minimum and maximum intervals between episodes, and the interval between the last episode of the study period and the next episode, respectively, after the end of the study period. As we need for the models approbation an example of behavior allowing us to get direct behavior rate, we take users' posting behavior in social network. For learning parameters and structure one of the models, testing models, data from the social network Vkontakte for December 2019 was collected. This data includes an information about posting on own users' "walls" for this month, i.e. posts quantity, time of last three posts, maximum and minimum time interval between posts for December 2019, and time of the first post starting from January 2020.

1. Introduction

Many sciences studying the behavior consider its characteristics. Intensity is one of the main characteristics of human behavior. In [Abr18, Aza16], taking into account the intensity of risky user behavior, conclusions are made about the success of socio-engineering attacks. In [Gar17], based on the respondents' data on the intensity and efforts of training, the number of training sessions in the future is estimated. In [Jil20] efforts on making improvements to farmers' markets are evaluated by accounting farmers' market shopping frequency and fruit and vegetable consumption among farmers' market customers before and after establishing improvements. In [Kim20] it is showed how intensity of parents' alcohol consumption impacts stress in their children.

The behavior rate evaluates the behavior intensity as a mean number of behavior episodes during a particular period [Tor19].

Russian Advances in Artificial Intelligence: selected contributions to the Russian Conference on Artificial intelligence (RCAI 2020), October 10-16, 2020, Moscow, Russia
alexandra.toropova@gmail.com (A.V. Toropova); tvt@dscs.pro (T.V. Tulupyeva)
0000-0001-7311-6192 (A.V. Toropova); 0000-0003-3630-7971 (T.V. Tulupyeva)

Obviously the direct observation is the most reliable way to obtain a behavior rate value [May19, Reh19], but it can not possible in many situations, because of high cost, time consumption or restrictions for monitoring private lives, so tools are needed to estimate it indirectly. One of the most popular approaches to get information about human behavior is self-reports survey [New19]. In self-reports respondents fill questionnaires, polls or surveys answering questions, that may relate to different types: open or closed (with limited choice) or rating scales (e.g. Likert scale). This method has certain limitations: respondents can give unreliable answers because of memory issues, faking answers to paint themselves in the best light etc. Another approach to get information about behavior intensity is "diary" method, when respondents note information about the behavior in a diary for a certain period of time and then researchers analyze these diaries.

In [Suv13a, Suv13b, Suv17, Suv14, Tor19], models based on Bayesian belief networks are proposed that use information about the last three episodes of behavior. However, all the proposed models take into account the moment of the interview, which is not an episode of the studied behavior, and therefore may give distortions in the behavior rate assessment. Thus, it would make sense to consider possible models that also take into account information about recent episodes of behavior, but do not include information about the interview itself.

This paper presents two models for evaluating the behavior rate with expert-defined and learned structures. Models are Bayesian belief networks [Tul19]. They include information about the intervals between the last three episodes of the study period, the minimum and maximum intervals between episodes, and the interval between the last episode of the study period and the next episode, respectively, after the end of the study period.

Data on posting in the social network Vkontakte [Vko20a] during December 2019 was collected to build the structure of one of the models, to learn and test the models.

2. Data

Collecting a fairly large amount of data with well-known information about the behavior intensity is not always possible for a number of reasons. Constant monitoring of a large group of subjects is practically impossible both from the technical and financial side. Conducting a survey or interviewing respondents can give no guarantee that the received information is correct. Using social networks is a good opportunity to get accurate data about users' behavior, because every action is fixed by time, and it is possible to get direct behavior rate value. Social networks creating an ecosystem for different types of behavior: making connections, liking, disliking, following, expressing emotions, making statements etc. Thus social networks provide a great opportunity for behavior research. However, this approach is also not perfect, since using the social media API also implies a number of limitations.

In our study, the data was collected from the social network Vkontakte [Vko20a], the largest in Russia [Sim20]. This social network also provides for users different types of activities. Each user of this network has a so-called "wall", on which one can publish posts, make reposts of other users' records, and can also leave comments on the "walls" of other users. Users also can follow each other, connect as friends, show relations with someone, like different types of data and other things, Users may have a "private" profile, in that case access to their "wall" is

restricted.

For approbation of the models we chose such type of behavior as publishing posts on user's own wall, i.e. posting. This is one of the simplest type of behavior that can be assessed properly, by considering records on user's wall excluding records made by others.

To get data from Vkontakte, a special program was written in C#. The "wall.get" method was used, which is provided by the VK API [Vko20b]. This method allows to get information about the last hundred records from the user's wall, and you can apply the condition that these records belong to the owner of the "wall", and not to other users. This number of records is sufficient if we consider one month as the study period. The use of this method is limited to 5000 calls per day.

December 2019 was taken as the study period. Accounts of users who provided access to their "wall" were randomly selected and processed. For each user, the following information was collected for December 2019: the time of the last three entries, the maximum and minimum time between entries, and the number of entries; in addition, the time of recording the first entry in 2020 was saved. Users with insufficient information were removed from the data-set. In this way, information about 6556 users was collected.

4556 records were used for learning models, and the remaining 2000 records were used for testing them

3. Models

We propose models based on Bayesian belief networks [Tul19]. This is a capable and popular tool that has found application in many branches of science [Tor15], it shows strong ability on combining data from different sources, doing probabilistic reasoning [Zha19]. In addition there are many powerful software packages [Bay20, Net20, Scu10] making it easier to work with this tool.

All calculations in this and the following sections were performed in R [R20] using the bnlearn package [Scu10], which provides work with Bayesian belief networks.

To work with Bayesian belief networks, all continuous data needs to be sampled. Therefore, the values of variables related to time (we use a day as the unit of measurement), i.e. t_next, t_12, t_23, t_min and t_max, were divided into the intervals $t_1 = (0; 0.1)$, $t_2 = [0.1; 0.5)$, $t_3 = [0.5; 1)$, $t_4 = [1; 7)$, $t_5 = [7; 10)$, $t_6 = [10; 20)$ and $t_7 = [20; \infty)$; and the values of the variable λ (we will measure the behavior rate as the number of posts divided by for the number of days in a month) were divided into the intervals $\lambda_1 = (0; 0.1)$, $\lambda_2 = [0.1; 0.2)$, $\lambda_3 = [0.2; 0, 3)$, $\lambda_4 = [0.3; 0, 5)$, $\lambda_5 = [0.5; 0.7)$, $\lambda_6 = [0.7; 1)$ and $\lambda_7 = [1; \infty)$.

3.1. Behavior Rate Model with an Expert-Defined Structure

Figure 1 shows a behavior rate model with an expert-defined structure, which is a Bayesian belief network. The calculation of conditional probability tensors that characterize transitions between network nodes is based on the data and is described in the next section.

The node λ characterizes the behavior rate, t_next is the interval between the last episode for the study period and the first episode after the end of the study period, 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 the behavior episodes for the study period, t_min and t_max are the minimum and maximum intervals between episodes for the study period, n is the number of episodes for the study period.

The proposed model is based on the model from [Suv13a]. The difference is that instead of information about the interval between the last episode and the moment of the interview, the model includes information between the last episode during the study period and the next episode that occurred after the study period. This makes sense, since the moment of the interview is not an episode of the studied behavior and does not provide any useful information about the latter.

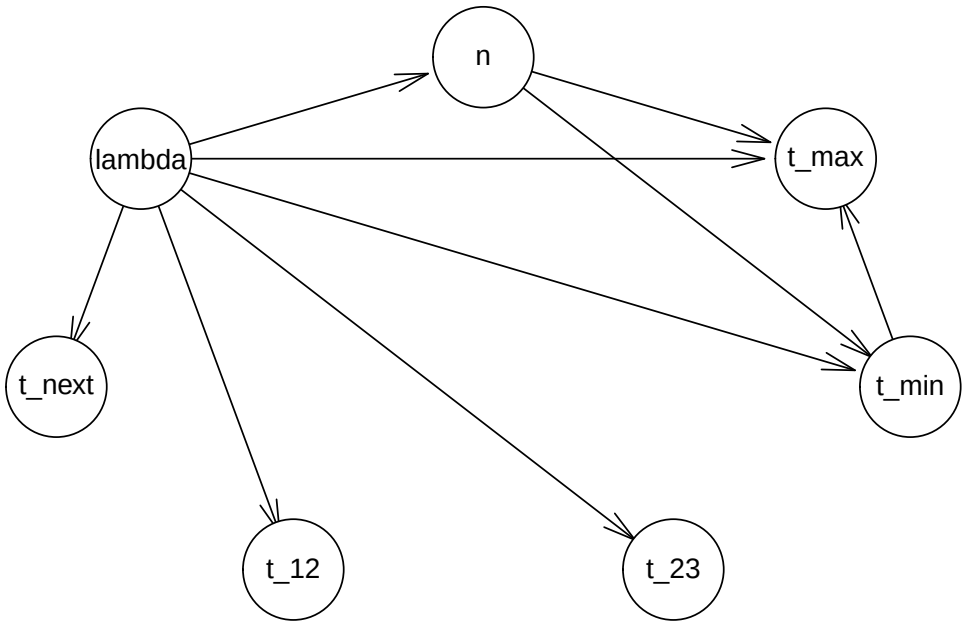

Figure 1: Behavior Rate Model with an Expert-Defined Structure

3.2. Behavior Rate Model with a Learned Structure

In order to construct the network structure from data, we used the Hill-Climbing greedy search on the space of directed graphs [Scu10]. It is one of score-based algorithms, which assign a score to each candidate Bayesian network with respect to the training data-set and try to maximize it. We used the Bayesian information criterion (BIC) as a quality score. BIC is equivalent to the Minimum Description Length (MDL) and is also known as Schwarz Information Criterion [Sch78, Scu10, Gam10].

Figure 2 shows the resulting structure. As we see connections between λ, t_next and n are saved. t_12 and t_12 are always between t_min and t_max, what caused occurrence of

arcs between these nodes. t_next does not depend on t_min and t_max, because these are extremum points only during the study period, and t_next occurs after it. The arc λ–t_next shows inverse dependence comparing with the expert-based structure, this can be connected with the specific characteristics of the training data-set.

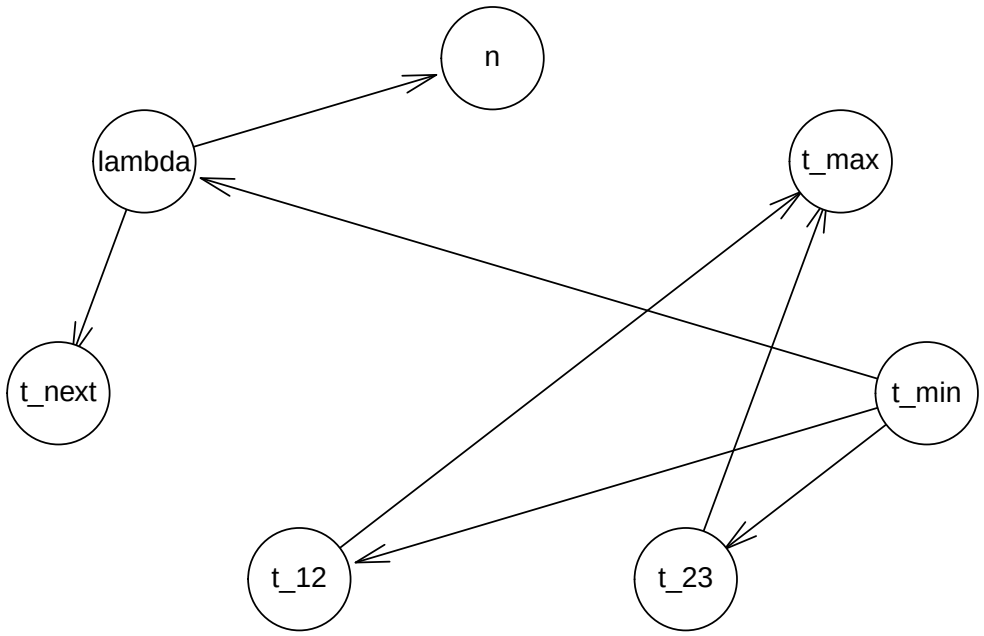

Figure 2: Behavior Rate Model with a Learned Structure

3.3. Learning Parameters

On this step we have two models' structures. To define the models completely, further learning of the models' parameters using the training data-set was conducted. In other words, tables of conditional probabilities were constructed for all pairs of network vertices connected by an arc. For example, table 1 is a conditional probabilities table for the pair $\lambda - t\_next$ (table 1) in a model with the expert-defined structure. After that we can use these models for making predictions of behavior rates.

Table 1
Conditional Probability Table

	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7
t_1	0.001	0	0.001	0.002	0	0	0.009
t_2	0.001	0.006	0.007	0.005	0.032	0.071	0.054
t_3	0.006	0.021	0.443	0.069	0.07	0.111	0.143
t_4	0.176	0.257	0.411	0.494	0.601	0.622	0.673
t_5	0.098	0.12	0.115	0.156	0.12	0.062	0.045
t_6	0.272	0.298	0.265	0.186	0.136	0.11	0.067
t_7	0.446	0.298	0.157	0.088	0.041	0.022	0.009

4. Comparison of the Models

Let us compare the received models. According to the construction algorithm, on the training set, the quality measure of the structure shown in fig. 2 is higher than the initial one set by experts (Fig. 1), as for BIC (-38789 and -54736, respectively), and for the maximum likelihood measure (-36552 and -39762). On the test dataset, the quality measures of the data-learned structure are also higher (BIC: -18379 and -30624; maximum likelihood: -16142 vs -17113).

However, since the main task of the models is to evaluate the behavior rate, let us move on to the next stage, namely, comparing their prediction quality.

After the models predict the behavior rates, these predictions can be compared with the known user posting intensities. Table 2 is a confusion matrix for the behavior rate model with the expert-defined structure, and table 3 is a confusion matrix for the behavior rate model with the learned structure. The rows represent the real intensities, and the columns represent the intensities predicted by the model.

Table 2
Confusion Matrix for the Behavior Rate Model with the Expert-defined Structure

		Predicted Behavior Rates						
		λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7
Behavior Rates	λ_1	115	125	22	9	8	19	7
	λ_2	61	352	51	95	12	29	16
	λ_3	6	119	66	141	12	4	20
	λ_4	2	43	54	177	23	21	27
	λ_5	0	8	7	67	31	18	37
	λ_6	0	2	3	25	12	14	40
	λ_7	0	1	0	16	14	16	53

Table 4 shows the main quality metrics: accuracy, average accuracy, precision, and recall for

Table 3
Confusion Matrix for the Behavior Rate Model with the Learned Structure

		Predicted Behavior Rates						
		λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7
Behavior Rates	λ_1	40	197	2	65	0	0	0
	λ_2	6	348	7	249	0	0	0
	λ_3	0	118	9	239	0	0	0
	λ_4	0	64	20	263	0	0	0
	λ_5	0	17	10	140	0	0	0
	λ_6	0	9	4	82	0	0	0
	λ_7	0	8	4	88	0	0	0

both the models.

Table 4
Comparison of Quality Metrics

	Accuracy	Avg. Accuracy	Precision	Recall
Behavior Rate Model with the Expert-defined Structure	0.404	0.83	0.404	0.357
Behavior Rate Model with the Learned Structure	0.332	0.809	0.332	0.212

As we can see from the table 4, the difference in quality metrics is not very large, but the behavior rate model with the expert-defined structure showed higher results. In addition, since in this case, the problem can be reduced to classification by seven classes, comparing table 2 and table 3 we can see that the model with the expert-defined structure, even in the case of an error, is likely to place the evaluated value in an adjacent class, while the model with the learned structure does not consider classes with high behavior rate (starting from λ_5).

5. Conclusion

Two models for evaluating the behavior rate with the expert-defined and the learned structures were presented. The models are Bayesian belief networks. They include information about the intervals in days between the last three episodes during the study period, the minimum and maximum intervals between the episodes, and the interval between the last episode during the study period and the next episode, respectively, after the end of the study period.

Data on posting in the social network Vkontakte during December 2019 was collected to build the structure of one of the models, to learn models' parameters and to test the models.

Despite the fact that the model with the learned structure showed higher metrics of structure quality, in the behavior rate predictions the model with the expert-defined structure showed better results. In this regard, it is recommended to use the behavior rate model with the expert-defined structure.

Further we plan to consider the other discretization of continuous data, because this can affect on the received results. Also it can be interesting to test models on another data-set obtained from the different source.

The application of this model can be found in many areas related to the behavior intensity, for example, in sociology, epidemiology, etc.

The main advantages of suggested models is that sufficiently accurate results can be obtained using a very small amount of data: information about last and extremum episodes can be remembered quite easily comparing remembering of all episodes of behavior. Thus researchers can use questionnaires with questions about this information for studying any behavior intensity.

In case of absence enough real data for training-set researchers can use synthesized data according to assumptions about the studied behavior.

As for social network behavior research these models also can be useful. As we could see social networks' API not always allow to collect all the data that researcher need, so using these models it is possible to get more information about some kind of user behavior. Besides that as the suggested models include the data about the next episode, which can happen in the future, using information about other nodes, we can predict approximate time of this next episode.

Acknowledgments

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.

References

[Abr18] M. V. Abramov, T. V. Tulupyeva, A. L. Tulupyev. Sotcioinzhenernye ataki: sotcialnye seti i ocenki zashchishchennosti polzovatelei. SPb.: GUAP,. 266 p. ISBN 978-5-8088-1377-5, 2018. (In Russian).

[Aza16] A. A. Azarov, T. V. Tulupyeva, A. V. Suvorova, A. L. Tulupyev, M. V. Abramov, R. M. Iusupov. Sotcioinzhenernye ataki: problemy‘ analiza. SPb.: Nauka. ISBN: 978-5-0203-9592-3, 2016. (In Russian).

[Bay20] BayesFusion. https://www.norsys.com/, last accessed 2020/04/20.

[Gam10] J. Gamez, J. Mateo, J. Puerta. Learning Bayesian networks by hill climbing: Efficient methods based on progressive restriction of the neighborhood. *Data Mining and Knowledge Discovery*, 22, 106–148, doi: 10.1007/s10618-010-0178-6, 2010.

[Gar17] D. Garcia, T. Danielem, T. Archer. A brief measure to predict exercise behavior: the Archer-Garcia ratio. *Heliyon*, doi: 10.1016/j.heliyon.2017.e00314, 2017.

[Jil20] S. B. Jilcott Pitts, Q. Wu, W. Gray, M. J. Lyonnais. Examining changes in farmers' markets and in customers' farmers' market shopping frequency and fruit and vegetable purchase and consumption: evaluation data from the Partnerships to Improve Community Health Project, 2014–2017 *Journal of Hunger and Environmental Nutrition*, 15(1), 107–117. DOI: 10.1080/19320248.2018.1512924, 2020.

[Kim20] S. Kim, W. Chae, S. H. Min, Y. Kim, S.-I. Jang. Alcohol consumption frequency of parents and stress status of their children: Korea national health and nutrition examination survey (2007–2016) *International Journal of Environmental Research and Public Health*, 17(1), doi:10.3390/ijerph17010257, 2020.

[May19] G. R. Mayer, B. Sulzer-Azaroff, M. Wallace. Behavior analysis for lasting change. Corn-wall-on-Hudson, NY: Sloan Publishing, 2019.

[Net20] Netica Bayesian network software package. https://www.norsys.com/, last accessed 2020/04/20.

[New19] 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, 347–354, doi:10.1016/j.jcbs.2018.11.005, 2019.

[R20] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/, last accessed 2020/04/20.

[Reh19] R. A. Rehfeldt. Clarifying the nature and purpose of behavioral assessment: A response to Newsome et al. *Journal of Contextual Behavioral Science*, 14, 37–39, doi:10.1016/j.jcbs.2019.09.001, 2019.

[Sch78] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.

[Scu10] M. Scutari. Learning Bayesian networks with the bnlearn R package. *Journal of Statistical Software* , 35, 2010.

[Sim20] SimilarWeb. https://www.similarweb.com/fr/top-websites/russian-federation, last accessed 2020/04/20.

[Suv13a] A. V. Suvorova. Modeli i algoritmy analiza sverkhkorotkikh granulyarnykh vremennykh ryadov na osnove bayesovskikh setey doveriya. PhD, Diss. [Models and Algorithms for analysis of super-short granular time series on the base of Bayesian belief networks.]. St.Petersburg, 2013. (in Russian).

[Suv13b] A. V. Suvorova. Socially significant behavior modeling on the base of super-short incomplete set of observations. *Information-measuring and Control Systems*, 9(11) 34–38, 2013. (in Russian).

[Suv17] A. V. Suvorova. Models for respondents' behavior rate estimate: bayesian network structure synthesis. *Proceedings of 2017 XX IEEE International Conference on Soft Computing And Measurements (SCM)*, 87–89, 2017.

[Suv14] 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. *Journal of Russian Association for fuzzy systems and soft computing*, 9(2) 115–129, 2014.

[Tor15] A. V. Toropova. Approaches to the data coherence diagnosis in bayesian belief network models. *SPIIRAS Proceedings*, 6(43), 156–178, 2015.

[Tor19] A. V. Toropova, T. V. Tulupyeva. Synthesis and learning of socially significant behavior model with hidden variables. *Advances in Intelligent Systems and Computing*, 875, 76–84, 2019.

[Tul19] A. L. Tulupyev, S. I. Nikolenko, A. V. Sirotkin. Osnovy teorii bayesovskikh setey: uchebnik. [Fundamentals of Bayesian Network Theory: A Textbook], St. Petersburg, SPbSU Publ, 399 p., 2019. (In Russian).

[Vko20a] Vkontakte. https://vk.com/, last accessed 2020/04/20.

[Vko20b] Vkontakte for Developers. https://vk.com/dev/methods, last accessed 2020/04/20.

[Zha19] J. Zhang, H. Yue, X. Wu, W. Chen. A brief review of Bayesian belief network *Proceedings of the 31st Chinese Control and Decision Conference*, 3910–3914, doi:10.1109/CCDC.2019.8832649, 2019.