Bayesian Belief Networks in Risky Behavior Modelling

Alena Suvorova and Tatiana Tulupyeva

Abstract The area of risky behavior modelling has unsolved issues in current practice: there is a need for numerical estimates of risky behavior rate. We propose the approach for risky behavior modelling in terms of Bayesian Belief Networks on the base of the data about behavior episodes. The paper includes the description of the model, results of model testing on automatically generated dataset and discussion of possible further development.

Keywords Bayesian belief network · Risky behavior · Rate estimates · Poisson random process

1 Introduction

Bayesian Belief Network (BBN) is a type of probabilistic graphical models that represents a set of random variables and their conditional dependencies [1]. Formally, BBN is a directed acyclic graph that should satisfy several requirements [1]. In general BBNs represent complex influencing factor relationships, they are applied in wide range of areas: finance, medicine, information technology [2–6]. BBNs are known as good representation of knowledge and decision support under

A. Suvorova (✉) · T. Tulupyeva
SPIIRAS, Saint Petersburg, Russia
e-mail: suvalv@gmail.com

T. Tulupyeva
e-mail: tvt100a@mail.ru

A. Suvorova
National Research University Higher School of Economics, Saint Petersburg, Russia

T. Tulupyeva
Saint Petersburg State University, Saint Petersburg, Russia

T. Tulupyeva
North-West Institute of Management of the Russian Presidential Academy
of National Economy and Public Administration, Saint Petersburg, Russia

A. Abraham et al. (eds.), *Proceedings of the First International Scientific Conference “Intelligent Information Technologies for Industry” (IITI’16)*, Advances in Intelligent Systems and Computing 451, DOI 10.1007/978-3-319-33816-3_10

uncertainty [7] and one of their important advantage is the ability to combine different source of information [8].

These features of BBN play an important role in research that involves both empirical data and expert knowledge [9]. One of such research areas is individual's risky behavior modelling. On the one hand, researchers collect data about respondents' risky behavior in different field studies; on the other hand, there are associations, causalities and assumptions based on results of previous studies, facts and theories from other research areas. To construct better models of respondents' behavior and, furthermore, to compute more accurate characteristics of that behavior, it is important to combine both data-based and expert-based information.

Knowledge of risky behavior characteristics supports decision making in many practical issues. Unusual behavior of the user of information system can be a marker that defines insecure events. Risky sexual behavior (e.g. having multiple sexual partners, unprotected sex) is widely known to be associated with high risk of sexually transmitted infections, including HIV-infection [10]. The "gold standard" for behavior rate measuring is a diary method, in other words, simple recording of episodes [11]. However this method is extremely time-consuming, resource-consuming and even hardly possible for many kinds of behavior [12], for example, due to private nature of the behavior [13] or social desirability bias. In [14] authors provided the method based on data about several behavior episodes.

Risky behavior studies were mostly focused on exploring factors associated with risky behavior or factors influenced by that behavior [15–17]. The results were mostly data-based with expert knowledge included in the form of variable selection. For those studies regression models were primary and dominating method especially in medicine and public health [16, 17]. Also the results did not provide any numerical characteristic of behavior, e.g. rate, frequency or risk that can be further implemented into automated system for decision making.

Human reliability analysis that studies causes, consequences and contributions of human failures in socio-technical systems [9] is close to risky behavior modelling in many parts: it deals with both expert information and empirical data, it faces with uncertainty of initial data, and it is related to human behavior. Human reliability analysis is based on many methods and now BBNs received an increasing attention [9].

However, this practice can not be applied directly to risky behavior modelling. The closest outcome in human reliability analysis that can be considered as behavior characteristic is probability of human error [9] but we need more detailed outcome for risky behavior rate estimate, not just "it happens with probability p". The other issue is an availability of data sources: there is no rather simply collected data from technical systems, the information about behavior comes from self-reports about risky behavior episodes.

The purpose of the paper is to describe the approach for risky behavior modelling in terms of Bayesian Belief Networks.

2 Risky Behavior Modelling

2.1 Initial Data

The model [18] is based on data about the length of intervals between three last episodes of risky behavior and the length of minimum and maximum intervals between episodes during period of interest *T*. The data about episodes in most applications is obtained from respondents' self-reports [14]. We assume that for each respondent occurrence of episodes follows Poisson random process: the occurrence of the next episode is independent from the previous ones, length of interval between concurrent episodes follows exponential distribution. This assumption corresponds to the features of risky behavior and, at the same time, allows less complicated calculations.

Adding data about minimum and maximum intervals decreases the influence of recent behavior represented by the last episodes. However, combining all the data about episodes leads to very complicated joint distribution [19] even in case of Poisson random process and requires much more calculation for behavior rate estimate. Any change or revision of the model, again, will require re-calculation of joint distribution (if it will be possible in elementary functions).

On the contrary, as it was mentioned earlier, BBN allows determining complex relationships in terms of simpler dependencies between small parts. Modelling risky behavior as BBN gives a way to add all available data into the model as well as include expert assumptions about relationships between them and their distributions. Revising the assumptions or adding new variables to the model requires re-arranging small part of the BBN only. Moreover, the existence of software tools (including freeware) for dealing with BBNs, for example [20] or [21], allows researchers focus on description of the model while calculations are performed automatically. Considering these advantages of BBNs we proposed a BBN for risky behavior modelling.

2.2 Model Description

The structure of BBN model is a graph $G(V, L)$ with vertices $V = \{t_{01}, t_{12}, t_{23}, t_{\min}, t_{\max}, \lambda, n\}$ and edges (or links) $L = \{(u, v) : u, v \in V\}$ (Fig. 1), where λ is random variable for behavior rate; t_{ij} is random variable for the length of the interval between *i*th and *j*th episodes from the end (0 corresponds to interview moment); $t_{\min}$ and $t_{\max}$ are random variables for the length of minimum and maximum intervals; *n* is random variable for the number of episodes during period of interest.

All variables were discretized. Conditional probabilities for variables were as follows [18] (where $l_s = 1, \ldots, k_s$, k_s was the number of disjunctive intervals for

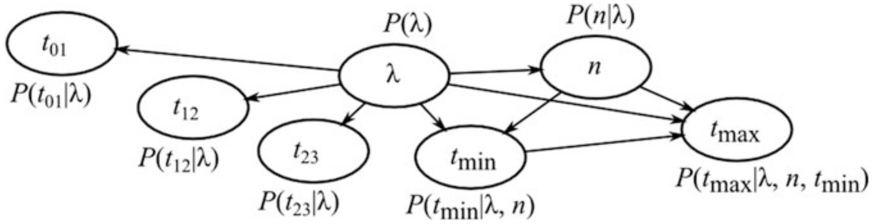

Fig. 1 Bayesian Belief network for risky behavior modelling

discretized $t_{j,j+1}$ variable; $s = 0, \ldots, 4$; $j = 0, 1, 2$; $i = 1, \ldots, m$, m was the number of disjunctive intervals for discretized λ variable):

$$p\left(t_{j,j+1}^{(l_j)}\middle|\lambda^{(i)}\right) = e^{-a\lambda^{(i)}} - e^{-b\lambda^{(i)}},\ t_{j,j+1}^{(l_j)} = [a;b);$$

$$p\left(t_{\min}^{(l_3)}\middle|n, \lambda^{(i)}\right) = e^{-an\lambda^{(i)}} - e^{-bn\lambda^{(i)}},\ t_{\min}^{(l_3)} = [a;b);$$

$$p\left(n\middle|\lambda^{(i)}\right) = \frac{\left(\lambda^{(i)}T\right)^n}{n!}e^{-\lambda^{(i)}T};$$

$$p\left(t_{\max}^{(l_4)}\middle|n, \lambda^{(i)}, t_{\min}^{(l_3)}\right) = e^{(n-1)\lambda^{(i)}t_{\min}^{(l_3)}}$$
$$\times\left(\left(e^{-\lambda^{(i)}t_{\min}^{(l_3)}} - e^{-\lambda^{(i)}b}\right)^{n-1} - \left(e^{-\lambda^{(i)}t_{\min}^{(l_3)}} - e^{-\lambda^{(i)}a}\right)^{n-1}\right),\ t_{\max}^{(l_4)} = [a;b).$$

For all further examples we used the following discretization: for the rate variable λ $\lambda^{(1)} = [0, 0.01)$, $\lambda^{(2)} = [0.01, 0.03)$, $\lambda^{(3)} = [0.03, 0.05)$, $\lambda^{(4)} = [0.05, 0.1)$, $\lambda^{(5)} = [0.1, 0.2)$, $\lambda^{(6)} = [0.2, 0.5)$, $\lambda^{(7)} = [0.5, 1)$, $\lambda^{(8)} = [1, \infty)$; for the variables $t_{j,j+1}$, $t_{\min}$, $t_{\max}$ $t^{(1)} = [0, 0.1)$, $t^{(2)} = [0.1, 1)$, $t^{(3)} = [1, 7)$, $t^{(4)} = [7, 30)$, $t^{(5)} = [30, 180)$, $t^{(6)} = [180, \infty)$.

The model was represented in GeNIe&Smile [20]. The calculations and statistical analysis were performed using Smile library [20] and R [22].

2.3 Testing Results

To test the model we used automatically generated dataset. First, we generate 200 values for behavior rate that followed Gamma distribution with the shape $k = 1.1$ and the scale $\theta = 0.1$. The parameters were chosen to produce more "real-behavior"-like dataset with behavior rate in most cases less than 1 and concentrated around 0.1 episodes per day. Next, for each rate value according to assumptions of the model we generated 20 "respondents" or 20 sequences of behavior episodes each of those for period of 180 days in total. In other words, we had 4000 sequences of episodes. Than we calculated lengths of minimum, maximum intervals between episodes and lengths of intervals between the last three episodes. After deletion of incomplete cases (e.g. cases with only one episode during 180 days) the final dataset consisted of 3672 cases (or “respondents”).

All variables were discretized as it was described in Sect. 2.2. The prior distribution of behavior rate for BBN model presented on Fig. 2: the probability of the last episode was set to 0.125, the remaining part was distributed uniformly according to the length of the interval.

We considered both individual predictions and prediction for the whole group. The predicted rate value for each case or “respondent” was the interval with maximum posterior probability. Table 1 shows the confusion matrix for the individual predictions. The average accuracy [23] for this eight-class classification was 87.7 %. Moreover, the false positive predictions in most cases belonged to adjacent intervals.

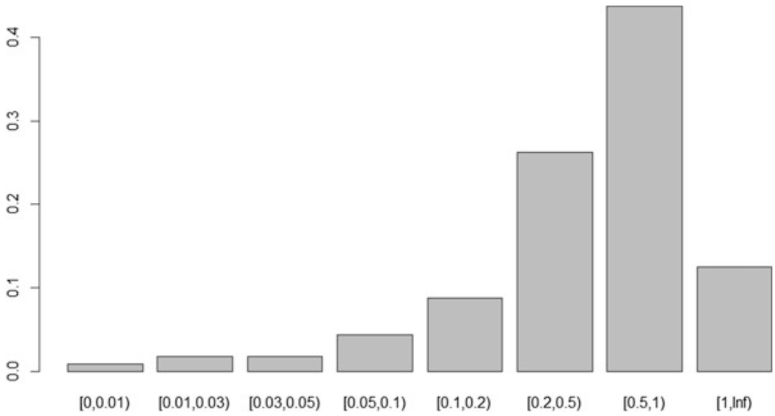

Fig. 2 Prior rate distribution

Table 1 Confusion matrix for the individual predictions

Initial value	Predicted value							
	$\lambda^{(1)}$	$\lambda^{(2)}$	$\lambda^{(3)}$	$\lambda^{(4)}$	$\lambda^{(5)}$	$\lambda^{(6)}$	$\lambda^{(7)}$	$\lambda^{(8)}$
$\lambda^{(1)}$	6	12	3	1	0	0	0	0
$\lambda^{(2)}$	83	310	85	36	1	0	0	0
$\lambda^{(3)}$	11	205	145	123	6	0	0	0
$\lambda^{(4)}$	0	121	249	470	153	7	0	0
$\lambda^{(5)}$	0	3	18	252	568	119	0	0
$\lambda^{(6)}$	0	0	0	6	285	369	0	0
$\lambda^{(7)}$	0	0	0	0	0	20	0	0
$\lambda^{(8)}$	0	0	0	0	0	0	0	0

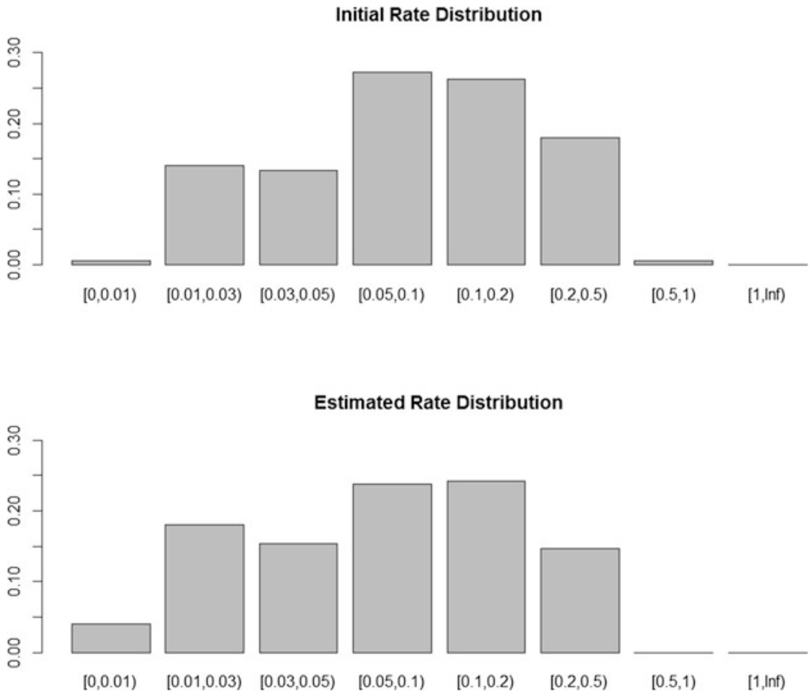

Fig. 3 Comparison of the initial and the estimated rate distributions

The initial rate distribution for generated dataset and comparison with the rate distribution estimated on the base of data about episodes is presented on Fig. 3. The estimated rate distribution was the result of inference on the BBN combined for the group of "respondents" according to Dirichlet distribution [24]. The χ^2 test for comparison the distributions showed that we did not reject the hypothesis about similarity of distributions ($\chi^2 = 48$, $df = 42$, $p-$value $= 0.24$). In other words, the estimated rate distribution looked similar to the initial distribution with major part of cases concentrated between 0.05 and 0.2 cases per day.

2.4 Real Data Example

To illustrate how the model works for real data we considered data about the last episodes of alcohol consumption and data about minimum and maximum intervals between the episodes during the last 6 months. The data about 380 respondents was collected in 2011 among patient of one of STD clinics in St.Petersburg, Russia. We used the same prior rate distribution that was described in Sect. 2.3 (Fig. 2). The estimated distribution presented on Fig. 4.

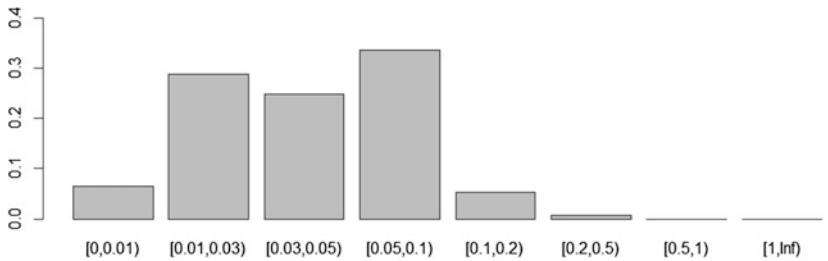

Fig. 4 Estimated rate distributions for alcohol consumption

According to the estimates we can conclude that with probability greater than 0.9 respondents in the sample drank alcohol rarely than one tome per 10 days. At the same time, the probability of really rare consumption was relatively low too that allowed to conclude that in general there was a need for behavior-changing events.

3 Conclusion

The area of risky behavior modelling has unsolved issues in current practice. To provide a tool for estimating behavior characteristics we proposed the model in terms of Bayesian Belief Networks that combined both empirical data and expert knowledge about behavior. Testing the model on automatically generated dataset showed good results. The rate estimates for real data illustrated the possible conclusions in practical issues. However, for practical usage the model requires more detailed testing with different initial rate distributions and, on the other hand, testing on data about real behavior with available data about episodes and real rate.

Acknowledgments This work was partially supported by the by RFBR according to the research projects No. 14-01-00580 and No. 16-31-60063.

References

1. Pearl, J.: Causality: Models, Reasoning, and Inference. Cambridge University Press, Cambridge (2000)
2. Du, Y., Guo, Y.: Evidence reasoning method for constructing conditional probability tables in a Bayesian network of multimorbidity. Technol. Health Care, 23(s1) (2015)
3. Jitwasinkul, B., Hadikusumo, B.H., Memon, A.Q.: A Bayesian belief network model of organizational factors for improving safe work behaviors in Thai construction industry. Saf. Sci. **82**, 264–273 (2016)
4. Trucco, P., Cagno, E., Ruggeri, F., Grande, O.: A Bayesian belief network modelling of organisational factors in risk analysis: a case study in maritime transportation. Reliab. Eng. Sys. Saf. **93**(6), 845–856 (2008)

5. Semakula, H.M., Song, G., Achuu, S.P., Zhang, S.: A Bayesian belief network modelling of household factors influencing the risk of malaria: a study of parasitaemia in children under five years of age in sub-Saharan Africa. Environ. Model Softw. **75**, 59–67 (2016)
6. Sun, L., Erath, A.: A Bayesian network approach for population synthesis. Transp. Res. Part C: Emerg. Technol. **61**, 49–62 (2015)
7. Neapolitan, R.E.: Learning Bayesian Networks. Pearson Prentice Hall (2003)
8. Tulupyev, A., Nikolenko, S., Sirotkin, A.: Bayesian Networks: A Probabilistic Logic Approach. Nauka, SPb (2006)
9. Mkrtchyan, L., Podofillini, L., Dang, V.N.: Bayesian belief networks for human reliability analysis: a review of applications and gaps. Reliab. Eng. Syst. Saf. **139**, 1–16 (2015)
10. Leigh, B.C., Stall, R.: Substance use and risky sexual behavior for exposure to HIV: issues in methodology, interpretation, and prevention. Am. Psychol. **48**(10), 1035 (1993)
11. Bolger, N., Davis, A., Rafaeli, E.: Diary methods: capturing life as it is lived. Annu. Rev. Psychol. **54**(1), 579–616 (2003)
12. Schroder, K.E., Carey, M.P., Vanable, P.A.: Methodological challenges in research on sexual risk behavior: II. Accuracy of self-reports. Ann. Behav. Med. **26**(2), 104–123 (2003)
13. Graham, C.A., Catania, J.A., Brand, R., Duong, T., Canchola, J.A.: Recalling sexual behavior: a methodological analysis of memory recall bias via interview using the diary as the gold standard. J. Sex Res. **40**(4), 325–332 (2003)
14. Tulupyeva, T., Paschenko, A., Tulupyev, A., Krasnoselskikh, T., Kazakova, O.: HIV Risky Behavior Models in The Context of Psychological Defense and Other Adaptive Styles. Nauka, SPb (2008)
15. Ramrakha, S., Caspi, A., Dickson, N., Moffitt, T.E., Paul, C.: Psychiatric disorders and risky sexual behaviour in young adulthood: cross sectional study in birth cohort. BMJ **321**(7256), 263–266 (2000)
16. Lemelin, C., Lussier, Y., Sabourin, S., Brassard, A., Naud, C.: Risky sexual behaviours: the role of substance use, psychopathic traits, and attachment insecurity among adolescents and young adults in Quebec. Can. J. Hum. Sex. **23**(3), 189–199 (2014)
17. Vacirca, M.F., Ortega, E., Rabaglietti, E., Ciairano, S.: Sex as a developmental transition: the direct and indirect roles of peers. Psychol. Sex. **3**(2), 108–122 (2012)
18. Suvorova, A.: Socially significant behavior modeling on the base of super-short incomplete set of observations. Inf. Measur. Control Syst. **9**(11), 34–38 (2013)
19. Stepanov, D.V., Musina, V.F., Suvorova, A.V., Tulupyev, A.L., Sirotkin, A.V., Tulupyeva, T. V.: Risky behavior Poisson model identification: heterogeneous arguments in likelihood. Trudy SPIIRAN **23**, 157–184 (2012)
20. GeNIe& SMILE: Decisions Systems Laboratory. School of Information Sciences. University of Pittsburg. http://genie.sis.pitt.edu/
21. AgenaRisk Bayesian Network Tool. http://www.agenarisk.com
22. Core Team. *R:* A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2015). http://www.R-project.org/
23. Sokolova, M., Lapalme, G.: A systematic analysis of performance measures for classification tasks. Inf. Process. Manage. **45**(4), 427–437 (2009)
24. Suvorova, A.V., Tulupyev, A.L., Sirotkin, A.V.: Bayesian belief networks for risky behavior rate estimates. Nechetkie sistemy i myagkie vychisleniya (Fuzzy Syst. Soft Comput.) **9**(2), 115–129 (2014)