Atlantis Studies in Uncertainty Modelling, volume 1
11th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT 2019)

Prediction quality of Bayesian belief network model for risky behavior: comparison of subsamples with different rates

Alena Suvorova[a] **and Alexander Tulupyev**[b]
[a]National Research University Higher School of Economics, St. Petersburg, Russia, suvalv@gmail.com
[b] St.Petersburg Institute for Informatics and Automation of RAS, St. Petersburg, Russia, alexander.tulupyev@gmail.com

Abstract

The study investigates the proposed approach for behavior modeling on the base of Bayesian belief networks that allows predicting behavior characteristics using small and incomplete data from surveys about behavior episodes. We explored the prediction quality of the models in case of rare behavior. The test dataset was automatically generated and included 24465 cases. During the experiment, we considered cases with different rates to compare prediction quality. Our findings suggest that the model had a good prediction quality especially for rare and frequent behaviors (about 92% accuracy) and lower measures for medium-rate behaviors (about 86% accuracy).

Keywords: Bayesian Belief Network, Machine learning, Behavior models, Risky behavior.

1 Introduction

Knowing behavioral characteristics allows you to make informed decisions, make predictions, plan costs, and monitor changes [3, 10]. One example is the modeling of risk behavior and the behavior parameter estimates including the assessment of the prevention program effectiveness [4]. The task of risk assessment in the areas of public health, epidemiology, cybersecurity includes numerical estimates of the behavior parameters [19]. For example, the feeding behavior of pregnant women can influence the development of the fetus [7]. When observing pregnancy, a specialist may require a quick assessment of the frequency of use of certain foods in order to appoint additional examinations or give dietary recommendations. An important feature in this case is the difficulty of recalling everyday actions [6].

Other examples of risk-taking behavior include the use of alcohol and other drugs (including injection drug use), unprotected sex with a random partner [19]. This behavior is an important object of study in the fields of epidemiology and public health, because it is associated with increased risks of mortality and disability and leads to the spread of diseases. A special feature is the individual's desire to give a socially expected answer to questions about this behavior [2].

To make predictions we have to combine expert knowledge and data from different sources (usually incomplete). Earlier studies proposed an approach for risky behavior modelling based on Bayesian belief network with data about several last behavior episodes [17].

Bayesian Belief Network (BBN) is a type of probabilistic graphical models that represents a set of random variables and their conditional dependencies [8]. It consists of two components: structure and parameters. A network structure is presented in the form of a directed acyclic graph where nodes correspond to the random variables and directed edges represent dependencies among variables. Parameters are represented as a set of conditional probability distributions, one for each variable, characterizing the dependencies represented by the edges [1].

The proposed model showed good prediction quality on the data about behavior on social networking site [16], the data about the last episodes are able to describe the behavior. However, the proposed model can be good for estimating frequent behavior and fail for rare behaviors or vice versa still being good on average.

The purpose of the paper is to compare the performance of the proposed model for different behavior rates to explore if there are any limitations for applying it in rare behaviors studies.

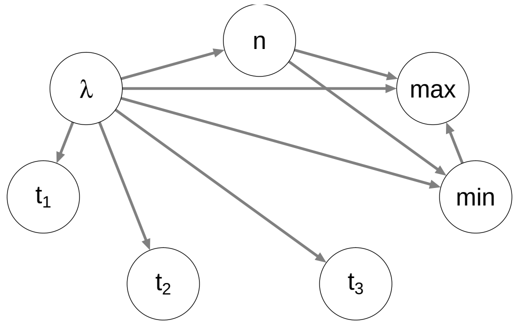

Figure 1: Expert structure for risky behavior model

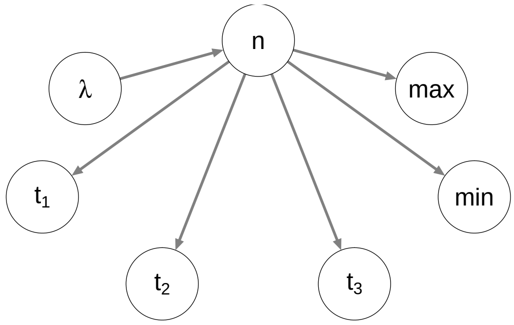

Figure 2: Simplified structure for risky behavior model

2 Model description

The input data for the model [17] include the lengths of intervals between three last episodes of risky behavior and the lengths of minimum and maximum intervals between episodes during a given period of interest. The data about episodes in most applications is obtained from respondents self-reports [18]. In addition, the model includes the latent variable n that corresponds to the number of episodes during the period of interest and the behavior rate, that is the key variable, the one we want to estimate. We assume that for each respondent occurrence of episodes follows Poisson random process.

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 [14] 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.

On the contrary, Bayesian belief networks allow determining complex relationships in terms of simpler dependencies. Modelling risky behaviour 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 [8].

Figure 1 represented the structure of BBN model proposed in [17]. The model structure is directed acyclic graph $G(V, L)$ with vertices V and edges $L = \{(u, v) : u, v \in V\}$, where $V = \{t_1, t_2, t_3, t_{\min}, t_{\max}, \lambda, n\}$ and λ is random variable for the behavior rate; t_i are random variables for the lengths of the i-th last interval between episodes; $t_{\min}$ and $t_{\max}$ are random variables for the length of minimum and maximum intervals; n is random variable for the number of episodes during the period of interest.

Further studies [18] explored other structures including those that were data-based and used structure-learning algorithms. The Hill-Climbing algorithm and other score-based methods in many experiments produced a simplified structure close to naive Bayes classifier (Figure 2). In this structure the behavior rate is related to number of episodes and all other variables depended on the number of episodes only. This structure has simple interpretation because the number of episodes can be directly calculated on the base of the rate. The simplified model showed the good prediction quality compatible to the quality of initial model [15], so we decided to include this structure to the experiments too.

For all further examples we discretized variables using the following intervals: $\lambda^{(1)} = [0; 0.05)$, $\lambda^{(2)} = [0.05; 0.1)$, $\lambda^{(3)} = [0.1; 0.15)$, $\lambda^{(4)} = [0.15; 0.2)$, $\lambda^{(5)} = [0.2; 0.3)$, $\lambda^{(6)} = [0.3; 0.5)$, $\lambda^{(7)} = [0.5; 1)$, $\lambda^{(8)} = [1; +\infty)$ for the rate variable; t_i, $i = 1, 2, 3$, $t_{\min}$ $t_{\max}$ $t^{(1)} = [0; 0.5)$, $t^{(2)} = [0.5; 1)$, $t^{(3)} = [1; 3)$, $t^{(4)} = [3; 7)$, $t^{(5)} = [7; 14)$, $t^{(6)} = [14; 30)$, $t^{(7)} = [30; 180)$, $t^{(8)} = [180; +\infty)$ for the intervals between episodes.

The discretization breaks were chosen by experts. The general idea was to make smaller intervals for more frequent values and use more interpretable breaks. For example, $\lambda \in [1; +\infty)$ means that there was at least one behavior episode per day. Furthermore, interpretability becomes more important for variables corresponded to interval lengths: $t_1 \in [14; 30)$ means that the last behavior episode was 2-4 weeks ago. The usage of weeks-months-year notation allows simplifying the questionnaire design for real-world applications. The study [15] showed that such kind of discretization worked better than equal-length intervals and was close in model performance metrics to quantile breaks those usually are unavailable for real data.

3 Methods

3.1 Dataset

Since collecting the real data about behavior episodes is extremely time-consuming especially for risky behaviors, we explored the models using automatically generated dataset. The generation process assumes that the behaviour 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 real-life risky behavior and it is widely used in previous studies [13]. The use of the automatically generated dataset provides an important advantage: we can compare the theoretical (the a priori given rate) and the estimated rate, while survey data usually do not have real rates or they are extremely biased [2]

On the first step we generated 1000 values for behavior rate following Gamma distribution (shape $\theta = 1.1$ scale $k = 0.3$). The choice of parameters value was aimed to get dataset that corresponded to real-life risky behavior where most of the respondents have less than one episode per day (96.7% of cases in our dataset) and on average an episode happens once in 34 days (mean rate value was 0.31).

Next step included generation of 25 respondents (sequences of behavior episodes) for each rate value for period of 365 days in total, that summed up to 25000 sequences of episodes. Then we calculated initial data for the model: 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 365 days) the final dataset included 24465 cases (or respondents).

To estimate prediction quality of the models we randomly select 10000 cases to the test dataset and do not use them in model learning.

3.2 Experiment design

We splitted the test dataset into three subsets: low rates ($\lambda \in [0; 0.1)$), high rates ($\lambda \in [0.5; +\infty)$), and medium rates ($\lambda \in [0.1; 0.5)$). On each iteration we randomly selected 5000 cases from our train dataset (14465 cases). Then we learned the parameters of Bayesian belief network for both structures described in Section 2: 1) the initial model, 2) the simplified model with fewer links. Next, we tested our models on three subsets with different rates and estimated prediction quality (accuracy, precision, recall, F1 score). All measures were calculated according to multi-class classification metrics (average accuracy, macro precision, macro recall, macro F1 score) [12], measures were averaged among subsets to take into account different number of classes in subsets. To eliminate the influence of particular training dataset we repeated the experiment 50 times. The calculations and statistical analysis were performed using R [9]; in particular, for Bayesian network analysis we used bnlearn [11] and gRain [5] packages.

4 Results

The average accuracy through 50 iterations is shown on Figure 3 (the results on the whole test sample are in bold). The both models produced the highest accuracy measures for high-rate subsample, a bit lower results were in low-rate subsample while the middle part showed the worst results comparing to others. Note that the gap between medium-rate and low- and high-rate subsamples was larger for the simplified model: the difference in average accuracy was about 10% between medium-rate and high-rate subsamples, while the same difference for the initial model was about 8%. In other words, the simplified model worked better for low- and high-rate behaviors but worse for medium-rate behavior compared to the model with initial structure.

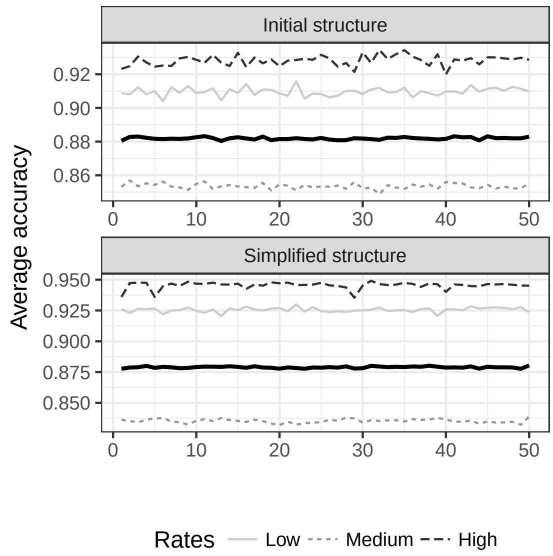

Figure 3: Average accuracy measures

The medium-rate subsample also had lower F1 scores compared to low- and high-rate subsamples (Figure 4).

Mean quality measures and the standard deviations (SD) for both model structures and various rate subsamples are summarized in Table 1. To compare all quality measures among the different behavior rates we run pairwise t-test for multiple comparisons with Bonferroni correction. All differences were statistically significant (p-value<0.01).

Thus, the extreme (low and high) rates were predicted better in our experiment according to all measures. This difference can be caused by discretization if it is

Subsample	Accuracy (SD)	Avg. accuracy (SD)	Precision (SD)	Recall (SD)	F1 (SD)
Expert structure					
High rates	0.71 (0.01)	0.93 (0.003)	0.31 (0.03)	0.66 (0.02)	0.42 (0.03)
Low rates	0.64 (0.01)	0.91 (0.002)	0.3 (0.01)	0.64 (0.009)	0.41 (0.01)
Medium rates	0.41 (0.006)	0.85 (0.002)	0.28 (0.005)	0.4 (0.006)	0.33 (0.005)
Total sample	0.53 (0.003)	0.88 (0.000)	0.53 (0.004)	0.53 (0.004)	0.53 (0.003)
Simplified structure					
High rates	0.78 (0.01)	0.95 (0.003)	0.31 (0.02)	0.54 (0.02)	0.39 (0.01)
Low rates	0.7 (0.007)	0.93 (0.002)	0.31 (0.01)	0.7 (0.007)	0.43 (0.01)
Medium rates	0.34 (0.006)	0.84 (0.002)	0.28 (0.007)	0.33 (0.007)	0.3 (0.006)
Total sample	0.52 (0.003)	0.88 (7e-04)	0.53 (0.01)	0.48 (0.006)	0.5 (0.004)

Table 1: Mean prediction quality measures

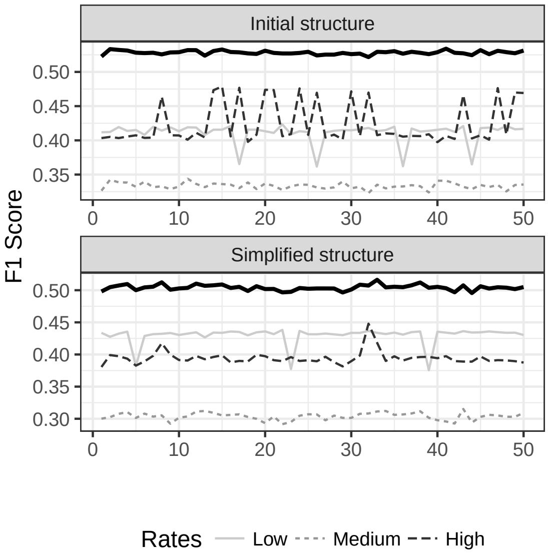

Figure 4: F1 scores

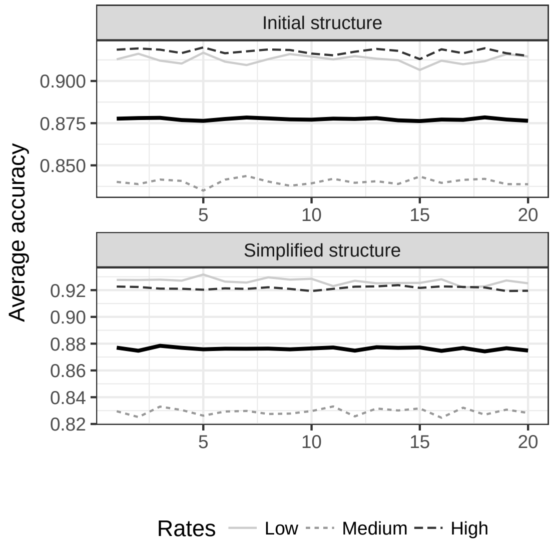

Figure 5: Average accuracy for model with quantile discretization

harder to distinguish between intervals in the "middle part" of the value range. At the same time, the low-rate intervals are also very close to each other. To verify this suggestion we repeated the experiment with quantile breaks for discretization, i.e using more information about initial rates (that is unknown when we explore real survey data). The new experiment showed the same pattern, all measures were higher for extreme rates and lower for medium rates (see Figure 5 as example)

5 Conclusion

The current study continues exploring the behavior model on the base of Bayesian belief networks that allows processing incomplete data about behavior episodes and predicting behavior characteristics.

Our initial suggestion that the model can fail for estimating rare behaviors was not confirmed in the experiment. Moreover, low-rate subsample had significantly higher results comparing to medium-rate one. The only limitation of estimating rare behaviors is extremely rare behavior: if there was only one episode during the period of interest the model does not have enough data to make prediction.

The model was trained on the dataset that followed a set of assumptions including the distribution of rates in population and better discretization intervals. If we explore more frequent behavior, using new training dataset for learning model parameters allows accounting for behavior characteristics. Additional knowledge about distributions, theoretical assumptions about behavior can lead to higher prediction quality.

In general, proposed model showed high prediction quality and has a great potential for analyzing real-life behavior problems.

Acknowledgement

This work was partially supported by the by RFBR according to the research project No. 18-01-00626 and Governmental contract (SPIIRAS) No. 0073-2019-0003

References

[1] A. Darwiche, Modelling and reasoning with Bayesian networks, Cambridge University Press, Cambridge, 2009.

[2] C. Davis, J. Thake, N. Vilhena, Social desirability biases in self-reported alcohol consumption and harms, Addictive behaviors 35 (4) (2010) 302–311.

[3] A. Do, E. Rosenberg, P. Sullivan, L. Beer, T. Strine, J. Schulden, J. Fagan, M. Freedman, J. Skarbinski, Excess burden of depression among hiv-infected persons receiving medical care in the united states: data from the medical monitoring project and the behavioral risk factor surveillance system, PloS one 9 (3) (2014) e92842.

[4] M. Fishbein, W. Pequegnat, Evaluating aids prevention interventions using behavioral and biological outcome measures, Sexually transmitted diseases 27 (2) (2000) 101–110.

[5] S. Højsgaard, Graphical independence networks with the gRain package for R, Journal of Statistical Software 46 (10) (2012) 1–26.
URL `http://www.jstatsoft.org/v46/i10/`

[6] J.-S.Shim, K. Oh, H. Kim, Dietary assessment methods in epidemiologic studies, Epidemiology and health 36.

[7] M. Mulligan, S. Felton, A. Riek, C. Bernal-Mizrachi, Implications of vitamin d deficiency in pregnancy and lactation, American journal of obstetrics and gynecology 202 (5) (2010) 429–e1.

[8] J. Pearl, Causality: Models, Reasoning, and Inference, Cambridge University Press, Cambridge, 2000.

[9] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria (2017).
URL `https://www.R-project.org/`

[10] J. Sallis, N. Owen, M. Fotheringham, Behavioral epidemiology: a systematic framework to classify phases of research on health promotion and disease prevention, Annals of Behavioral Medicine 22 (4) (2000) 294–298.

[11] M. Scutari, J.-B. Denis, Bayesian networks: with examples in R, Chapman and Hall/CRC, 2014.

[12] M. Sokolova, G. Lapalme, A systematic analysis of performance measures for classification tasks, Information Processing & Management 45 (4) (2009) 427–437.

[13] D. Spiegelman, E. Hertzmark, Easy sas calculations for risk or prevalence ratios and differences, American journal of epidemiology 162 (3) (2005) 199–200.

[14] D. Stepanov, V. Musina, A. Suvorova, A. Tulupyev, A. Sirotkin, T. Tulupyeva, Risky behavior poisson model identification: heterogeneous arguments in likelihood, Trudy SPIIRAN 23 (2012) 157–184.

[15] A. Suvorova, Exploring bayesian belief network for risky behavior modelling: discretization and latent variables, in: The Second International Scientific and Practical Conference "Fuzzy Technologies in the Industry FTI 2018". CEUR Workshop Proceedings, 2018, pp. 63–70.

[16] A. Suvorova, A. Tulupyev, Evaluation of the model for individual behavior rate estimate: Social network data, in: Soft Computing and Measurements (SCM), 2016 XIX IEEE International Conference, 2016, pp. 18–20.

[17] A. Suvorova, A. Tulupyev, A. Sirotkin, Bayesian belief networks for risky behavior rate estimates, Nechetkie sistemy i myagkie vychisleniya [Fuzzy Systems and Soft Computing] (9 (2)) (2014) 115–129.

[18] A. Suvorova, T. Tulupyeva, Bayesian belief networks in risky behavior modelling, in: Proceedings of the First International Scientific Conference Intelligent Information Technologies for Industry (IITI16), 2016, pp. 95–102.

[19] T. Tulupyeva, A. Paschenko, A. Tulupyev, T. Krasnoselskikh, O. Kazakova, HIV risky behavior models in the context of psychological defense and other adaptive styles, Nauka, SPb, 2008.