Empirical approach to assessing sensitivity of local posteriori inference of algebraic Bayesian network

Arseniy Zavalishin[a,b] and Anatoliy Maksimov[b]

[a] *Saint Petersburg State University, 11 Lieutenant Shmidt Emb., St Petersburg 199034, Russia*
[b] *St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), 14th line. V.I., 39, St. Petersburg, 199178, Russia*

Abstract
The article presents the results of the study of the upper assessment of the sensitivity of solving the second problem of local posterior inference in algebraic Bayesian networks to variations in probability estimates in the knowledge pattern. The result is the top estimate found. The theoretical and practical significance of the study is to determine the permissible error in the data when working with algebraic Bayesian networks. Note that this task is considered for the first time.

Keywords 1
Knowledge pattern, sensitivity assessment, algebraic Bayesian networks, empirical approach, local inference.

1. Introduction

One of the problems solved in the framework of studying various models of machine learning, in particular, belonging to the class of probabilistic graphic models, is the problem of studying the sensitivity of the model to variations in input data [1]. The reasons for the special attention to this task lie in the means and time that must be spent on obtaining the information on the basis of which the model is formed. For example, often its production is associated with the purchase and development of more high-tech equipment, conducting a large number of experiments, interviewing experts, etc. [2]. All this, as a rule, leads to the consumption of available resources. Given the limitations of these resources, the sensitivity of the model can play a large role in planning the budget and allocating time for the implementation of its stages [1]. Knowing the sensitivity of the model to input data, it is possible to estimate in advance the required accuracy of the obtained data and allocate the necessary resources for it [1]. Therefore, the task of studying sensitivity is very important in the study of models working with data [1]. Some of these models are algebraic Bayesian networks belonging to the class of logic-probabilistic graphic models [2, 3, 4, 5, 6]. They are a non-directional graph, at the vertices of which are knowledge patterns [2]. The mathematical model of the latter is the ideal of conjuncts, each of which is assigned some estimate of the probability of their truth, and the estimates can be both scalar and interval [2, 7]. Interval estimates are an advantage of algebraic Bayesian networks, since they allow you to work with data that may be inaccurate, non-numerical or incomplete [12].

Due to various reasons, information about the subject area in the process of working with algebraic Bayesian networks will change. The data come into the knowledge pattern in the form of a certificate with updated estimates of the probability of truth of conjuncts [2, 6, 8]. After that, estimates of the probabilities of the truth of conjuncts in knowledge pattern are updated by solving the second problem of local posterior inference [2]. Often changing information can, for example, arise when studying socioengineering attacks when work is done with people [9, 10, 11, 12].

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: adz@dscs.pro; agm@dscs.pro
ORCID: 0000-0003-2756-460X; 0000-0002-7542-3302

As part of the study of the model of algebraic Bayesian networks, the task of studying the sensitivity of the second set local posteriori inference to variations in estimates of the probability of truth of conjuncts in a knowledge pattern appears. In other words, the effect of probability estimates errors in the knowledge pattern on the result of updating the probability estimates of the truth of the conjuncts included in it is investigated. The solution of this problem will allow, if inaccuracies in the estimates in the knowledge pattern are detected, to predict in advance the errors in the estimates obtained after the solution of the second posterior output problem, and, depending on the result of forecasting and the tasks set, to decide on the allocation or non-allocation of additional resources for the elimination of errors. In the present work, the upper estimate of the norm of the difference of the vectors of the estimates of the probability of the truth of the conjuncts of the final knowledge patterns after solving the second problem of local posterior inference for various knowledge patterns is investigated.

The purpose of the work is to find empirical upper estimates of the sensitivity of the second task of posteriori inference in a knowledge pattern to errors in estimates of the probabilities of the truth of conjuncts in the initial knowledge pattern. Obtaining empirical estimates allows now to assess the permissible error in estimates of conjuncts in a knowledge pattern, as well as to make assumptions about theoretical estimates of sensitivity.

2. Assessments in the Knowledge Pattern

Algebraic Bayesian networks, as mentioned earlier, are a non-directional graph, at the vertices of which are knowledge patterns [2, 6]. We formulate the definitions and approvals necessary to describe the work done.

Definition. The alphabet is called the set of atomic propositional formulas [13].

Definition. Conjunct is the conjunction of a number of atomic variables [13].

Definition. The literal $\tilde{x}$ means that in its place in the formula can be either x, or $\bar{x}$ [13].

Definition. A quant over an alphabet C is a conjunct that for any atomic formula of the alphabet contains either this formula or its negation.

Definition. The mathematical model of the knowledge pattern is the ideal of conjuncts built over the alphabet [2, 6]. Each conjunct is compared with an estimate of the probability of the truth of this conjunct, which can take both scalar and interval values [2, 6].

Thus, the knowledge pattern is written in the form of a pair $\langle C, P_c \rangle$, where C={c_1, c_2, … , c_m} is the alphabet over which the knowledge pattern is built, P_c and the vector of estimates of the probabilities of the truth of conjuncts in the knowledge pattern [6]:

$$P_c=(p(\omega), p(c_1),\dots, p(c_m), p(c_1c_2), \dots , p(c_1c_2\dots c_m))^T.$$

Definition. To determine the consistent knowledge pattern, we define the matrix **I** [13]:

$$I = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}.$$

By defining a matrix I, we can define a matrix $I_n=I^{[n]}$ where the degree means the Kronecker degree, wherein the matrix I_n is reversible [13].

Definition. A knowledge pattern $\langle C, P_c \rangle$ with scalar estimates is called consistent if it satisfies the condition [13]:

$$I_n \times P_c \geq \langle 0 \rangle^{2^n},$$

where $\langle 0 \rangle^{2^n}$ is a height 2^n vector consisting of 0 [13].

Definition. Another way to represent a consistent knowledge pattern is a pair $\langle C, P_c \rangle$, where C={c_1, c_2, … , c_m} is the alphabet over which the knowledge pattern is built, and P_q the vector of probability estimates of all quanta over the alphabet C [13]:

$$P_q = \begin{pmatrix} p(c_1c_2\dots c_n) \\ p(c_1c_2\dots \bar{c}_n) \\ \dots \\ p(\bar{c}_1\bar{c}_2\dots \bar{c}_n) \end{pmatrix}.$$

Definition. The vector P_q is constrained by [13]:

$$(\forall k \in 1..2^n) P_q[k] \geq 0, \quad (1)$$

$$\sum_{k=1}^{2^n} P_q[k] = 1. \quad (2)$$

Statement. Estimates of quant probabilities can be expressed through estimates of conjunct probabilities as follows [13]:

$$I_n \times P_c = P_q.$$

Remark. Since the matrix I_n is reversible, we can express conjunct estimates through quant estimates multiplying both parts on the left by I_n^{-1}.

Having formulated the necessary statements and definitions, you can go to setting the task.

2.1. Problem statement

Consider the setting of the problem to be solved.

Let be given two consistent knowledge patterns $\langle C, P_c^{(1)} \rangle$ and $\langle C, P_c^{(2)} \rangle$ over the alphabet C={c_1, c_2, … , c_m}:

$$P_c^{(1)}=(p_1(\omega), p_1(c_1), \ldots, p_1(c_m), p_1(c_1c_2), \ldots, p_1(c_1c_2\ldots c_m))^T,$$
$$P_c^{(2)}=(p_2(\omega), p_2(c_1), \ldots, p_2(c_m), p(c_1c_2), \ldots, p_2(c_1c_2\ldots c_m))^T,$$

There is also a certificate above the alphabet $X=\{x_1, x_2, \ldots, x_n\}$, such that $X \subset C$:

$$E_V = \begin{pmatrix} p(x_1x_2 \ldots x_n) \\ p(x_1x_2 \ldots \bar{x}_n) \\ \ldots \\ p(\bar{x}_1\bar{x}_2 \ldots \bar{x}_n) \end{pmatrix}.$$

which are subject to the following conditions [13]:

$$(\forall k \in 1..2^n) E_{V_i}[k] \geq 0, \quad (1)$$

$$\sum_{k=1}^{2^n} E_{V_i}[k] = 1. \quad (2)$$

The problem consists in finding the upper estimate of the norm of the difference of the vectors of estimates of the probability of truth of conjuncts of knowledge patterns obtained from the initial ones by solving the second problem of posteriori inference after receiving a given certificate.

3. Finding empirical estimates

A common method used for primary sensitivity studies of models is to find empirical sensitivity estimates by starting the model multiple times with a change in input data [1]. The results of the model launch are then compared to each other and the results of the comparisons are analyzed [1].

To find empirical estimates of the sensitivity of the second problem of local posterior inference, a consistent knowledge pattern $\langle C, P_c^{(0)} \rangle$ was fixed over the alphabet C={c_1, c_2, c_3}, such that $P_c^{(0)} = (0.125, 0.125, \ldots, 0.125)$. This knowledge pattern $I_3 \times P_c^{(0)} = P_q^{(0)}$ was chosen so that the truth of all quanta over the alphabet C was equal. Then a certificate was recorded Ev:

$$E_V = \begin{pmatrix} p(c_1c_2) \\ p(c_1\bar{c}_2) \\ p(\bar{c}_1c_2) \\ p(\bar{c}_1\bar{c}_2) \end{pmatrix} = \begin{pmatrix} 0.30 \\ 0.15 \\ 0.23 \\ 0.32 \end{pmatrix}.$$

$P_c'^{(0)}$ result of solving the second problem of local posterior output for selected knowledge patterns and evidence.

Then, for comparison, a set of *m=100000* consistent knowledge patterns $i = 1^{m\{P_c^{(i)}\}}$ was generated, for each of which there was also a second local posterior inference task. The result was a set $i = 1^{m\{P_c'^{(i)}\}}$.

For each knowledge pattern from the resulting sets, the following values were found:

$$\left\|P_C^{(0)} - P_C^{(i)}\right\|_2 \qquad (dif_x)$$

$$\left\|P_C'^{(0)} - P_C'^{(i)}\right\|_2 \qquad (dif_p)$$

Figure 1 shows the plotting of the relationship between dif_x and dif_p.

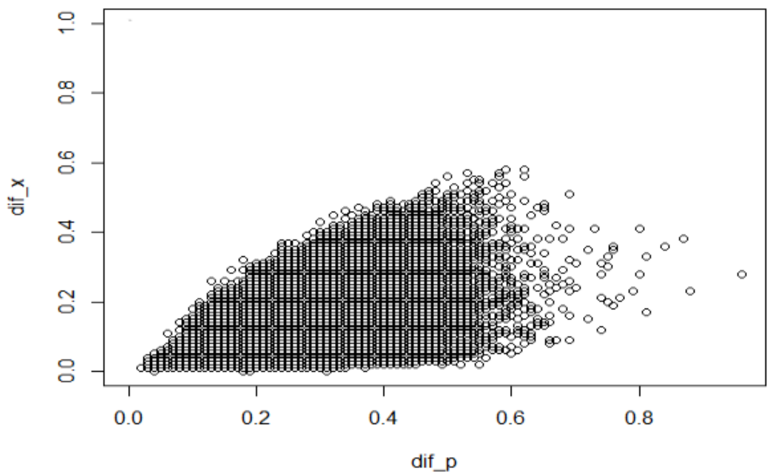

Figure 1: Results of the experiment.

In addition, the maximum ratio was found:

$$\max_i \left[\frac{\left\|P_C^{(0)} - P_C^{(i)}\right\|_2}{\left\|P_C'^{(0)} - P_C'^{(i)}\right\|_2} \right] = 1.9.$$

As a result, it can be seen that the variation of probability estimates in the knowledge pattern after solving the second problem of local posterior inference is no more than 2 times the variation of probability estimates in the original knowledge pattern.

4. Conclusion

In the work, an experiment was proposed and conducted to find empirical estimates of the sensitivity of the second problem of local posteriori inference to variation in estimates of the probabilities of conjuncts in a knowledge pattern. The obtained estimates suggest that the variation of probability estimates in the knowledge pattern after solving the second problem of local posterior inference is no more than 2 times the variation of probability estimates in the original knowledge pattern. This result will make it possible now to assess errors in the estimates of knowledge patterns after solving the second problem of local posterior inference, and can also set the direction of studying theoretical evaluation.

5. Acknowledgement

Work is performed within the project on the SPIIRAS state task No.0073-2019-0003 and with financial support of the Russian Federal Property Fund (grant No.18-01-00626)

6. References

[1] C. Bishop, Pattern Recognition and Machine Learning, Springer-Verlag New York, 2006.

[2] A. L. Tulupyev, S. I. Nikolenko, A. V. Sirotkin, Fundamentals of the theory of Bayesian networks: textbook, St. Petersburg: Publishing House S.-Peterb. Un, 2019.

[3] N. A. Kharitonov, A. G. Maximov, A. L. Tulupyev, Algebraic Bayesian networks: the use of parallel computing while maintaining various degrees of consistency, International Conference on Information Technologies. Springer, Cham, 2019.

[4] D. G. Levenets, M. A. Zotov, A. V. Romanov, A. L. Tulupyev, A. A. Zolotin, A. A. Filchenkov, Decremental and incremental reshaping of algebraic Bayesian networks global structures , Proceedings of the First International Scientific Conference “Intelligent Information Technologies for Industry”(IITI’16). Springer, Cham, 2016.

[5] T. D. Nielsen, F. V. Jensen, Bayesian networks and decision graphs, Springer Science Business Media. (2009)

[6] A. L. Tulupyev, D. M. Stolyarov, M. V. Mentyukov, Representation of the local and global structure of the algebraic Bayesian network in Java applications, Proceedings of SPIIRAN, 5 (2007).

[7] N. J. Nilsson, Probabilistic logic , Artificial intelligence, 28, 1 (1986).

[8] V. V. Oparin, A. A. Filchenkov, A. V. Sirotkin, A. L. Tulupyev, Matroid representation of a family of adjacent graphs over a set of knowledge pattern, Scientific and technical bulletin of information technologies, mechanics and optics, 4(68) (2010).

[9] M. V. Abramov, Automation of analysis of social networks to assess the security against socioengineering attacks, Automation of management processes, 1(51) (2018).

[10] A. Khlobystova, M. Abramov, A. Tulupyev, An approach to estimating of criticality of social engineering attacks traces , Studies in Systems, Decision and Control, 199, 2019.

[11] A. A. Korepanova, V. D. Oliseenko, M. V. Abramov, A. L. Tulupyev, Application of machine learning methods in the task of identifying user accounts in two social networks, Computer tools in education, 3 (2019). doi:10.32603/2071-2340-2019-3-29‒43 (In Russian)

[12] A. Suleimanov, A. Abramov, A. Tulupyev, Modelling of the social engineering attacks based on social graph of employees communications analysis, 2018 IEEE Industrial Cyber-Physical Systems (ICPS). IEEE, 2018. Springer Science Business Media, 2009.

[13] A. L. Tulupyev, Algebraic Bayesian networks: local logic-probabilistic conclusion. Tutorial, Ed. 2nd, rework. and additional. St. Petersburg: VBM, 2019.