Australian Journal of Basic and Applied Sciences, 5(8): 832-837, 2011
ISSN 1991-8178

Development of Proficient Effort Estimation Bayesian Belief Network Through Meta-Model Conversion

[1,2]Abou Bakar Nauman [1]Romana Aziz

[1]Dept of Computer Sciences COMSATS Institute of Islamabad, Pakistan,
[2]Sarhad University of Science and Information Technology, Peshawar, Pakistan

Abstract: Effort estimation is one of the core components of project management and it is being practiced in many ways. One of the effort estimation methodologies which is becoming popular is Bayesian Belief Networks (BBNs), a probabilistic model based on Bayesian theorem. In this paper we discuss how a simple estimation model can be developed based on an existing mathematical estimation model. A mathematical model with focused scope is selected and converted into Bayesian network. Different scenarios are applied to show how this model can benefit in project management estimation as well as in risk analysis.

Key words: Effort estimation; Bayesian Networks; Methametical models

INTRODUCTION

Project management is one of the most important activity in iterative software development. Planning the resources, cost and schedule is the most important part of Project Management in software engineering discipline (William and Miller, 1978; Kurt Bittner, Ian Spence, 2006). The planning is done with the help of estimations applied on given requirements, resulting estimated effort, cost and schedule. One of the key artifacts of management is status assessment, which consists of important tasks like metrics collection, schedule tracking and risk management Walker Royce, (2005). The status assessment helps the managers to compare the current state of project with the planned states and manage rest of the project accordingly. In the water fall model the management discipline covers the whole project, however in iterative development, each iteration is treated as a mini-project, and hence the management is done at the iteration level along with project level. As the project starts the manager has to estimate total effort, and cost of the project. A good project manager will try to control the activities of the project keep the actual effort under the estimates (William and Miller, 1978; Kurt Bittner, Ian Spence, 2006).

The project manager performs estimation of the whole project. He performs a size estimation process by using some method like Function Point. By doing this he gets an estimate of size according to given requirements. He uses this size information to produce an estimate for the required effort, this is done by using some productivity factors. The productivity factors include organizational and technical factors. The effort estimate also helps to estimate the required cost and time.

Bayesian Belief Networks:

In recent years a new methodology has emerged which solves many complex problems and provides accuracy and flexibility at the same time. The use of BBN in different phases of project management is really encouraging (Norman Fenton, *et al.,* 2007; Neil, *et al.,* 2000; Hackerman, 1995) and some researchers have successfully adopted this methodology in risk management and estimation. Fenton, N. Walker Royce, (2005) developed a software defects prediction model with the help of Bayesian Networks. This model was developed in such a way that it involved many important factors of software development lifecycle and project manager is able to predict the number of defects easily. In another research Sunita Devani-chullani (Chulani *et al.,* 2000) analyzed software cost and quality models with the help of Bayesian theory. This research showed a gain in accuracy when the Bayesian theory was applied on COCOMO costing model. In another research Pendharkar, P.C *et al.,* (2005) developed a probabilistic model for software development effort. This research proved that Bayesian networks can be used for effort estimation, and the results from these BBN's are comparative with the actual estimations. The literature is rich with the encouraging results achieved with the use of Bayesian networks in project management and risk analysis.

Corresponding Author: Abou Bakar Nauman, Dept of Computer Sciences COMSATS Institute of Islamabad, Pakistan, E-mail: csit@suit.edu.pk, romana@comsats.edu.pk.

Effort estimation:

The estimation is a process of producing a result on the basis of the values of different input variables. These variables are connected with each other with different kinds of relationships. Logically these variables are connected in a network, and every variable has influence in the final output through that network. Different estimation models include different variables and compute the values so that most accurate estimation can be achieved. In all sort of models, either it is mathematical model or expert judgment based, factors of software project management are included Bohem *et al.*, (2000). The use of causal model or BBN is supported by some researchers in effort estimation and project management. The causal models represent the basic logical relationship among the factors in a particular domain. As the mathematical models also represent the relationship among different factors, the relationship can also be represented in causal network. The parametric information can be used to populate the node probability tables to complete the Bayesian belief network. This article demonstrates that the Bayesian model can also be developed from an existing meta model which is useful in developing the model quickly and achieving the benefits of Bayesian models with existing theories. One way to improve the working of existing estimation models is to transform the models to BBN

Model under Discussion:

This section provides a brief review of an estimation model described in (Richard Fairly, 1997). This article elaborates the use of COCOMO 1 model for estimation and risk management. The article (Richard Fairly, 1997) discusses this estimation model and identifies the causal relationship among different factors. This article is selected for two reasons, first it elaborates a workable risk management/estimation model, secondly the factors are minimized and their relationships are provided with statistical data elements. COCOMO 1.0 regression equation is:

$$Effort = 3.6 * (Size^{1.25}) * EAF$$

EAF is product of 15 factors; however the author identified five most significant factors.

1. Complexity
2. Time constraint
3. Storage constraint
4. Platform Experience
5. Software Tools

Size was the sixth factor which was used. The authors identify the relationship among these factors as:

"These factors are interrelated: if algorithms are complex, code size is likely to increase, if size increases, more memory and execution time will be required. With more experience on the target processor architecture and with better software tools, the team might better control the code size, execution time and memory equipment"(Richard Fairly, 1997).

The individual relationships are also identified among these factors as.

Size:

the author estimated that the size for the particular case study has these parameters.

Size of Telecom project code was to be from 9KLOC to 15KLOC, with 10KLOC as the most likely value.

Complexity:

a probability density function was identified with these parameters Normal distribution of 1.0 to 1.6, and the most likely value is 1.3.

Storage Constraint:

The storage constraint is indirectly dependent of size. From the size it can be estimated that how much memory is used and then the STORAGE coefficients can be selected according to %age of memory usage. Equation for memory usage was identified as

$$\%age\ of\ Memory = 100*[16*SIZE] / 256$$

Where as the relation between %age of memory and Storage Constraint coefficient was identified as

Memory %	<50	70	85	95
$STOR	1.00	1.06	1.21	1.56

Similarly the *Time constraint* is indirectly dependent on Size factor. A factor Percentage of time used was introduced as directly dependent on size and Time constraint is dependent on %age of Time used. Equation of %age of Time use is identified as

%age TimeUsed = 100* [(1/2) * (1/3) * (4 * SIZE)] / 10

The relation between %ageTimeUsed and Time constraint is given below.

Time Used %	<50	70	85	95
Time	1.00	1.11	1.30	1.66

Proposed Model Conversion:

We have converted the above information in collective manner and tried to build a single tool to apply the relationships given above. A BBN is constructed from this information and can be seen here, the table below provides information to show how NPT's were developed for the above mentioned factors.

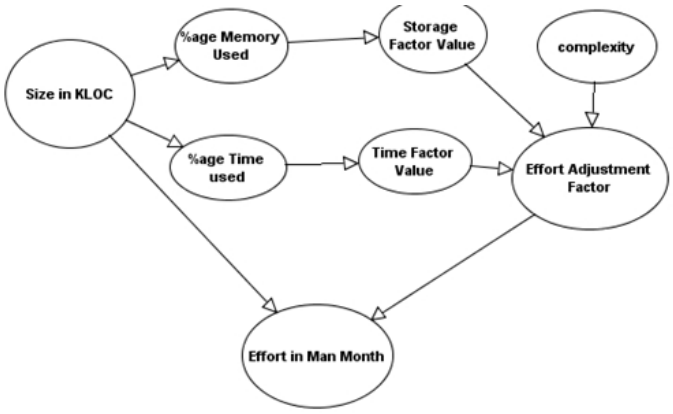

Fig. 1: Causal network for estimation

The factors of Platform Experience and Tool were not considered in this example as the intention was to show how we can translate the given information into BBN Node probability tables. This small example shows that if we have the information about the relationship of certain factors, we can easily use the information like mean values, ranges and tabular relationships to construct the Node Probability Tables.

Table 1: Detail of nodes

Node/Factor	NPT type	Mean	Ranges
Size	Normal Distribution	10	9-15
Complexity	Normal Distribution	1.3	1.0-1.6
%age Memory	Arithmetic Equation	%age of Memory = 100*[16*SIZE] / 256	0-100
Stor constraint	Manual	Conditions applied according to Table (a)	Discrete values 1.0,1.06,1.21,1.56
%age TimeUsed	Arithmetic Equation	%age TimeUsed=100* [(1/2) * (1/3) * 4* SIZE)] / 10	0-100
Time Constraint	Manual	Conditions applied according to Table (a)	Discrete values 1.0, 1.11, 1.30, 1.66
EAF	Arithmetic expression	Complexity * Time * Storage	0-4
Effort	Arithmetic Expression	Effort = 3.6 * (Size^1.25) * EAF	0-900

usability:

The conversion of mathematical model into BBN has two distinct advantages, first the model is easy to use and secondly it can give better results. In this article we are not comparing the results to show if the results are compareable or not. The scope of this article is to show that complex mathematical models can be used with ease, by converting them into a BBN model.

Simulation 1:

The simulation is based on the normal data, i.e. the information provided in the example. As the data is based on the existing information so this information will be called as Prior information. The BBN is displayed

in graph mode, i.e. the graphs are displayed for each node. The project manager from this BBN can see how the different factors e.g. size, complexity and effort are connected with each other. The curves of graph are also helpful to visualize the trend lines of each variable. The effort node is the resultant node which shows a normal distribution trend. The advantage of probabilistic approach is that in addition to the exact resultant value we can also see the other values and probabilities. As software engineering is a tough discipline and one may find an unplanned activity that needs to be done, so the manager can easily visualize what will happen if he meets some undesirable event?

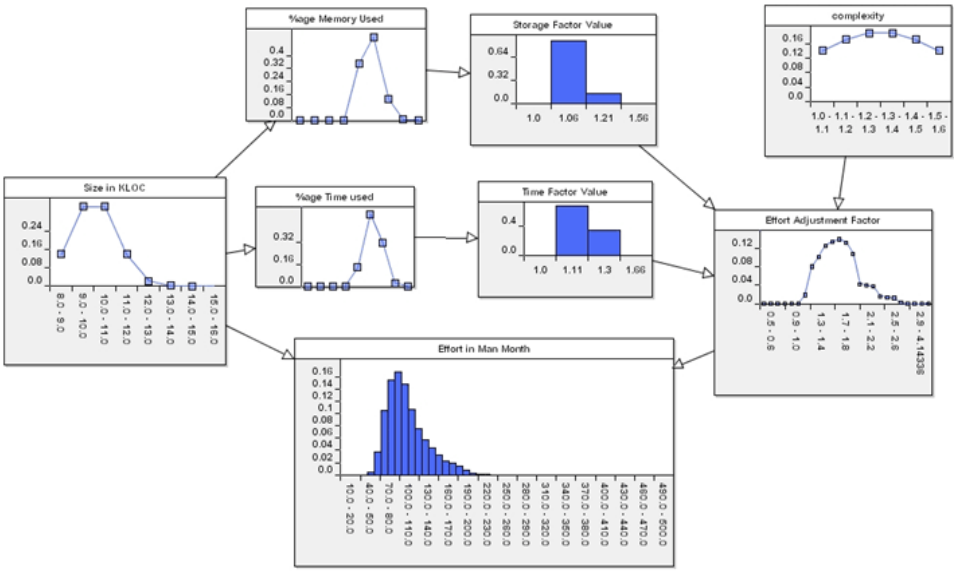

Fig. 2: Causal network for estimation

Simulation 2:

The simulation 2 is evidence of effort estimation on the basis of new information or needs. In this simulation the Size factor has a new fixed value and Complexity also has a new fixed value. The effect of these values can be observed on the effort estimate. One can see a changed trend line in comparison with the old trend line. It is to be noted that the effect of change is also on input variable like complexity which shows that the model understands the relationship between each factor, and if a change occurs in one factor the effect will be on all factors including input variable(s). The effect of change is also dependent on the information provided in NPT. Hence the change is not like a meta-model but the change is distributed on each factor with probabilistic approach.

Simulation 3:

This simulation can be called as Reverse estimation, i.e. the effort value is fixed and it is observed that what would be the values of the other factors if we need that the effort be 60. The model provides us the values which should be at the complexity and size factors if we can only provide a 60 man/month effort.

Simulation 4:

This simulation is like risk analysis approach. In this simulation the complexity is fixed at the highest value, i.e. we want to observe the effect if the complexity is accidentally or abnormally increased to its highest point. The effect of the new value can be observed at the effort node. As the size is not changed so the original distribution is observed there. The effect of complexity is also not on Time and Storage factors.
Causal network for estimation

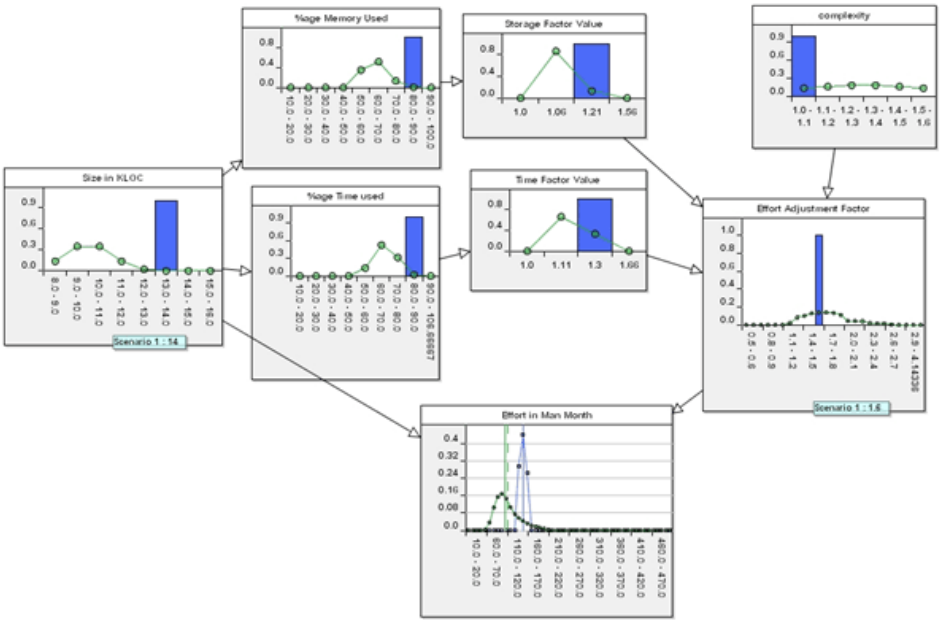

Fig. 3: Causal network for estimation

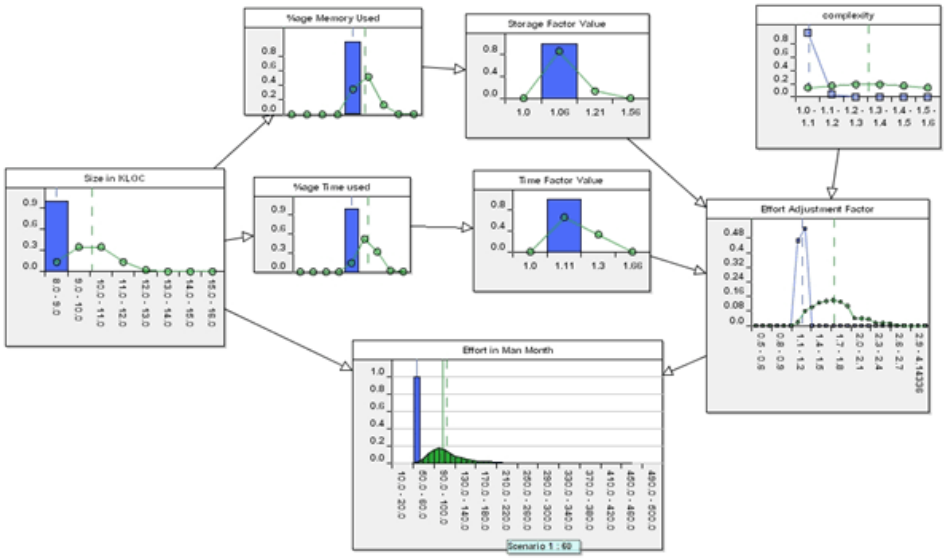

Fig. 4: Causal network for estimation

Discussion:

The mathematical models need lot of calculation to reach final values, however the BBN can easily perform calculations and help the project manager in estimation process. The above mentioned BBN based model is a good way to estimate the effort as well as apply risk management. The estimation model which was originally represented in mathematical model needed lot of calculations. The meta-model is combination of 5-6 math equations and one has to put value of each factor and calculate the outputs by solving all the equations. By meta-model we will get only one value, however by using the proposed BBN based model we can get more than one values. It is also significant that the meta-model has to be re-executed when the value of some factor is changed slightly. This puts a lot of overhead on the project manager which can effect on the estimation decision. On the other hand the project manager also has to record and arrange different results of the calculations. As the model represents an unseen network of dependencies, a non-graphical representation of the estimation model may not help to draw a complete sketch of estimation causes and effects.

The meta-model represents that the major input factors e.g. values of size and complexity, have probabilistic distributions, so it is most appropriate that these factors should be represented in probabilistic model. The model provides that the Size and Complexity has probability distribution with certain Mean and ranges, however meta-model is not able to represent these facts.

Thus the development of this kind of model in BBN is more useful and helpful to the project manager. The development of the new model can be divided in two steps, first a graphical representation is to be made and then the node probability tables are to be build, the simulations given above demonstrate both steps. A simple graphical representation of the model makes it easy for the project manager to understand the network of dependencies. Next step was to build the node probability tables. The probability distributions of Size and Complexity were given in the article, so these factors were represented by Normal distribution with given mean and ranges. The other factors were dependent on these two factors with the mathematical equations given to show their relationship. These equations were used as the arithmetic expressions for their NPTs. The resultant node Effort depends on the probabilistic values of Size and Complexity so the effort node also represents the probabilistic nature of results.

Conclusion:

The different simulations are presented here to show different scenarios. The feature of multiple scenarios is represented by the software of AgenaRisk[12], which helps a project manager to assume and apply different scenarios and execute the model to get the results with different assumptions. The results from different scenarios make is easy for project manager to make decisions considering all possible situations and values of different factors.

The BBN based model is hence very helpful in different types of situations and with the same ease the model can be used in different projects. Already some models exist for estimation but this model has some unique features. This model converts an existing Meta-model to Bayesian network, which is an encouraging phenomenon as this exercise can be done to develop BBN for different problems easily if mathematical model for that problem already exists. This model uses arithmetic expressions rather than manual entry for NPTs and makes it easy for project manager to adopt. This is a ready to use model in comparison with some models which still need a lot of effort to use.

REFERENCES

Agena, 2009. Bayesian network and simulation software, http://www.agenarisk.com/ ,accessed on 18-feb.

Bohem, B. et al., 2000. " Software development cost estimation approaches-A survey",

Chulani, Sunita. Barry Boehm; Bert Steece, 2000. "Bayesian Analysis of Empirical Software Engineering Cost Models".

Hackerman, D., 1995.. "A tutorial on learning Bayesian networks", Technical report, Microsoft Press.

Kurt Bittner, Ian Spence, 2006. "Managing Iterative Software Development Projects", Addison Wesley Professional.

Jensen, F.V., 1996 An Introduction to Bayesian Networks, UCL Press.

Norman Fenton, Neil, M., W. Marsh, P. Hearty, D. Marquez, P. Krause, R. Mishra, 2007. Predicting software defects in varying development lifecycles using Bayesian nets", Information and Software Technology, 49 Issue: 1.

Neil, Martin, Fenton, Norman; Nielson, Lars, 2000. Building large-scale Bayesian networks, Journal of Knowledge engineering review, 15 Issue: 3.

Pendharkar, P.C., G.H. Subramanian, J.A. Rodger, 2005. A Probabilistic Model for Predicting Software Development Effort, IEEE Transactions on Software Engineering, 31(7): 615-624.

Richard Fairly, 1997. "Risk Management for software Development", pages 387-400. In Richard H. Thayr, "Software engineering Project management" IEEE Computer Society. Press Los Alamitos, CA, USA.

Walker Royce, 2005.' Software project management a Unified frame work" Pearson education press.

William, B., Miller, 1978. " Fundamentals of Project management" Journal of Systems management, Nov.