Terentiev O.M., Makohon R.O.
Igor Sikorsky Kyiv Polytechnic Institute, Institute for Applied System Analysis, Kyiv, Ukraine

Analysis of causality Bayesian network usage for data-mining purposes

Modern world is constantly progressing, so enormous amounts of information are being generated by various areas of the human life. And the speed if it's creation is only increasing, due to concepts like an internet of things, globalization and developments in the computer miniaturization. All this raw data contains useful information that may help improve performance of the analyzed subject. This confirms the relevance of the data-mining methods usage to find hidden relations within the dataset variables, determine yet unknown dependencies and patterns.

The purpose of this research is to analyze the causality Bayesian network usage as a data-mining tool for the real economical problem, namely the estimation of the person's creditworthiness, discover it's strengths and weaknesses.

As an algorithm for the causality Bayesian network creation the heuristic one was chosen. It is based on such values estimations as the MI (Mutual Information) value and the MDL (Minimum Description Length) one. This model was described in works of researchers like Zheng Y., Kwoh C.K. [1] and Heckerman D., Geiger D., Chickering D.M. [2].

The heuristic method was used as a way to drastically decrease the number of needed computational powers to build the accurate Bayesian network. The following table shows the amount of possible networks that machine learning algorithm may build out of number of nodes in a raw data.

Table 1. Exhausted search disadvantages

Number of nodes	Number of acyclic models
1	1
2	3
3	25
4	543
5	29281
6	3781503
...	..
15	$2.38*10^{41}$
20	$2.34*10^{72}$

The computer program builds the Bayesian network as an acyclic directed graph. Each node there stands for some variable in Bayesian sense. It may be either observable quantile or latent, unknown, hypothetic variables in the training data. Each edge stands for a conditional dependency of the children node from its parent one. So there is an opportunity to calculate all children probabilities, if parent's probabilities are known. The graph nature of the instrument makes the output results visually convenient for perception and analysis.

The heuristic method if based on the mutual information value between pairs of variables in a dataset (1).

$$MI(x^i, x^j) = \sum_{x^i, x^j} P(x^i, x^j) * \log \frac{P(x^i, x^j)}{P(x^i)P(x^j)}. \quad (1)$$

This value provides information about the order of building the optimal dataset. The method begins from the pair of nodes that has the biggest value of MI. There are three possible outcomes for each such pair: two ways of one variable being parent for the other one, and one opportunity for them not to be directly connected. So the method chooses the best outcome of these three ones, using the value of MDL, based on coding theory, suggested by Shennon. It states that there is a fixed minimal length of the code that encodes some message, that fully depends on the message information. So if the code is based on the incorrect perception of the message, its length would be

bigger than the optimal one. When applied to Bayesian networks, this function is determined as

$$L(j,g,x^n) = H(j,g,x^n) + \frac{k(j,g)}{2} * \log(n). \quad (2)$$

So having built Bayesian network by minimizing MDL on each step of the algorithm, optimal structure will be achieved.

The SAS Base programming language was used for building and learning of the analyzed Bayesian networks. No ready-made functions were used to ensure code's high performance and it's conformity with the explored learning method. The data set, which was used to create the network structure, was provided by the forecasting the probability of non repayment of loan competition, organized by https://sascompetitions.ru. It originally contained 34 variables in total of 1787571 observations.

For the accuracy checking purposes more than 10 different test data sets were analyzed by the coded Bayesian network. These examples were provided by SASHelp datasets. Final resulting network structures were validated by comparison with the corresponding GeNIe (http://genie.sis.pitt.edu/) and BayesiaLab (http://bayesia.com/) results. One of the analyzed examples contained 8 nodes, that would take over $7,83 * 10^{11}$ probable networks to examine for the exhausted search. The iterative heuristic search used 28 iterations to consider 120 different possible structures. The needed structure was built at the 81-th try on a 15-th iteration. Here is the graph representation of the program result:

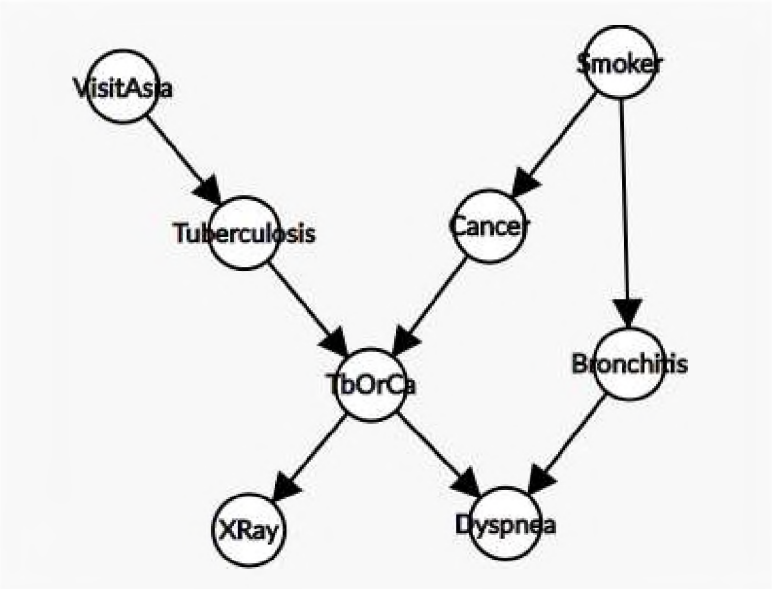

Figure 1. Example of program result network

It is important to state that implemented model allows building Bayesian network using training data, unlimited by the size of the data set and the number of the input variables. And the accuracy of implemented algorithm was equal to the one of the compared program products.

Conclusion. The causality Bayesian network building program product was developed, that allows construction of the network structure, using training data. The accuracy and the performance of the implemented algorithm was successfully tested in comparison with the existing program solutions. The product had demonstrated it's ability to fulfill the described task.

References. **1.** Zheng Y. and Kwoh C.K. "Improved MDL Score for Learning of Bayesian Networks", Proceedings of the international conference on artificial intelligence in science and technology, pp. 98–103, 2004. **2.** Heckerman D., Geiger D., Chickering D.M., "Learning Bayesian Networks: The Combination of Knowledge and Statistical Data," Machine Learning, vol. 20, no. 3, pp. 197–234, September 1995. [Online]. Available: https://link.springer.com/article/10.1007/BF00994016. [Accessed March 30, 2018].