Bayesian Networks for Enterprise Risk Assessment: Application to the one Parent Case

Concetto Elvio Bonafede
Data mining lab
University of Pavia
bonafede@eco.unipv.it

Paolo Giudici
Data mining lab
University of Pavia
giudici@unipv.it

Abstract

According to different typologies of activity and priority, risks can assume diverse meanings and it can be assessed in different ways.

In general risk is measured in terms of a probability combination of an event (frequency) and its consequence (impact). To estimate the frequency and the impact (severity) historical data or expert opinions (either qualitative or quantitative data) are used.

In the case of enterprise risk assessment the considered risks are, for instance, strategic, operational, legal and of image, which many times are difficult to be quantified. So in most cases only expert data, gathered in general by scorecard approaches, are available for risk analysis.

The Bayesian Network is a useful tool to integrate different information and in particular to study the risk's joint distribution by using data collected from experts.

In this paper we want to show a possible approach for building a Bayesian networks in the particular case in which only prior probabilities of node states and marginal correlations between nodes are available, and when the variables have only two states.

1 Introduction

A Bayesian Net (BN) is a directed acyclic graph (probabilistic expert system) in which every node represents a random variable with a discrete or continuous state [2, 3]. The relationships among variables, pointed out by arcs (see figure 1), are interpreted in terms of conditional probabilities according to Bayes theorem.

With the BN is implemented the concept of conditional independence that allows the factorization of the joint probability, through the Markov property, in a series of local terms that describe the relationships among variables:

$$f(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} f(x_i | pa(x_i))$$

where $pa(x_i)$ denotes the states of the parents of the variable X_i (child) [1, 2, 3, 6]. This factorization enable us to study the network locally.

One of the problems of a BN is that it requires an appropriate database to extract the conditional probabilities (parameter learning problem) and the network structure (structural learning problem)[1, 3, 13, 17] .

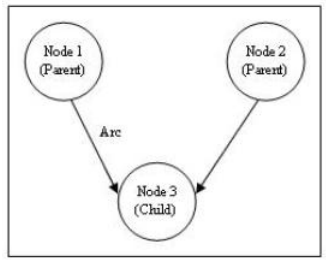

Figure 1: A simple bayesian network. The nodes 1 and 2 (parents) are the predecessors of the node 3 (child).

The objective is to find the net that best approximates the joint probabilities and the dependencies among variables.

The learning approaches can be distinguished according to whether the structure of the net is known or not and if the data are completely or partially observable [13].

After we have constructed the network one of the common goal of bayesian network is the probabilistic inference to estimate the state probabilities of nodes given the knowledge of the values of others nodes. The inference can be done from children to parents (this is called diagnosis) or vice versa from parents to children (this is called prediction) [13].

As the case of learning also for the inference we need a good and efficient methods to calculate the node states in particular when we analyze large networks and database [2, 13, 16].

However in many cases the data are not available because the examined events can be new, rare, complex or little understood. In such conditions experts' opinions are used to collect information that will be translated in conditional probability values or in a certain joint or prior distribution (Probability Elicitation) [17, 20].

There are many problems associated with the gathering of expert probabilities, which many times are affected from biases that can be classified in three principal classes: motivational, over confidential and cognitive [11, 12].

Such problems are more evident in the case in which the expert is requested to define too many conditional probabilities due to the number of the variable's parents. So, when possible, is worthwhile to reduce the number of probabilities to be specified by assuming, for instance, some relationships that impose bonds on the interactions between parents and children.

Among the most used methods for the case of discrete variables there is the noisy-OR defined by Pearl (1988) and its variations and generalizations as noisy-MAX and others [3, 9, 10, 14, 17].

In the business field, Bayesian Nets are a useful tool for a multivariate and integrated analysis of the risks, for their monitoring and for the evaluation of intervention strategies (by decision graph) for their mitigation [3, 5, 7].

Enterprise risk can be defined as the possibility that something with an impact on the objectives happens, and it is measured in terms of combination of probability of an event (frequency) and of its consequence (impact).

To estimate the frequency and the impact distributions historical data as well as expert opinions are typically used [4, 5, 7, 8]. Then such distributions are combined to get the loss distribution.

In the case of enterprise risk assessment the considered risks are, for instance: strategic, operational and legal risks, which many times are difficult to be quantified. So, in general, only data gathered from experts are available for risk analysis.

In this context Bayesian Nets are a useful tool to integrate historical data with those coming from experts which can be qualitative or quantitative [20].

Enterprise risk assessment is a part of enterprise risk management (ERM) whose guidelines are developed by the Committee of Sponsoring the Organization of the Treadway Commission (COSO) that creates standard for enterprises, non-profit organizations and public corporations [15].

The ERM procedure handles the creation of a structure that manages the uncertainties with the relative risks and the associated opportunities.

The ERM must identify the potential harmful events for the organization. It has to manage the consequent risks that the organization can accept, to get value and guarantee the attainment of the business goals.

2 Our proposal

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

What we present in this work is the construction of a Bayesian Net for having an integrated view of the risks involved in the building of an important structure in Italy, where the risk frequencies and impacts were collected by an ERM procedure unsing expert opinions.

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 risks.

In our work we have considered only the frequencies of risks and no impacts. With our BN we can construct the risks' joint probability and the impacts could be used in a later phase of scenario analysis to evaluate the loss distribution under the different scenarios [5].

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 considered risk in the network there will be one node with two states ($one \equiv Y$ and $zero \equiv N$).

Table 1: Expert values database structure (Learning table)

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

The task is, therefore, 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.

The main ideas for finding a way to construct a BN have been: first to find the joint probabilities as functions of only the 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. Afterwards we have understood as to create the equations, for the case with dependent parents we have used the local network topolgy to set the equations.

We have been able 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 we have 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 [13]. 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 [18, 19].

We have analyzed first the case with only one parent to understand the framework, then it has been seen what happens with two independent parents and then dependent. Finally we have used the analogies between the cases with one and two parents for setting the equations for three parents.

In this paper for lack of space, we show only the calculi and the result for the case with one parent. To see whats happen to the case with two and three parents we refer the reader to a more extended paper [21].

2.1 One parent case

The case with one parent (figure 2) is the simplest. Let P(F) and P(C) be the marginal probability:

- 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;

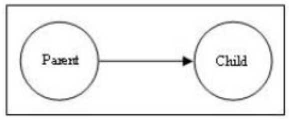

Figure 2: One parent scheme.

The conditional probability table (CPT) and the joint probability table look as in tables 2 and 3.

Table 2: Conditional Probability Table

$P(C\|F)$	F =Y	F =N
C=Y	α_1	α_2
C=N	α_3	α_4

Table 3: Joint Probability Table

$P(C,F)$	F =Y	F =N
C=Y	c_1	c_2
C=N	c_3	c_4

The two matrices can be related using the conditional probability formula:

$$P(F,C) = P(C)P(F|C)$$

Marginalizing joint probabilities and using the previous formula, two systems of equations can be obtained:

$$\begin{array}{l||l}
\textit{CPT equation system} & \textit{Joint equation system} \\
\alpha_1 x + \alpha_2(1-x) = y; & P(F=Y) = c_1 + c_3 = x; \\
\alpha_3 x + \alpha_4(1-x) = 1-y; & P(F=N) = c_2 + c_4 = 1-x; \\
\alpha_1 + \alpha_3 = 1; & P(C=Y) = c_1 + c_2 = y; \\
\alpha_2 + \alpha_4 = 1 & P(C=N) = c_3 + c_4 = 1-y;
\end{array}$$

Such systems, of course, don't have an unique solution because one equation is dependent from the others. Hence one more equation (independent) is required. To look for such equation we can use the correlation given by the expert. This correlation is defined formally by:

$$\rho = \frac{Cov[X,Y]}{\sqrt{Var[X]Var[Y]}}$$

Replacing in the correlation formula with the values of variances and covariance calculated for binary variable we obtain:

$$\rho = \frac{c_1 - xy}{\sqrt{x(1-x)y(1-y)}} \Rightarrow c_1 = \rho M + xy; \text{ and } M = \sqrt{x(1-x)y(1-y)}$$

In this way there is one more equation to be inserted in the joint system and so we get the joint equation system:

$$\begin{aligned}
c_1 &= \rho M + xy; \\
c_2 &= y - \rho M - xy; \\
c_3 &= x - \rho M - xy; \\
c_4 &= 1 - y - x + \rho M + xy;
\end{aligned}$$

Once the joint probabilities are found, the CPT can be obtained from conditional probability formula.

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 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:

$$\begin{aligned}
\rho &> -\frac{xy}{M} = A; \\
\rho &< \frac{x(1-y)}{M} = B; \\
\rho &< \frac{y(1-x)}{M} = C; \\
\rho &> \frac{y + x(1-y) - 1}{M} = D;
\end{aligned}$$

Establishing the maximum between A and D and the minimum between B and C the correlation can be shown to be constrained to:

$$\rho \in [max(A, D); min(B, C)] ;$$

What we have obtained is a first solution for the problem with one parent; we can calculate the CPT for a full network (with one parent) using the previous equations' system for every two nodes and truncating the correlation every time is out of the interval range. In this way the error is limited between nodes.

A second solution to our problem can be found using the derivative of the child's probability in function of the parent's probability and replacing the conditional probabilities in function of the joint ones in the CPT system.

More precisely, from the system of conditional probabilities we have:

$$\tfrac{\partial f}{\partial x} = \tfrac{\Delta y}{\Delta x} = \alpha_1 - \alpha_2 = \alpha_4 - \alpha_3 = k;$$

In this manner one more equation is obtained and the CPT equation system becomes:

$$\begin{aligned}
\alpha_1 x + \alpha_2(1-x) &= y; \\
\alpha_1 - \alpha_2 &= k; \\
\alpha_1 + \alpha_3 &= 1; \\
\alpha_2 + \alpha_4 &= 1;
\end{aligned}$$

Note that also in this case we cannot have any conditional probability values given the marginal ones (it is a problem of consistence). This can be seen using

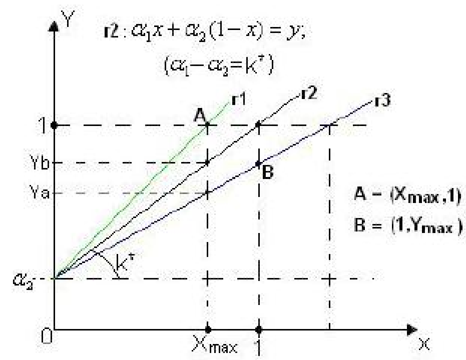

Figure 3: Graph that shows how the child's marginal probability (Y) varies with that of the parent (X) in function of the conditional probabilities.

the slope coefficient k. For example, in figure 3 we show that with different values of k, only one probability, either that of the child or that of the parent, can be one, (points A or B).

When k is zero the child is independent from the parent and its probability to be in state Y is $\alpha_2 = \alpha_1$.

Besides, there is only one value for k, given α_2, such that when P(F = Y) is one also P(C = Y) is one (k^*). Furthermore it can be seen that if P(F=Y) is fixed at the value X_{max} the maximum value for P(C=Y) varies with k among (X_{max}, Y_a), (X_{max}, Y_b) and A.

The coefficient k is a function of the correlation and of the marginal probability. This can be seen by replacing the value of the conditional probabilities in function of the joint probabilities:

$$\begin{aligned} k &= \alpha_1 - \alpha_2 = \frac{c_1}{c_1 + c_3} - \frac{c_2}{c_2 + c_4} = \frac{\rho M + xy}{x} - \frac{y - \rho M - xy}{1 - x} = \frac{\rho M}{x(1-x)} \\ &= \rho \sqrt{\frac{Var[C]}{Var[F]}} \end{aligned}$$

As we are in the binary case, we can consider the $odds_{ratio}$ as dependence measure. The odds are defined as:

$$odds(A|B) = \frac{P(A|B)}{P(A^c|B)} = \frac{P(A, B}{P(A^c, B)}$$

The $odds_{ratio}$ is:

$$odds_{ratio} = \frac{odds(X = 1|Y = 1)}{odds(X = 1|Y = 0)} = \frac{P(X|Y)P(X^c|Y^c)}{P(X^c|Y)P(X|Y^c)};$$

When two nodes are independent then:

$$odds_{ratio} = \frac{P(X)P(X^c)}{P(X)P(X^c)} = 1$$

The odds for our scheme (in figure 2) are:

$$\theta_{11} = \frac{c_1}{c_2}; \theta_{10} = \frac{c_2}{c_4}; \theta_{01} = \frac{c_3}{c_1}; \theta_{00} = \frac{c_4}{c_2};$$

Dividing the numerator and denominator for $c_2 c_3$ and replacing the formulas of the odds, k is found in function of the θ_{ij}:

$$k = \frac{c_1c_2 + c_1c_4 - c_1c_2 - c_2c_3}{c_1c_2 + c_1c_4 + c_3c_4 + c_2c_3} = \frac{\theta_{11}/\theta_{10} - 1}{\theta_{11} + \theta_{11}/\theta_{10} + 1 + \theta_{10}}$$

If the parent and the child are independent, then:

$$\theta_{11}/\theta_{10} = 1 \Leftrightarrow k = 0 \Leftrightarrow \rho = 0$$

2.1.1 Example

Consider table 4, as an example of expert assignment, that we used to learn the network structure and parameters in figure 4.

Table 4: Learning table for the network in figure 4.

PARENT	CHILD	CORRELATION	PARENT FREQ.	CHILD FREQ.
NODE 1	NODE 2	0.8	0.85	0.45
NODE 2	NODE 3	0.5	0.45	0.65

Using one of the two systems developed before, the CPTs become as in tables 5 and 6. Moreover, the correlation ρ_{12} is shifted from 0.8 to 0.38 because its value is outbound, instead for the correlation ρ_{23} the value 0.5 is admissible.

The resulting network (using the software GeNIe and the values in tables 5 and 6) is in figure 4.

The marginal probabilities are those of the experts but if the node one is set with another probability, for example $P(N1 = y) = 1$, then the propagation of the probability given from the network will be:

$$P(N1 = y) = 1 \Rightarrow P(N2 = y) = 0.5294 \Rightarrow P(N3 = y) = 0.6880.$$

Using such values we can calculate the incremental ratios by the formula:

$$k^C_{ij} = (y^j_{new} - y^j_{old})/(x^i_{new} - x^i_{old});\ j > i;$$

and compare them with the theoretical values (in table 7) calculated by using the values in table 4 and the following formula:

$$k^T_{ij} = \rho_{ij}\sqrt{\frac{Var[Node_j]}{Var[Node_i]}};\ j > i;$$

In table 7 the results are reported; the theoretical and the calculated values can be considered quite close each other.

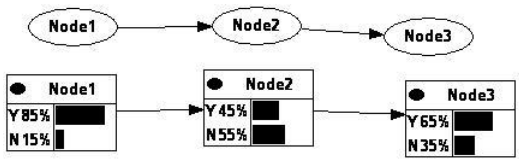

Figure 4: Network given from the table 4 whose CPTs (tables 5 and 6) are calculated by the joint systems.

Table 5: CPT for nodes 1 and 2.

$P(N_2\|N_1)$	$N_1 = Y$	$N_1 = N$
$N_2 = Y$	0.5294	0
$N_2 = N$	0.4706	1

Table 6: CPT for nodes 2 and 3.

$P(N_3\|N_2)$	$N_2 = Y$	$N_2 = N$
$N_3 = Y$	0.9137	0.4343
$N_3 = N$	0.0863	0.5657

3 Conclusion

What we have presented here is a simplest case but we have also extended the calculi to when there are two and three parents (independent or not) [21].

However, with two parents the DB's available correlations are only pairwise and we miss an important data which is the joint moment among the three nodes. The joint moment in this situation is considered as a project parameter, in fact we have more than a result when the joint moment varies. For choosing among such solutions we use the mutual information as index, by selecting the joint moment's value such that gives the minimal mutual information among the three nodes [18, 19].

One more problem is to look for admissible correlations (between each parent and common child) that in this case are inside an area instead of an interval as the case with one parent.

For the case with three parents we have to manage more missing data (the all joint moment among parents' pair and the child). To treat this case we have used again the mutual information to look for one solution [21].

In practice we can generalize the equation system to the case with more than three parents but the raising evident difficulties of setting the joint moments have led us to stay inside the three parents' case by separating the parents with different effects on the child (for example as the incremental ratios).

So, to develop a network we propose to use, separately, firstly the equations and procedure for the one parent; secondly those for two parents distinguishing when they are dependent and not. Finally we use the equations and the procedures for the three parents case by distinguishing also in this situation between dependent and independent parents [21].

Table 7: Comparison between theoretical and calculated incremental ratios.

Incremental ratio (k_{ij})	k_{12}	k_{23}
Theoretical	0.5294	0.4793
Calculated	05293	0.4798

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}) [18, 19].

Acknowledgments
The authors acknowledge financial support from the MIUR-FIRB 2006-2009 project "Data mining for small and medium enterprises" and MUSING project "MUlti-industry, Semantic-based next generation business INtelliGence", contract number 027097.

References

[1] Heckerman D. (1996). A tutorial on learning with Bayesian networks. *Microsoft Research tech. report MSR-TR-95-06.* Revised November 1996, from http://research.microsoft.com.

[2] Cowell R.G., Dawid A. P., Lauritzen S.L. and Spiegelhalter D.J. (1999). *Probabilistic Networks and Expert Systems.* New York, USA: Springer.

[3] Jensen F.V. (2001). *Bayesian Networks and Decision Graphs.* New York, USA: Springer.

[4] Cruz M.G. (2002). *Modeling, measuring and hedging operational risk.* West Sussex, England: John Wiley and Sons.

[5] Alexander C.(Ed.). (2003). *Operational Risk, regulation analysis and management.* London, England: Prentice Hall.

[6] Giudici P. (2003). *Applied Data Mining.* West Sussex, England: John Wiley and Sons.

[7] Cruz M.G.(Ed.). (2004). *Operational risk modeling and Analysis.* London, England: Risk Books.

[8] Fanoni F., Giudici P. and Muratori G.M. (2005). *Operational risk: measurement, modelling and mitigation.* Milan, Italy: Il Sole 24 Ore.

[9] Zagorecki A., and Druzdzel M. (20049. An Empirical Study of Probability Elicitation under Noisy-OR Assumption. In *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2004)*, pp 800-885.

[10] Francisco J.D., and Severino F.G. (2003). Efficient Computation for the Noisy MAX. *International journal of intelligent systems*, Vol. 18, pp 165-177.

[11] Wiegmann D.A. (2005). Developing a Methodology for Eliciting Subjective Probability Estimates During Expert Evaluations of Safety Interventions: Application for Bayesian Belief Networks. *Aviation Human Factors Division*, October, from www.humanfactors.uiuc.edu.

[12] Daneshkhah A.R. (2004). Uncertainty in Probabilistic Risk Assessment: A Review. *The University Of Sheffield*, August 9, from http://www.shef.ac.uk/beep/publications.html.

[13] Murphy K.P. (2001). An introduction to graphical models. *A Brief Introduction to Graphical Models and Bayesian Networks*, May 10, from http://www.cs.ubc.ca.

[14] Agnieszka 0., Druzdzel M. and Wasyluk H. (2001). Learning Bayesian network parameters from small data sets: application of Noisy-OR gates. *International journal of Approximate Reasoning*, Vol. 27, pp 165-182.

[15] COSO (2004). Enterprise Risk Management Integrated Framework: Executive Summary. *Committee of Sponsoring Organizations of the Treadway Commission*, September 2004, from http://www.coso.org/.

[16] Jaakkola T.S. and Jordan M.I. (1999). Variational probabilistic inference and the QMR-DT database. *Journal of Artificial Intelligence Research*, Vol. 10, pp 291-322.

[17] Druzdzel M.J. and van der Gaag L.C. (2000). Building Probabilistic Networks: Where Do the Numbers Come From? *IEEE Transactions on Knowledge and Data Engineering*, Vol. 12(4), pp 481-486.

[18] Williamson J. (2000). Approximating discrete probability distributions with Bayesian networks *IEEE Transactions on Information Theory*, Vol. 14(3), pp 462-467.

[19] Kleiter G.D. 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*, pp 985-990.

[20] Druzdzel M.J. and van der Gaag L.C. (1995). Elicitation of Probabilities for Belief Networks: Combining Qualitative and Quantitative Information *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pp 141-148.

[21] Bonafede C.E., Giudici P. (2006). *Construction of a bayesian network for a project of enterprise risk management*. Technical report, University of Pavia, Italy.