

A Bayesian Network for integration of risk assessment ¹

Concetto Elvio Bonafede ²

Dipartimento di Statistica ed Economia Applicate “Libero Lenti”,
Università di Pavia, e-mail: ingc.bonafede@gmail.com

Abstract: The goal of this work is to introduce a methodological approach for the construction of a bayesian network that integrates the different risks involved in a engineering project for a public structure.

In this case the data are the single expert opinions on the different categories of risk (for instance structural risks, legal, environmental, etc...) and the links among the various risky events (nodes) with the respective degree of correlation.

The links among the different nodes are already defined, therefore the purpose is to understand as, (on the base of such links and the single expert observations about probabilities of the risks and their correlations), we can obtain the conditional probability tables that characterize the relationships among the nodes (risks).

Keywords: Bayesian network, Risk assessment, Optimization, Data Mining

1. Introduction

To learn a Bayesian Net we need to have an adequate database to find the structure and the parameters (the conditional probability table or CPT). Unfortunately, in the real world it is very difficult to find good data, especially when the studied problem is complex.

We have constructed the network by using an already existing data base (DB) where the available information are the risks with their frequencies, impacts and correlation among them. In total there are about 300 main risks involved in the project.

In this work we have considered only the frequencies of risks and no impacts. Impacts could be used in a later phase of scenario analysis to evaluate the loss distribution under the different scenarios.

In Table 1 there is the DB structure used for network learning and in which each risk is considered as a binary variable (one if the risk exists (*yes*) and zero if the risk doesn't exist (*not*)). Therefore, for each risk in the network there will be one node with two states (*one* $\equiv Y$ and *zero* $\equiv N$).

The task is to find the conditional probabilities tables by using only the correlations and the marginal frequencies. Instead, the net structure is obtained from Table 1 by following the node relationships given by correlations.

Table 1: *Expert values database structure (Learning table)*

PARENT	CHILD	CORRELATION	PARENT FREQ.	CHILD FREQ.
RISK A	RISK B	$\rho_{AB} = 0.5$	$P(\text{risk A} = \text{Yes})=0.85$	$P(\text{risk B} = \text{Yes})=0.35$
RISK A	RISK C	$\rho_{AC} = 0.3$	$P(\text{risk A} = \text{Yes})=0.85$	$P(\text{risk C} = \text{Yes})=0.55$

¹This work has been supported by MUSING 2006 contract number 027097, 2006-2010.

2. The methodology

The main ideas for finding a way to construct a BN have been: first to find the joint probabilities as functions of only correlations and the marginal probabilities; second to understand how the correlations are linked with the incremental ratios or the derivatives of the child's probabilities as functions of the parent's probabilities. This choice is due to the fact that parent and child interact through the values of conditional probabilities; the derivatives are directly linked to such probabilities and, therefore, to the degree of interaction between the two nodes and, hence with the correlation.

It is possible to calculate the CPT up to three parents for each child. Although there is the possibility to generalize to more than three parents, it is necessary to have more data which are not available in our DB. So when four or more parents are present it has been decided to divide and reduce to cases with no more than three parents.

To approximate the network we have "separated" the nodes that give the same effects on the child (as for example the same correlations) by using auxiliary nodes (Jensen, 2001; Giudici, 2006). When there was more than one possible scheme available we have used the mutual information (MI) criterion as a discriminating index by selecting the approximation with the highest total MI; this is the same to choose the structure with the minimum distance between the network and the target distribution (Kleiter and Jirousek, 1996; Williamson, 2000).

In this paper is shown the result for the cases of one and two parents. For more details see the technical report and publication (Bonafede and Giudici, 2007a,b): *Construction of a bayesian network for a project of enterprise risk assessment* and *Bayesian Network for enterprise risk assessment*.

2.1 One parent case

The case with one parent is the simplest. Let $P(F)$ and $P(C)$ be the marginal probability given from expert (as in Table 1 where there are three nodes (A, B, C) and theirs correlations):

- For the parent, F, we have: $P(F=Y)=x$, $P(F=N)=1-x$;
- For the child, C, we have: $P(C=Y)=y$, $P(C=N)=1-y$;

The equations to find either the conditional probabilities or the joint probabilities are:

$$\begin{array}{l} \text{CPT equation system} \\ \alpha_1 x + \alpha_2 (1-x) = y; \\ \alpha_1 - \alpha_2 = k; \\ \alpha_1 + \alpha_3 = 1; \\ \alpha_2 + \alpha_4 = 1; \end{array} \quad \left\| \quad \begin{array}{l} \text{Joint equation system} \\ c_1 = \rho M + xy; \\ c_2 = y - \rho M - xy; \\ c_3 = x - \rho M - xy; \\ c_4 = 1 - y - x + \rho M + xy; \end{array} \right.$$

where $k = \rho \sqrt{\frac{Var[C]}{Var[F]}}$; whit α_i and c_i we indicate respectively the conditional and the joint probabilities.

Considering that probabilities c_i and α_i must be positive either the marginal probabilities or the correlation value should be constrained. If the marginal probabilities are fixed then the correlation values must be constrained, which will be normally the case, as estimates of probabilities are more reliable.

It is not possible to have any value of correlation given the marginal probabilities. Indeed, as we want to maintain the marginal probabilities as fixed by the expert, correlation limits can be determined as follows:

$\rho > -\frac{xy}{M} = A$; $\rho > \frac{y+x(1-y)-1}{M} = D$; $\rho < \frac{x(1-y)}{M} = B$; $\rho < \frac{y(1-x)}{M} = C$;
and the correlation interval will be: $\rho \in [\max(A, D); \min(B, C)]$.

2.2 Two parents case

This case is more complicated than the one before. In this situation a further difficulty is that the given expert correlations are only pairwise marginal and, therefore, we need more information to find the CPT.

For example the joint moment among the nodes which is not in the DB. Consequently there can be more than one CPT, corresponding to different values of the joint moment, for the same marginal correlation and probability.

The joint moment becomes thus a project parameter to be set by using an appropriate criterion. We define the standardized joint moment among three variables to be:

$$\rho_{N_i N_j N_k} = \frac{E[(N_i - E[N_i])(N_j - E[N_j])(N_k - E[N_k])]}{\sqrt[3]{\text{Var}[N_i]\text{Var}[N_j]\text{Var}[N_k]}}.$$

To choose among such CPTs we have used the total mutual information (I_{total}) by selecting that CPT with the $\rho_{N_i N_j N_k}$ that gives the minimum I_{total} .

In this case we have to distinguish between independent and dependent parents. The solutions are:

<i>CPT equation system for independent parents</i> $f(x, z) = (\alpha_1 - \alpha_2 - \alpha_3 + \alpha_4)xz + (\alpha_2 - \alpha_4)x + (\alpha_3 - \alpha_4)z + \alpha_4 = y$ $\frac{\partial f}{\partial x} = (\alpha_1 - \alpha_2 - \alpha_3 + \alpha_4)z + (\alpha_2 - \alpha_4) = \frac{(\rho_{AF})(M_{AF})}{x(1-x)}$ $\frac{\partial f}{\partial z} = (\alpha_1 - \alpha_2 - \alpha_3 + \alpha_4)x + (\alpha_3 - \alpha_4) = \frac{(\rho_{BF})(M_{BF})}{z(1-z)}$ $\frac{\partial^2 f}{\partial x \partial z} = \frac{\partial^2 f}{\partial z \partial x} = (\alpha_1 - \alpha_2 - \alpha_3 + \alpha_4) = \frac{(\rho_{ABF})(M_{ABF})}{x(1-x)z(1-z)}$ $\alpha_1 + \alpha_5 = 1$ $\alpha_2 + \alpha_6 = 1$ $\alpha_3 + \alpha_7 = 1$ $\alpha_4 + \alpha_8 = 1$	<i>Joint equation system for dependent parents</i> $c_1 = \rho_{ABF}M_{ABF} + xyz + (\rho_{AF}M_{AF})z + (\rho_{BF}M_{BF})x + (\rho_{AB}M_{AB})y$ $c_1 + c_5 = \rho_{BF}M_{BF} + zy$ $c_1 + c_3 = \rho_{AF}M_{AF} + xy$ $c_1 + c_2 = \rho_{AB}M_{AB} + xz$ $c_1 + c_2 + c_3 + c_4 = x$ $c_1 + c_5 + c_6 + c_2 = z$ $c_1 + c_3 + c_5 + c_7 = y$ $c_8 = 1 - \sum_{i=1}^7 c_i$
---	--

where $M_{BF} = \sqrt{z(1-z)y(1-y)}$, $M_{AF} = \sqrt{x(1-x)y(1-y)}$, $M_{AB} = \sqrt{x(1-x)z(1-z)}$ and $M_{ABF} = \sqrt{x(1-x)z(1-z)y(1-y)}$. As before the α_i and c_i are respectively the conditional and the joint probabilities.

The problem is now setting the marginal correlations when those given from experts are not consistent with the marginal probabilities. Differently from the case with one parent where the correlation belongs to an interval, with two parents the admissible pairs (ρ_{AF}, ρ_{BF}) can be shown to belong to an area.

To approach this problem we have decided to decrease the values of the two correlations ρ_{BF} and ρ_{AF} with a fixed step by maintaining their relative difference. At each step we verified the existence of a value of ρ_{ABF} which supports the new pair (ρ_{AF}, ρ_{BF}) . If it exists the process is stopped, otherwise it goes to the next step; and so on.

If the correlation ρ_{AB} is different from zero (dependent parents), we can set it in advance using the interval obtained for the case of one parent; afterward the ρ_{AB} 's value is used into the joint equation system. Then we can work again only on the pair (ρ_{AF}, ρ_{BF}) by considering the same procedure for independent parents and selecting $\rho_{N_i N_j N_k}$.

2.3 Three parents case

As before two equation systems are obtained. One system for the case with independent parents by which the CPT is directly calculated; another one when there are some correlations between parents and in this case the joint probabilities are calculated instead of the conditional ones. To define the equation systems the analogies between the cases with one and two parents have been exploited. For more details see Bonafede and Giudici (2007b).

3. Conclusions

We have seen that using the equation systems for conditional and joint probabilities the CPTs can be obtained. The method can be generalized to the case with more than three parents, but there are problems in setting more parameters (standardized joint moment) and in looking for more complicated feasible marginal correlation areas. So to develop a network:

- firstly we set the equations and procedure for the one parent;
- than those for two parents distinguishing when they are dependent and not;
- and finally, when possible, we use the equations and the procedures for the three parents case by distinguishing also in this situation between dependent and independent parents; otherwise we split one parent from the others by using the mutual information as splitting index (Bonafede and Giudici, 2007b; Williamson, 2000; Kleiter and Jirousek, 1996).

We remark that we need to reduce to a more simple case those configurations with more than three parents. We can achieve this trying to estimate a local approximate structure, with only one, two and three parents, by "separating" those that give different effects on the child (as for instance different incremental ratios). If there are more schemes available for the substitution we select that with the highest MI (I_{total}) (Bonafede and Giudici, 2007b; Williamson, 2000; Kleiter and Jirousek, 1996).

References

- Bonafede C. and Giudici P. (2007a) Bayesian network for enterprise risk assessment, *Physica A*, 382, 22–28.
- Bonafede C. and Giudici P. (2007b) Construction of a bayesian network for a project of enterprise risk assessment, *Technical Report via <http://www-1.unipv.it/dipstea/workingpapers/35.pdf>*, 35, 1–13.
- Giudici P. (2006) *Applied Data Mining: Statistical Methods for Business and Industry*, Willey.
- Jensen F. (2001) *Bayesian Networks and Decision Graphs*, Springer.
- Kleiter G. and Jirousek R. (1996) Learning bayesian networks under the control of mutual information, *Proceedings in Information Processing and Management of Uncertainty in Knowledge-Based Systems*, 985–990.
- Williamson J. (2000) Approximating discrete probability distributions with bayesian networks, *IEEE Transactions on Information Theory*, 14, 462–467.