Local Parameter Training of Algebraic Bayesian Networks: Conjugate Distributions and Expert Knowledge With Uncertainty

Nikita A. Kharitonov[a], Tulupyev Alexander[a,b]

[a]St. Petersburg State University, St. Petersburg, Russia
[b]Laboratory of Theoretical and Interdisciplinary Problems of Informatics, St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, St. Petersburg, Russia

Abstract

In the work the local parametric training of Algebraic Bayesian networks is considered. The theorem about the change of Dirichlet distribution parameters during transition from a priori to a posteriori probability distribution on propositional quantum formulas is formulated and proved. The proof is based on the conjugation property of the multinomial and Dirichlet distributions.

Keywords

machine learning, probabilistic graphical models, Algebraic Bayesian networks, parametric training, Dirichlet distribution, multinomial distribution

1. Introduction

One of the main points in machine learning models research is the training of the model — parametric synthesis and structural synthesis. There is a complex mathematical apparatus that determines the correctness of operations behind specific algorithms and calculations.

This is also correct for Algebraic Bayesian networks. The object of the research is approach to local parametric training of Algebraic Bayesian network, what is training of network represented by knowledge pattern. The a priori distribution of probabilities presented in a knowledge pattern is studied and approaches to the formation of the a posteriori distribution resulting from learning are considered.

2. Related works

Neural networks [1, 2, 3, 4, 5] and probabilistic graphical models are one of the machine learning models. Belief Bayesian networks, [6, 7, 8], Markov networks [9, 10, 11] and Algebraic Bayesian networks [12], which are the subject of the research, belong to the class of probabilistic graphical models.

Russian Advances in Artificial Intelligence: selected contributions to the Russian Conference on Artificial intelligence (RCAI 2020), October 10-16, 2020, Moscow, Russia
nak@dscs.pro (N.A. Kharitonov); alt@dscs.pro (T. Alexander)
0000-0001-7531-941X (N.A. Kharitonov); 0000-0003-1814-4646 (T. Alexander)

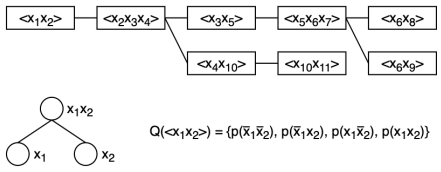

Figure 1: Example of algebraic Bayesian network and the presentations of the first knowledge pattern in it as the ideal of conjuncts and vector of quantum probabilities

Approach for generating the structure of Bayesian network basing on con-straint-based, assessment-based and search-based methods is described in [6]. The use of Belief Bayesian networks for prediction of students expulsion is the subject of research [7]. The detection and prediction of emergencies on atomic power plants by Bayesian network was studied in [8].

One of the main ideas during the work with probabilistic graphical models is their decomposition in smaller parts, which describe some information about the object [13]. These parts are knowledge patterns in Algebraic Bayesian network theory [12]. Fig. 1 presents an example of one of the possible Algebraic Bayesian network graphical representation: the non-directional graph with knowledge patterns in nodes.

The representation of knowledge pattern in the form of proposal formulas-quantum set is used in the context of this research [12], moreover, each propositional formula has scalar or interval probabilities estimate in N.Nilsson interpretation [14]. Other possible representations of knowledge pattern are ideals of conjuncts or disjuncts [12]. The work [15] has description of transition matrix between the set of propositional-formulas-quanta, ideal of conjuncts and ideal of disjuncts.

Main operations of probabilistic-logical inference during the work with Algebraic Bayesian networks are consistency maintaining, a priori and a posteriori inference [16].

Consistency maintaining is verification of correctness of probability estimates which are presented in the network and their changing if it is necessary and possible [17].

A priori inference is receipt of probability estimates of some propositional formula based on the estimates presented in the network [12].

A posteriori inference is the process of changing estimates in network basing on some information, which is presented as evidence. Also the evidence probability is calculated in this process. The study [18] describes the sensitivity of local a posteriori inference.

When working with Algebraic Bayesian networks it is necessary to receive the probabilities of elements in network. The research [19] provides an approach for obtaining algebraic Bayesian network with interval estimates basing on data set with missing values. The approach for receiving quantum probabilities with scalar estimates is investigated in [13] is the subject of this research.

3. Local parameter training of Algebraic Bayesian networks

Tasking of local parametric training of algebraic Bayesian networks is proposed in the paper [13].

x_1	x_2	x_3	$\hat{x}_1\hat{x}_2\hat{x}_3$
0	1	0	010
?	1	0	?10
?	?	1	??1
1	0	0	100
?	1	?	?1?
1	1	1	111
0	0	?	00?
?	?	0	??0
0	1	1	011

Table 1
The example of conversion data set presented as the set of atom variables to the quanta data set (? — missing value)

The input data set is the set of quanta with missing values, in other words, it is the set of implementations of the conjunctions of random binary elements.

For illustrative purposes, let us consider random binary elements above the alphabet x_1, x_2, x_3. We will be interested in the implementations of their conjunctions $\hat{x}_1\hat{x}_2\hat{x}_3$. The designation $\hat{x}$ is used to represent a random binary element that may be set as false or true (0 and 1 respectively for convenience). In this case the implementation (observation) can be either complete or incomplete.

Here are some examples of complete (without missing values) implementations for x_1, x_2, x_3 :

- 001,
- 010,
- 110,
- 111.

Examples of incomplete (with missing values) implementations(the missing value is marked with ?):

- 00?,
- 0?0,
- ??0,
- ?1?.

It is possible that line in input data set is presented as the set of atom variables with 0, 1 or missing value. That case can be easily converted to the quanta. The example for several lines is presented in table 1.

When formalizing local parametric machine training with Dirichlet distribution (single and conjugated distributions) the first step of the training is the creation of "a priori" knowledge pattern with equal quanta probabilities. Application of such uniform distribution on quanta

corresponds to the implicit hypothesis that in conditions of complete uncertainty we can not give preference to any quantum.

Although this is the subject of another study, it should be noted that this hypothesis is not always true. For example, expert knowledge expressed as incomplete, inaccurate, nonnumerical information may affect the choice of a priori distribution. It may become known from an expert that features of a subject area are such that $p(x_1) \leq p(x_2), p(x_2) \geq 0.75$. The information obtained (in fact, initially - knowledge obtained from the Expert Advisor with uncertainty) may significantly affect the choice of a priori distribution.

According to the definition of quanta probability, the probability of each element will be $1/k$, where k is the number of quanta in knowledge pattern, $k = 2^n$, where n is a number of atoms in alphabet, under which knowledge pattern is built.

During the first step of the algorithm the knowledge pattern is trained on a part of data set without missing values.

During the second step the part of data set with missing values is used. On each step the quantum with missing values is converted to the set of quanta without missing values, which are not contradictory to the given. That means that these quanta are acquired from given by all possible replacements of missing value by 0 or 1. Basing on the received set of quanta the probabilities in the knowledge pattern are changed: for those quanta, which are in set, the probabilities increase with the preservation of their ratio, for other they proportionally decrease.

4. Probability distribution

Theorem

After the training of knowledge pattern by proposed method the probabilities of quanta in knowledge pattern will be described by the Dirichlet distribution.

Theorem

Proof

Approaches to the analysis of what happens at receipt of additional data with conjugated distributions are considered deeply enough in modern probability and machine learning theories; in the proof we will use receptions from[20].

The probabilities of quanta in "a priori" knowledge pattern can be described by Dirichlet distribution with parameter vector set as unit vector. The probability density function in that case is a gamma function with knowledge pattern dimensional as parameter:

$$\mathrm{Dir}(Q_k) = \frac{\Gamma(\sum_{i=0}^{k} \alpha_i)}{\prod_{i=0}^{k} \Gamma(\alpha_i)} \prod_{i=0}^{k} \mathrm{p}(q_i)^{\alpha_i - 1} =$$

$$= \frac{\Gamma(k)}{\prod_{i=0}^{k} \Gamma(1)} \prod_{i=0}^{k} \mathrm{p}(q_i)^0 = \Gamma(k),$$

where Q_k is the quanta vector with k dimensional, Γ – gamma function and $\{\alpha_i\} = \{1\}$ is the distribution parameter vector.

Let us take a look at the data set. Every line of data set (which is presented as quantum q with possible missing values) can be converted to the vector of k dimension by next rule:

- this is the 1 value on position i, if quantum q_i is not contradict to the q;
- this is the 0 value in other cases.

Index of quantum is its binary representation converted to decimal number system.

This interpretation of data set allows to consider each input line, corresponding to quantum without missing values, as random experiment, the result of which is the one of the quanta. The 1 value in the relevant vector shows, which quantum is the result.

Note, that each of this experiments is independent on others, as data set has no information about the relationship between trials. Thus, the part of data set without missing values can be considered as n_{full} independent experiments with k possible results. That means that it can be described by multinomial distribution.

Rows corresponding to quanta with missed values need to be considered separately, since for such rows the number of outcomes is more than one, what is impossible in a multinomial distribution.

Each such line l can be converted to the set of vectors $\{r\}$ of length k, such so $|r_i| = 1 \forall i$ and $\sum_i r_i = l$. This set of vectors can be similarly presented as the set of independent experiments, described by multinomial distribution.

Thus, on each step the likelihood function on each step has multinomial distribution.

According to the Bayes theorem the j step of training is:

$$\mathrm{p}_j(Q|x) = \frac{\mathrm{p}_j(x|Q)\mathrm{p}_j(Q)}{\int_Q \mathrm{p}_j(x|Q)\mathrm{p}_j(Q)dQ},$$

where $\mathrm{p}_j(Q)$ is a priori distribution, which is equal to the a posteriori distribution for each $j \geq 1$. For $j = 0$ the a priori distribution is Dirichlet distribution: $\mathrm{p}_0(Q) = \mathrm{Dir}(Q_k)$.

The Dirichlet distribution is conjugate prior to the multinomial distribution, so, a posteriori distribution on each set is Dirichlet distribution.

Thus, the result of training will be the knowledge pattern with quanta probabilities which have Dirichlet distribution.

The theorem is proved

Let us consider the knowledge pattern KP built over the atoms x_1, x_2. In case there is no information about the subject area, the a posteriori distribution of quanta will set their probabilities equal to 0.25. However, some information which is known before training, expressed, for example, in expert estimates, may affect the a priori distribution. Let $\mathrm{p}(x_1) \leq 0.5$. In this case the a priori probability of quanta is: $\mathrm{p}(00) = \mathrm{p}(10) = 0.375, \mathrm{p}(10) = \mathrm{p}(11) = 0.125$.

It should be noted that initial constraints have only limited the class of probability distributions from the whole simplex of possible distributions and narrowed interval estimates. The a priori distribution was chosen as the mass center of the resulting figure in hyperspace. Generally speaking, when considering a priori distribution issues, there are tasks of constructing a canonical representative — a knowledge pattern with scalar estimates — for an initial knowledge pattern with interval estimates. In addition, the gamma distribution parameters can be affected by the "uncertainty level" of the initial knowledge pattern — the extent to which possible

distributions of probabilities (compatible with the knowledge pattern) relative to the canonical representative vary. However, both formalization of this remark and consideration of the degree of uncertainty are tasks of other studies.

Let us suppose that the training resulted in a knowledge pattern with a probability vector of quanta $\mathbf{P}_q$. In works[12, 13] the approach to obtaining the probabilities of conjuncts $\mathbf{P}_c$ and the vector of probability of disjuncts $\mathbf{P}_d$ is described. This approach allows to convert the result of training proposed above to the vector of conjuncts or disjuncts.

In order to do this, we define the matrices $\mathbf{J}_n$ and $\mathbf{K}_n$:

$$\mathbf{J}_1 = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \mathbf{J}_{i+1} = \begin{bmatrix} \mathbf{J}_i & \mathbf{J}_i \\ \mathbf{0} & \mathbf{J}_i \end{bmatrix},$$

$$\mathbf{K}_1 = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}, \mathbf{K}_{i+1} = \begin{bmatrix} \mathbf{K}_i & \mathbf{K}_i \\ \mathbf{K}_i^{\circ} & \mathbf{E} \end{bmatrix},$$

where $\mathbf{E}$ is a matrix with all elements equal to 1, $\mathbf{K}_i^{\circ}$ is received from $\mathbf{K}_i$ by replacing first row elements with 0.

The probabilities of conjuncts and disjuncts can be calculated in the next way:

$$\mathbf{P}_c = \mathbf{J} \times \mathbf{P}_q;$$

$$\mathbf{P}_d = \mathbf{K} \times \mathbf{P}_q.$$

Conclusion

The study covers local parametric learning of Algebraic Bayesian networks. The training process is described and the theorem on the probabilities of quanta in knowledge pattern having Dirichlet distribution is proved.

Further researches in this direction are studying the distribution of probabilities for a knowledge pattern represented as an ideal of conjuncts or disjuncts, as well as researching of the families of distributions received in training with the expert's knowledge, for example training with obtaining interval estimates.

Acknowledgments

The work was carried out with the financial support of the Russian Foundation for Basic Research (project No. 18-01-00626) and within the framework of the project on the state assignment of SPIIRAS No. 0073-2019-0003.

References

[1] R. D'souza, P.-Y. Huang, F.-C. Yeh, Structural analysis and optimization of convolutional neural networks with a small sample size, Scientific reports 10 (2020) 1–13. doi:https://doi.org/10.1038/s41598-020-57866-2.

[2] A. M. M.F. Kazemi, M.A. Pourmina, Novel neural network based ct-nsct watermarking framework based upon kurtosis coefficients, Sensing and Imaging 21 (2020) 7. doi:NovelNeuralNetworkBasedCT-NSCT.

[3] Y. Li, J. Xiang, Existence and global exponential stability of anti-periodic solutions for quaternion-valued cellular neural networks with time-varying delays, Advances in Difference Equations 2020 (2020) 47. doi:https://doi.org/10.1186/s13662-020-2523-4.

[4] Y. Sidi, S. Selouani, B. Zaidi, A. Bouchair, Improving dysarthric speech recognition using empirical mode decomposition and convolutional neural network, Eurasip Journal on Audio, Speech, and Music Processing 2020 (2020) 1–7. doi:https://doi.org/10.1186/s13636-019-0169-5.

[5] M. Stimberg, D. Goodman, T. Nowotny, Brian2genn: accelerating spiking neural network simulations with graphics hardware, Scientific reports 10 (2020) 1–12.

[6] J. Dai, J. Ren, W. Du, V. Shikhin, J. Ma, An improved evolutionary approach-based hybrid algorithm for bayesian network structure learning in dynamic constrained search space, Neural computing & applications 32 (2020) 1413–1434. doi:https://doi.org/10.1007/s00521-018-3650-7.

[7] D. Delen, K.Topuz, E. Eryarsoy, Development of a bayesian belief network-based dss for predicting and understanding freshmen student attrition, European journal of operational research 281 (2020) 575–587. doi:https://doi.org/10.1016/j.ejor.2019.03.037.

[8] Y. Zhao, J. Tong, L. Zhang, G. Wu, Diagnosis of operational failures and on-demand failures in nuclear power plants: An approach based on dynamic bayesian networks, Annals of nuclear energy 138 (2020) 107181.

[9] G. Masetti, L. Robol, Computing performability measures in markov chains by means of matrix functions, Journal of Computational and Applied Mathematics 368 (2020) 112534.

[10] M. Shi, Y. Tang, X. Zhang, Y. Zhang, J. Xu, Modeling and simulation of packet delivery rate in lte-v network based on markov chain, Tsinghua Science and Technology 25 (2020) 357–367. doi:https://doi.org/10.26599/TST.2018.9010142.

[11] Z. Wang, W. Yang, Markov approximation and the generalized entropy ergodic theorem for non-null stationary process, Proceedings of the Indian Academy of Sciences: Mathematical Sciences 130 (2020) 13. doi:https://doi.org/10.1007/s12044-019-0542-4.

[12] A. Tulupyev, S. Nikolenko, A. Sirotkin, Bayesian Belief Networks: Probabilisticlogic Approach [in Russian], SPb.: Nauka, Saint-Petersburg, Russia, 2006.

[13] A. Tulupyev, A. Sirotkin, S. Nikolenko, Bayesian Belief Networks [in Russian], SPbSU Press, Saint-Petersburg, Russia, 2009.

[14] N. Nilsson, Probabilistic logic, Artificial Intelligence 28 (1986) 71–87.

[15] E. Malchevskaya, A. Zolotin, A. Tulupyev, Algorithms of the a posteriori inference in the algebraic bayesian networks: refining of the matrix-vector representation (in russian), in: In: Fuzzy systems and soft computing. Industrial applications. (FTI-2017), 2017, pp. 376–388.

[16] A. Zolotin, E. Malchevskaya, N. Kharitonov, A. Tulupyev, Local and global logical-probabilistic inference in the algebraic bayesian networks: matrix-vector description and

the sensitivity questions (in russian), Fuzzy systems and soft calculations. Tver: TvGTU (2017) 133–150.

[17] N. Kharitonov, E. Malchevskaia, A. Zolotin, M.Abramov, External consistency maintenance algorithm for chain and stellate structures of algebraic bayesian networks: Statistical experiments for running time analysis, in: Proceedings of the third international scientific conference intelligent information technologies for industry (IITI'18), Springer, Cham, 2019, pp. 23–30.

[18] A. Zolotin, A. Tulupyev, Sensitivity statistical estimates for local a posteriori inference matrix-vector equations in algebraic bayesian networks over quantum propositions, estnik St. Petersburg university-mathematics 51 (2018) 42–48. doi:https://doi.org/10.3103/S1063454118010168.

[19] N. Kharitonov, A. Maximov, A. Tulupyev, Algebraic bayesian networks: Naïve frequentist approach to local machine learning based on imperfect information from social media and expert estimates, Communications in Computer and Information Science 1093 (2019) 234–244.

[20] M. DeGroot, Optimal Statistical Decisions, McGraw-Hill, 1970.