Algebraic Bayesian networks: consistent fusion of partially intersected knowledge systems

A Tulupyev[1]**, N Kharitonov**[1] **and A Zolotin**[1]

[1] TICS Lab, SPIIRAS, 39 14th Line, St. Petersburg, Russia

Abstract. In this paper, approaches to the synthesis of a consistent system of probability estimates of propositional formulas for two different partially overlapping data sets based on the theory of algebraic Bayesian networks are presented in this paper. Areas in which the results of this article can be applied are described. An example of synthesis for a particular algebraic network is given.

1. Introduction

With the expansion and the observed acceleration of the digitization of all spheres of the economy (perhaps it would be more accurate to talk about the digitalization of the national economy and even more broadly - about its informatization and about the information society) data sets (including large data sets - " big data ") are becoming more accessible, in fact, obtained not through specially planned expensive experiments or field research, but simply as a result of a" regular "day-to-day economic, organizational, administrative UPE and administrative or other activities. Moreover, the accumulation of data is more likely to be a "side effect" of such activity, rather than specifically planned taking into account the subsequent needs for analyzing these data, searching for patterns in these data, extracting knowledge from these data for the purpose of improving or optimizing processes, or for the preparation and adoption solutions.

This state of affairs, in addition to the positive "availability" of data sets, entails two negative effects. First, the data sets turn out to be fragmentary, often it is impossible to form a single sample, since data from one source can cover one part of the parameters of the observed object or process, the other part of the parameters, these parts may intersect, but prove to be unsuitable for the formation of a "combined" sample element, which would be desirable to be formed on the basis of two or more data sources. Such a situation can arise for a variety of reasons, for example, when the values of the parameters in different sources were recorded at different frequencies or, in general, in different periods of observation. Secondly, the received data sets may be "inadequate" for preparation and decision making simply because some parameters were not planned to be registered, the need for access to their values arose only when the data sets accumulated in information systems were recognized as a useful resource and Analytical efforts began to be made in their attitude.

The aim of this paper is to propose an approach to the synthesis of a consistent system of probability estimates of propositional formulas for two different partially overlapping data sets, based on the theory of algebraic Bayesian networks [1, 2].

2. Examples of applications

This situation arises in a wide range of areas: in epidemiology (in the study of risky behavior and the search for ways to develop preventive programs[3]), information security (in particular, in analyzing the security of information system personnel against social engineering attacks[4]), in advisory systems used in online and offline trade[5], accompanied by physical culture and sports activities (including the training process in the sport of higher achievements), in assessing the reliability of systems, in psychology[6], pedagogy and sociology (including, in the analysis of social networks[7]), in ecology[8].

In particular, when solving the problem of automating the analysis and evaluation of the degree of protection of information system personnel from social engineering attacks, the construction of a graph of social connections fetching highly uncoordinated data from the social network (possibly even from several). However, such data still does not contain parameters that could be used to estimate the effectiveness of preventive impact on personnel (for example, educating or personnel training), since, as noted, when creating social networks, the need to solve such a problem was not taken into account and the corresponding possibility was not pawned. At the same time, the situation is not as hopeless as it might seem, since it is possible to take advantage of the knowledge obtained from other sources, including expert psychologists and experts in the field of information security. This knowledge can help "build bridges" between what has been extracted from the "big data" contained in social networks and the parameters on the basis of the assessment of the values of which it is possible to prepare and make decisions. But the workable "conjugation" of knowledge from heterogeneous sources that becomes a key fundamental scientific problem, the approaches to the solution of which must be found, described and automated.

To emphasize the importance of the problem for a broad class of scientific research and areas of the digital economy, let us give one more example. There are widely distributed advisory systems that, according to the basket of goods in the online stores, tell the user what other goods were bought (or looked at) by other users with a similar set of goods in the basket. Such advisory systems use explicitly or implicitly machine learning of probabilistic graphical models where the trained probabilistic graphical model acts as the core of the advisory system. However, such an advisory system shows the same drawbacks as in the previous example:

- 1) it will not be able to include a new product in its recommendations, since there simply was no data on the sales of this new product;
- 2) it will not be able to respond to scenario-based queries , when the customer indicates that he wants to repair the apartment, waiting for the list of goods, or that he wants to prepare a light dinner for 10 people.

In this case, in the raw data that the corresponding system has accumulated, there are no reasons for decision making however, as in the first example, the way to cope with such class of tasks mightbe to include expert knowledge

3. Using of Probabilistic Graphical Models

We will express common problems that relate the above tasks from the areas of the digital economy and science in terms of the Bayesian networks theory (including machine learning of Bayesian networks) and related models of complex knowledge systems with uncertainty, adhering to the probabilistic, logical-probabilistic, relational-probabilistic and probabilistic-algebraic approaches in order to explicate them.

Bayesian networks (Bayesian belief networks, algebraic Bayesian networks, other related models) can be automatically constructed ("machine-trained") according to data from every single data source obtained from various automated information system. (Of course, open theoretical, algorithmic and technological issues remain even at this stage, but nevertheless we will assume that a satisfactory result of machine learning is available to us). As noted these information sources do not allow them to be directly and consistently combined into a single source, so Bayesian networks are constructed separately. The resulting Bayesian networks will also have an intersection over a set of vertices, because information sources intersect in terms of parameters (variables, attributes). At the same time, both the structure on the set of vertices from the intersection, and the probability estimates at such vertices, most likely, do not coincide. The problem of "merging" of intersecting but not coincident networks was not posed and solved neither in the theory of Bayesian networks of confidence, nor in the theory of algebraic Bayesian networks, nor in theories of related models, although the need for solving such a fundamental problem in the context described above, dictated by digitalization economy, is obvious. (More precisely, we are talking about a series of fundamental problems, because there are questions about finding consistent estimates of probabilities, both conditional and marginal, about finding structures that connect the vertices of networks that appeared in the intersection, about spreading the influence of matching probabilities and structures in place intersections on those parts of networks that did not enter the intersection, etc.).

However, the problem is not confined to the merging of existing networks. As noted, experts can be considered as an information source. Their contribution will result in the addition / completion of Bayesian networks (or related models) with new vertices and a new structure that forms connections both within the new set of vertices and vertices from this set with vertices of the previously "machine-trained" Bayesian network (Bayesian networks). As a rule, in this case, on the one hand, experts will receive incomplete, inaccurate, non-numeric information about both the structure "inside" and "outside", and about the probabilities that characterize the vertices and these connections "inside" and "outside". And according to such data (for example, only data on partial orders of probability estimation can be available, but not accessible by numerical estimates of probability itself), it will be necessary to "machine learn" the expert part of the Bayesian network (or related model), as well as the part that is responsible for the links between the "expert component" and the data-learned from automated information systems. On the other hand, it will be necessary to organize a dialogue for an iterative approximation to a satisfactory final result between the expert and the system providing the "completion" / pre-training of the Bayesian network (or related model).It also raises a number of fundamental and technological issues, starting from the questions of visualization of structures and values of parameters and ending with the issues of resolving collisions, modifying (eliminating) unallowable structural elements, minimizing the required operations, including operations to modify networks in terms of their intersection, conjugation, and parts that are not included into the intersection, but are modified "secondary" because of the types of modifications already listed.

Thus, a "bottleneck" has been revealed, which prevents the technological breakthrough in the application of approaches (to solving problems from a number of spheres of the digital economy and science), methods, models, algorithms, technologies and systems of machine learning and other methods of intellectual analysis. This “bottleneck” is described below: neither the known theoretical developments nor the existing software systems provide the ability to synthesize new Bayesian networks (and related models) on the basis of overlapping but not consentient Bayesian networks (and related models), and also do not provide the possibility of completion / training of these networks and models based on inaccurate, incomplete and non-numeric information coming from experts. Elimination of this bottleneck due to the development of appropriate theories, algorithms complexes, and then, on their basis, data mining systems, will open up within the developing digital economy a broad market for the application of computer-based training and data mining systems based on Bayesian networks (and related models), since all other prerequisites exist for this, including accumulated and accumulating data sets in information sources.

4. Combining two algebraic Bayesian networks

There is a simplest example of possible approaches to combining algebraic Bayesian networks in this part of the article.

Let us consider two algebraic Bayesian networks, each of one knowledge pattern with two atoms in it (figure 1).

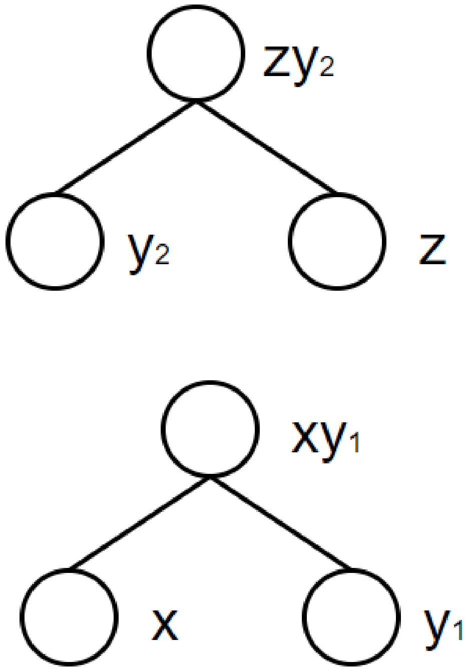

Figure 1. Explored algebraic Bayesian networks

Conjuncts will have following interval estimates in this case:

$$p(x) \in [a_x, b_x];$$
$$p(y_1) \in [a_{y_1}, b_{y_1}];$$
$$p(y_2) \in [a_{y_2}, b_{y_2}];$$
$$p(z) \in [a_z, b_z];$$
$$p(xy_1) \in [a_{xy_1}, b_{xy_1}];$$
$$p(y_2z) \in [a_{y_2z}, b_{y_2z}].$$

There are two possible ways to combine these algebraic Bayesian networks which are characterized by the complexity of the resulting integrated network and the completeness of the information provided. They both are represented in the next two subsections.

4.1. The first method

Atoms y_1 and $y_2$are replaced by $y$in the first case. y has the interval estimate of probability of truth which is equal to intersection of probabilities y_1 and y_2:

$$p(y\quad) \in [max(a_{y_1}, a_{y_2}), min(b_{y_1}, b_{y_2})].$$

The resulting algebraic Bayesian network is shown on figure 2.

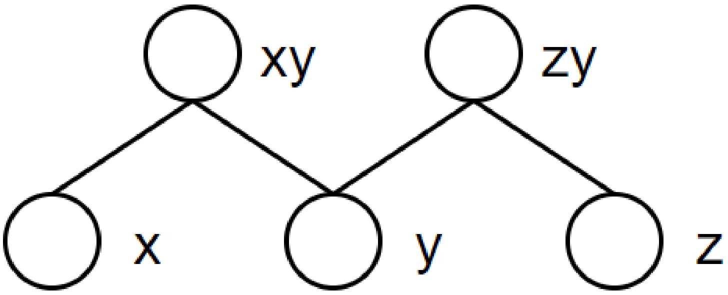

Figure 2. The first method of combining algebraic Bayesian networks.

The probabilities of conjuncts will satisfy the following inequalities:

$$
\begin{aligned}
&1 - p(x) - p(y) + p(xy) \geq 0;\\
&p(x) - p(xy) \geq 0;\\
&p(y) - p(xy) \geq 0;\\
&p(xy) \geq 0;\\
&1 - p(z) - p(y) + p(zy) \geq 0;\\
&p(z) - p(zy) \geq 0;\\
&p(y) - p(zy) \geq 0;\\
&p(zy) \geq 0.
\end{aligned}
$$

4.2. The second method

$y_1$and $y_2$are both added to intersection in the second case. It complicates the structure of resulting algebraic Bayesian network (figure 3), but allows to see the effect of combining of $y_1$and y_2(conjuncts $xy_1y_2$and zy_1y_2), and the effect of these parameters separately (conjuncts xy_1 and xy_2 , zy_1 and zy_2).

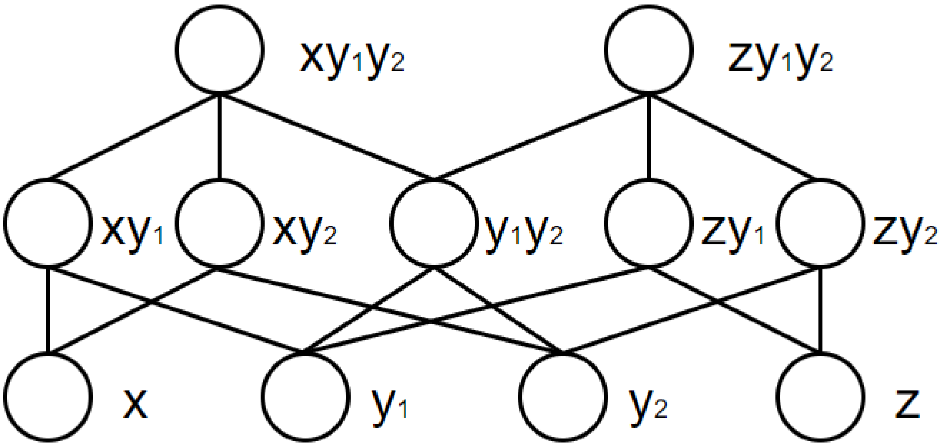

Figure 3. The second method of combining algebraic Bayesian networks.

The probabilities of conjuncts will satisfy the following inequalities:

$p(xy_1y_2) \geq 0;$
$p(xy_1) - p(xy_1y_2) \geq 0;$
$p(xy_2) - p(xy_1y_2) \geq 0;$
$p(y_1y_2) - p(xy_1y_2) \geq 0;$
$p(x) - p(xy_1) - p(xy_2) + p(xy_1y_2) \geq 0;$
$p(y_1) - p(xy_1) - p(y_1y_2) + p(xy_1y_2) \geq 0;$
$p(y_2) - p(xy_2) - p(y_1y_2) + p(xy_1y_2) \geq 0;$
$1 - p(x) - p(y_1) - \quad p(y_2) - p(xy_2) + p(xy_1) + p(xy_2) + p(y_1y_2) - p(xy_1y_2) \geq 0;$
$p(zy_1y_2) \geq 0;$
$p(zy_1) - p(zy_1y_2) \geq 0;$
$p(zy_2) - p(zy_1y_2) \geq 0;$
$p(y_1y_2) - p(zy_1y_2) \geq 0;$
$p(z) - p(zy_1) - p(zy_2) + p(zy_1y_2) \geq 0;$
$p(y_1) - p(zy_1) - p(y_1y_2) + p(zy_1y_2) \geq 0;$
$p(y_2) - p(zy_2) - p(y_1y_2) + p(zy_1y_2) \geq 0;$
$1 - p(z) - p(y_1) - \quad p(y_2) - p(zy_2) + p(zy_1) + p(zy_2) + p(y_1y_2) - p(zy_1y_2) \geq 0.$

4.3.Comparison of methods

So, the first method has the simpler computations, but it is less informative: the construction of a more complex algebraic Bayesian network makes it possible to reveal dependencies on different measurements of the same value. In addition, the second method either guarantees that the existing constraints correspond to a non-empty set of probability distributions (possibly with only one element), or allows one to conclude that the available data set is inconsistent and additional efforts are required to harmonize the information obtained from different sources.

5. Conclusion

A method is proposed for synthesizing a consistent system of probability estimates of propositional formulas for two different partially overlapping sets of data. It is based on the construction of two algebraic Bayesian networks, which will overlap partially, because the original data sets also intersect, their connection over coincident vertices and the subsequent application of algorithms to maintain the consistency of algebraic Bayesian networks. It should be noted that if the resulting network is acyclic or allows a successful conversion to acyclic, then according to [1] it will be possible to maintain its consistency cheaper than if it had to be immersed in an encompassing knowledge pattern.

6. References

[1] Tulupyev A L, Nikolenko S I and Sirotkin A V 2006 “Bayesian belief networks: probabilistic-logic approach,” SPb.: Nauka p 607 (In Russian)

[2] Levenets D G et al. Decremental and incremental reshaping of algebraic Bayesian networks global structures. // Proceedings of the First International Scientific Conference “Intelligent Information Technologies for Industry” (IITI’16) 2016 pp 57–67

[3] Kugbey N et al. International note: Analysis of risk and protective factors for risky sexual behaviours among school-aged adolescents. // Journal of AdolescenceVolume 68. 2018 pp 66-69

[4] Abramov M.V and Azarov A.A. Identifying user's of social networks psychological features on the basis of their musical preferences // Soft Computing and Measurements (SCM), 2017 XX IEEE International Conference on. – IEEE, 2017 pp 90–92

[5] Post Y L et al. Using probabilistic neural networks to analyze First Nations' drinking water advisory data. // Journal of Water Resources Planning and Management. Volume 144. 2018. Issue 11, 1.

[6] Davey C G et al. A brain model of disturbed self-appraisal in depression. // American Journal of Psychiatry. Volume 174, Issue 9. 2017 pp 895-903

[7] Bagretsov G I, Shindarev N A, Abramov M V and Tulupyeva T V Approaches to development of models for text analysis of information in social network profiles in order to evaluate user's vulnerabilities profile // Soft Computing and Measurements (SCM), 2017 XX IEEE International Conference on. – IEEE, 2017 pp 93–95

[8] Santos, R.A.L., Mota-Ferreira, M., Aguiar, L.M.S. Predicting wildlife road-crossing probability from roadkill data using occupancy-detection models // Science of the Total Environment. Volume 642. 2018 pp 629-637

Acknowledgments

This work was partially supported by the by RFBR according to the research project №. 18-01-00626, as well as the state task SPIIRAS № 0073-2018-0001.