MaCO: A Tool For Aiding Management Of Cost Overrun Of A Software Development Project

Kawal Jeet1, Renu Dhir2

1Department of Computer Science, DAV College, Jalandhar, India
2Department of Computer Science & Engineering, NIT, Jalandhar, India
1kawaljeet80@yahoo.com, 2dhirr@nitj.ac.in

Abstract-‘Cost Overrun’ is one of the major factors for the failure of a software development project. The major reason for this difficulty is poor analysis and management of risk factors. This paper aims to find out the ways to manage and predict the overrun in this cost by using Influence Diagram (ID). Two major risk factors having adverse impact on the cost are decided to be: (1) creeping user requirements (2) use of unnecessary features in the project. An integration of impacts of these factors with experts’ opinion and with the existing databases (consisting of probability of occurrence of risk factors) helped us to create an ID based system that is capable to model the cost overrun. This system can be used by the software manager at any stage during software development.

Key Words: Bayesian Network, Cost Management, Influence Diagram, K2 Algorithm, Project Size, Risk Management, Software Development.

1. Introduction

Software development projects (SDP) often exceed time and budgetary allocations and sometimes even fail to meet the user’s requirements. There are several factors that are responsible for this failure [1]. These are called risk factors and are possible future event that, if occurs, will lead to an undesirable outcome [2]. Some of the key factors responsible for this overrun are [3] creeping user requirements; requirement instability and involvement of unnecessary features, staff training etc. Numbers of subjective techniques were developed and used in software industry to manage the cost but they lack strong scientific theory and clarity. So these are obsolete. We have developed a scientific tool named ‘Management of Cost Overrun’ (MaCO) which is based on theory of Risk Management [7] and utilizes the Influence Diagram ID [6].

1.1. Risk Management

Software risk management is a key discipline for making effective decisions and communicating the results within software organizations [7]. The purpose of risk management is to identify potential managerial and technical problems before they occur so that actions can be taken that reduce or eliminate the likelihood or impact of these problems. This process consists of the following activities:
— Potential problems will be identified
— The likelihood and consequences of these risks will be understood
— The priority order in which risks should be addressed will be established
— Treatment alternatives appropriate for each potential problem above its risk threshold will be recommended
— Appropriate treatments will be selected for risks above their thresholds
— The effectiveness of each treatment will be monitored
— Information will be captured to improve risk management policies
— The risk management process and procedures will be regularly evaluated and improved

The focus of this research is risk prioritization which is a sub activity of risk analysis phases. It helps in reducing the number of failed SDP by early identification of high risk elements and ID has been used for the same.

1.2. Influence Diagrams

An influence diagram is a simple visual representation of a decision problem. Influence diagrams offer an intuitive way to identify and display the essential elements, including decisions, uncertainties, and objectives, and how they influence each other [4]. It provides a clear, graphical picture of a problem and helps in presenting important relationships and relevance. It provides a means to compare alternatives. The modeling is based on probabilistic theory. According to [6], Influence diagrams use shapes called
nodes and arrows called arcs, which enable the diagram to function as a graphical representation of a system. Nodes represent system variables while arcs represent influences between variables. The network contains circles or ovals called chance nodes which represents a discrete random variable, diamond called utility node which represents the desirability of different event combinations involved in the network. Normally, an arc in an ID denotes an influence. For instance if there is an arc from M to O it means the node at the tail (M or N in Figure 1) of the arc influences the value (or the probability distribution over the possible values) of the node at the head (O in Figure 1) of the arc.

![Figure 1 Cause-effect relationship](image)

According to Baye’s Theorem [5]

\[ P(C|A) = \frac{P(A|C)P(C)}{P(A)} \]  

(1)

In this formula O is the hypothesis required to be tested. M is the evidence that confirms or disconfirms the hypothesis.

2. Motivation

Since the early 80’s, IDs have been used in a wide variety of applications such as diagnosis of breast cancer [8], management of training of the staff involved in development of a software project [9], etc. With respect to the present research area, many opaque and empirical studies have been done for the management of cost of a SDP [10], [11]. The authors of [12-13] describe the way to compress the cost of a SDP. These studies provided illuminating insights into management of cost but are not enough in explaining its true impact. Moreover, the uncertainty in occurrence of these risk factors is also not taken into consideration. In this paper we identify the risk factors having an influence on cost of a SDP. By using ID, we can calculate the exact slippage in cost.

3. Design of MaCO

According to [14], 30% or more projects failed due to budget overrun. According to [15-17], and based on extensive interviews we have conducted with 45 software professionals, it has been identified that the following risk factors have more adverse impact on cost of a SDP than others. Brief descriptions of these factors are as follows:

- Creeping User Requirement: User keeps on posing the requirements about the project throughout the system development [18].
- Unnecessary features: Adding more functionality/features than actually required [19].

The aim of designing MaCO using ID is to calculate the increase (in lakhs) in the cost of the project due to the risk factors mentioned above. These risk factors are presented as chance nodes as shown in Figure 2. ‘Project Size’ also represented as chance node as it is dependent on the probability of occurrence of these risk factors. Slippage in ‘Cost’ is driven by the parent node ‘Project Size’ and is represented by diamond shaped node.

![Figure 2 Basic ID with probability tables to calculate ‘Cost Overrun’](image)

3.1 Use of SMILE

SMILE and java has been used to develop MaCO. SMILE is developed by Decision Systems Laboratory (DSL) [6]. It provides platform independent library built in C++ classes for reasoning in BNs and IDs. It helps in building probabilistic models graphically. SMILE libraries can be accessed from within Java applications by using a Java Native Interface (JNI) library called jSMILE (used to make ID). The interface for MaCO is developed by using NETBEANS (an IDE for java) [20]. SMILE wrappers are used for dealing with databases.

3.2 Measurement Scale for Influence

For each node (of ID) of MaCO, a measurement scale i.e. a categorization (such as frequent, probable etc.) of possible outcomes is required. With the help of experts, following five categories have been identified for the nodes discussed in the beginning of section 3.

- Frequent: If the risk factor occurs very often.
- Probable: If it occurs less frequently.
Occasional: If it occurs at a normal frequency.
Remote: If it occurs less.
Improbable: If the risk rarely occurs in the present SDP.

Since all the risk factors do not have the same effect on Project Size, so severity level of each risk factor has to be identified.

The measurement scale for severity of risk factor is developed by extensive interviews form 45 software engineers. According to these interviews and with reference to [2], the default scale for the above is found to be:
Catastrophic: If the risk factor has very severe impact/loss.
Critical: If loss is lesser severe.
Serious: If loss due to the risk factor is normal.
Minor: If consequence or loss is less,
Negligible: If consequence or loss is least.

### 3.3 MaCO Automation

By using NETBEANS (for IDE), jSMILE (for ID) and SMILE wrappers (for dealing with databases), MaCO model has been developed with the following functionalities.

Project managers (users) can enter the impacts of the risk factors responsible for overrun in the cost of project, as shown in Figure 3.

As shown in Figure 4, after accepting inputs from the user, MaCO asks for another set of inputs. The user can set the evidence (i.e., probability of occurrence of the risk factors in the project) using this interface.

MaCO generates the ID (.xdsl file) in the background, and ‘Cost Overrun’ is displayed as shown in Figure 5.

### 3.4 Generating conditional probability table

Impacts of risk factors entered as shown in Figure 3 are normalized. This means their relative strength of influence on child node is calculated [21]. These normalized values are used to generate CPT for node ‘Project Size’ which is further required to generate ID. This can be done using function normalized as discussed below.

\[
\text{Input: } \text{Impact of each risk factor involved } i_1,i_2,\ldots, i_n \\
\text{Output: Normalized weights of given risk factors } w_1,w_2,\ldots,w_n
\]

For each of the risks involved, \( t_i \) is assigned value on the basis of the severity of its impact.

\[
t_i = \begin{cases} 
1 & \text{if impact is negligible} \\
4 & \text{if impact is critical} \\
3 & \text{if impact is severe} \\
2 & \text{if impact is minor} 
\end{cases}
\]

For all the risk factors \( i=1,2,\ldots,n \), the relative weights \( w_i \) comes out to be in the range \([0,1]\). Sum of relative weights of all the risk factors comes out to be 1.

\[
0 \leq w_i \leq 1 \\
w_1 + w_2 + \ldots + w_n = 1 \quad (2)
\]

Function normalize is

\[
\text{end function}
\]

For instance if the impact of various risk factors on delay in ‘Cost’ is as entered in Figure 3 then relative or normalize weights are calculated as discussed below.

For ‘Creeping User Requirements’ the impact is entered to be ‘Critical’. So as discussed in function normalize  \( t_1 = 4 \).

Similarly, for ‘Unnecessary features’ the impact is entered to be ‘Severe’. So \( t_2 = 3 \).

Corresponding relative weights calculated by using function normalize is

\[
w_1 = t_1/(t_1+t_2) = 4/(4+3) = 4/7 \\
w_2 = t_2/(t_1+t_2) = 3/(4+3) = 3/7 \quad (3)
\]

After getting normalized weights we can countercheck the weights by calculating \( \sum_{i=1}^{n} w_i \) which comes out to be \( w_1 + w_2 \) (calculated above) = 7/7 = 1 as discussed in (2).

K2 algorithmic technique [22] could be used to generate CPT for ‘Project Size’. K2 is an algorithm for constructing a BN from a database of records. For the current ID, this database is built after analyzing past historic data and expert interviews. One of such a sample is Table 1. Each row of this table indicated the probability of occurrence of all the risk factors and ‘Project Size’ in a particular case. In order to calculate the joint CPT for ‘Project Size’ with respect to all the risk factors occurring simultaneously [22], temporarily ID’s and hence CPT’s are to be constructed for each individual relationship (indicated in Figure 5) by using K2 technique. CPT is developed for: ‘Creeping User Requirements’--‘Project Size’, ‘Unnecessary Features’--‘Project Size’, as shown in Table 2, 3. Let \( c_1, c_2 \) denotes matrices holding the CPT (as shown in Table 2, 3) of child nodes from each of the temporary ID’s built as mentioned above. For instance from temporary BN containing ‘Unnecessary Features’ and ‘Project Size’ we can get CPT \( c_2 \) as \( p(\text{Project Size}|\text{Unnecessary Features}) \) as discussed in section 1.2.

Similarly \( c_1 \) can be calculated.

CPT for ‘Project Size’ can be calculated by using following formula. According to [21],
\[ p(x^q \mid y_1^{q_1}, y_2^{q_2}) = \sum_{j=1,2,3,4,5} w_j^i \cdot p(x^q \mid y_1^{q_j}, y_2^{q_j}) \]  

where \( q, q_1, q_2 \) = {Frequent, Probable, Occasional, Remote and Improbable}. 

\( x \) = Project Size 
\( y_1 \) = Creeping User Requirement 
\( y_2 \) = Requirement Instability 

\( w_j \) is normalized impact of risk factor \( j \) as obtained by function normalize discussed in section 3.3. 

This formula indicates joint probability of possible outcome of factor \( x \) to be \( q \) provided \( y_1, y_2 \) has possible occurrence to be \( q_1, q_2, q_3, q_4, q_5 \) respectively is equal to summation of multiplication of each risk factor's normalized weight and probability of occurrence of \( x \) with outcome \( q \) with respect to risk factor and its outcome. 

For instance, 

Input: Impacts of risk factors as entered in Figure 3. Probability of occurrence of these risk factors as entered in Figure 4. 

Output: Probability of occurrence of ‘Project Size’ to be Occasional. 

Impacts entered in Figure 3 are normalized as 4/7 and 3/7 as mentioned above. 

From table 2, it is obtained that if probability of occurrence of ‘Creeping user requirement’ is ‘Probable’ then probability of occurrence of ‘Project Size’ to be occasional is 0.166667. 

From table 3, it is obtained that if probability of occurrence of ‘Unnecessary Features’ is ‘Remote’ then probability of occurrence of ‘Project Size’ to be occasional is 0.285714. 

As mentioned above, the normalized weights \( w_1 \) and \( w_2 \) will be 4/17 and 3/17. So by using (4) 

\[ p(x^{\text{Occasional}} \mid y_1^{\text{Probable}}, y_2^{\text{Remote}}) = \frac{4/7 \cdot 0.25 + 3/7 \cdot 0.142857}{0.14285714285714285714285714 + 0.061224428571428571428571429 + 0.20408157142857142857142843} \]  

is as shown in Table 5. 

### 3.5 Evaluation of the Result 

MaCO uses the above built ID and calculates the actual slippage in ‘Cost’ (in lakhs) due to already mentioned risk factors. ‘Cost Overrun’ is a utility node which represents the desirability of different event combinations involved in the network. It is driven by ‘Project Size’, as shown in Figure 2 and as discussed in section 3. For instance, if probability of occurrence of risk factors is as entered in Figure 4 then ‘Project Size’ for each possible outcome as obtained from its CPT populated in section 3.4 and are shown in Table 7 is 

A utility table is associated with this node which holds the ‘Cost Overrun’ in lakhs for each possible outcome of ‘Project Size’. This table could be populated with values obtained by expert interviews. For instance if the utility table is as shown in Table 9 then it means if Project Size happens to be ‘Frequent’ then ‘Cost Overrun’ will be by 12 months. For the case discussed above, the slippage in ‘Cost’ or increase in project ‘Cost’ will be 1.591836 + 2.387754 + 1.224492 = 6.01042 lakhs as shown in Figure 6. 

<table>
<thead>
<tr>
<th>Project Size</th>
<th>Unnecessary Features</th>
<th>Creeping User Req.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequent</td>
<td>Occasional</td>
<td>Remote</td>
</tr>
<tr>
<td>Probable</td>
<td>Remote</td>
<td>Probable</td>
</tr>
<tr>
<td>Improbable</td>
<td>Remote</td>
<td>Improbable</td>
</tr>
<tr>
<td>Remote</td>
<td>Probable</td>
<td>Probable</td>
</tr>
</tbody>
</table>

**Table 2 CPT for Project Size depending on Creeping User Requirement**

<table>
<thead>
<tr>
<th>Creeping Requirement</th>
<th>Frequent</th>
<th>Probable</th>
<th>Occasional</th>
<th>Remote</th>
<th>Improbable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequent</td>
<td>0.166667</td>
<td>0.125</td>
<td>0.166667</td>
<td>0.25</td>
<td>0.285714</td>
</tr>
<tr>
<td>Probable</td>
<td>0.166667</td>
<td>0.25</td>
<td>0.166667</td>
<td>0.125</td>
<td>0.142857</td>
</tr>
<tr>
<td>Occasional</td>
<td>0.166667</td>
<td><strong>0.25</strong></td>
<td>0.333333</td>
<td>0.125</td>
<td>0.142857</td>
</tr>
<tr>
<td>Remote</td>
<td>0.333333</td>
<td>0.25</td>
<td>0.166667</td>
<td>0.125</td>
<td>0.285714</td>
</tr>
<tr>
<td>Improbable</td>
<td>0.166667</td>
<td>0.125</td>
<td>0.166667</td>
<td>0.375</td>
<td>0.142857</td>
</tr>
</tbody>
</table>

**Table 3 CPT for Project Size depending on Unnecessary Features**

<table>
<thead>
<tr>
<th>Unnecessary Features</th>
<th>Frequent</th>
<th>Probable</th>
<th>Occasional</th>
<th>Remote</th>
<th>Improbable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequent</td>
<td>0.166667</td>
<td>0.142857</td>
<td>0.285714</td>
<td>0.142857</td>
<td>0.25</td>
</tr>
<tr>
<td>Probable</td>
<td>0.166667</td>
<td>0.142857</td>
<td>0.142857</td>
<td>0.285714</td>
<td>0.125</td>
</tr>
<tr>
<td>Occasional</td>
<td>0.166667</td>
<td>0.285714</td>
<td>0.142857</td>
<td>0.142857</td>
<td>0.25</td>
</tr>
<tr>
<td>Remote</td>
<td>0.166667</td>
<td>0.285714</td>
<td>0.285714</td>
<td>0.142857</td>
<td>0.25</td>
</tr>
<tr>
<td>Improbable</td>
<td>0.333333</td>
<td>0.142857</td>
<td>0.142857</td>
<td>0.285714</td>
<td>0.125</td>
</tr>
</tbody>
</table>
Table 4 Partial CPT for Project Size

<table>
<thead>
<tr>
<th>Risk Factor</th>
<th>Frequent</th>
<th>Probable</th>
<th>Occasional</th>
<th>Remote</th>
<th>Improbable</th>
<th>Frequency</th>
<th>Probability</th>
<th>Outcome (Value)</th>
<th>Normalized Weights</th>
<th>Contribution of Risk Factor (Value x Impact)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Creeping User</td>
<td>0.166667</td>
<td>0.156463</td>
<td>0.217687</td>
<td>0.156463</td>
<td>0.202381</td>
<td>0.142857</td>
<td>0.132653</td>
<td>0.193878</td>
<td>0.132653</td>
<td>0.178571</td>
</tr>
<tr>
<td>Requirements</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.142857</td>
<td>0.132653</td>
<td>0.193878</td>
<td>0.132653</td>
<td>0.196429</td>
</tr>
<tr>
<td>Unnecessary Features</td>
<td>0.166667</td>
<td>0.156463</td>
<td>0.217687</td>
<td>0.156463</td>
<td>0.202381</td>
<td>0.142857</td>
<td>0.132653</td>
<td>0.193878</td>
<td>0.132653</td>
<td>0.178571</td>
</tr>
<tr>
<td>Probable</td>
<td>0.166667</td>
<td>0.156463</td>
<td>0.217687</td>
<td>0.156463</td>
<td>0.202381</td>
<td>0.142857</td>
<td>0.132653</td>
<td>0.193878</td>
<td>0.132653</td>
<td>0.178571</td>
</tr>
<tr>
<td>Occasional</td>
<td>0.261905</td>
<td>0.312925</td>
<td>0.312925</td>
<td>0.251701</td>
<td>0.297619</td>
<td>0.214286</td>
<td>0.265306</td>
<td>0.214286</td>
<td>0.265306</td>
<td>0.204082</td>
</tr>
<tr>
<td>Remote</td>
<td>0.261905</td>
<td>0.312925</td>
<td>0.312925</td>
<td>0.251701</td>
<td>0.297619</td>
<td>0.214286</td>
<td>0.265306</td>
<td>0.214286</td>
<td>0.265306</td>
<td>0.204082</td>
</tr>
<tr>
<td>Improbable</td>
<td>0.238095</td>
<td>0.156463</td>
<td>0.217687</td>
<td>0.14881</td>
<td>0.14881</td>
<td>0.214286</td>
<td>0.132653</td>
<td>0.193878</td>
<td>0.132653</td>
<td>0.125</td>
</tr>
</tbody>
</table>

Table 5 Calculation of single conditional probability for Project Size with outcome to be occasional and impacts and probabilities as entered in Figure 3 and Figure 4

<table>
<thead>
<tr>
<th>Risk Factor</th>
<th>Frequent</th>
<th>Probable</th>
<th>Occasional</th>
<th>Remote</th>
<th>Improbable</th>
<th>Probability</th>
<th>Outcome (Value)</th>
<th>Normalized Weights</th>
<th>Contribution of Risk Factor (Value x Impact)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Creeping User</td>
<td>0.25 (Table 2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4/17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Requirement</td>
<td>0.142857 (Table 3)</td>
<td>3/17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unnecessary Features</td>
<td>0.261905</td>
<td>0.312925</td>
<td>0.312925</td>
<td>0.251701</td>
<td>0.297619</td>
<td>0.214286</td>
<td>0.265306</td>
<td>0.214286</td>
<td>0.265306</td>
</tr>
<tr>
<td>Probable</td>
<td>0.166667</td>
<td>0.156463</td>
<td>0.217687</td>
<td>0.156463</td>
<td>0.202381</td>
<td>0.142857</td>
<td>0.132653</td>
<td>0.193878</td>
<td>0.132653</td>
</tr>
<tr>
<td>Occasional</td>
<td>0.261905</td>
<td>0.312925</td>
<td>0.312925</td>
<td>0.251701</td>
<td>0.297619</td>
<td>0.214286</td>
<td>0.265306</td>
<td>0.214286</td>
<td>0.265306</td>
</tr>
<tr>
<td>Remote</td>
<td>0.261905</td>
<td>0.312925</td>
<td>0.312925</td>
<td>0.251701</td>
<td>0.297619</td>
<td>0.214286</td>
<td>0.265306</td>
<td>0.214286</td>
<td>0.265306</td>
</tr>
<tr>
<td>Improbable</td>
<td>0.238095</td>
<td>0.156463</td>
<td>0.217687</td>
<td>0.14881</td>
<td>0.14881</td>
<td>0.214286</td>
<td>0.132653</td>
<td>0.193878</td>
<td>0.132653</td>
</tr>
</tbody>
</table>

Table 6 Utility table for ‘Cost Overrun’

<table>
<thead>
<tr>
<th>Project Size</th>
<th>Frequent</th>
<th>Probable</th>
<th>Occasional</th>
<th>Remote</th>
<th>Improbable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>12</td>
<td>9</td>
<td>6</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>
4. Conclusion

This paper outlines the system which helps the project managers for estimating the slippage in ‘Cost’ that might occur due to various risk factors identified above. MaCO provides the following contributions for modeling the management process:

- It provides a graphical view of the problem
- It reduces large volumes of data required in the management process.
- It is capable of handling risk factors varying with every project.
- It helps in detecting cost overruns.
- It provides sensitivity analysis to examine the effect of particular uncertain events within the final outcome.

Although the response for a live cost management is encouraging but depending on the organization, some more factors could be added. As a result, this work is scalable.

References

[21].Balaram Das, “Generating Conditional Probabilities for Bayesian Networks: Easing the Knowledge Acquisition Problem”.