International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 9 (2019) pp. 2247-2251

A Decision Tree Algorithm Based Rule Induction Framework: An Approach to Knowledge Mining for Risk Assessment and Product Estimation

Deepanshu Sharma
Research Scholar, Department of Computer Science & Applications,
Himalayan Garhwal University
Uttarakhand, India.

Dr. Harsh Kumar
Dean Research, Department of Computer Science & Applications,
Himalayan Garhwal University
Uttarakhand, India.

Abstract

This paper presents an analytical view giving insights of the need for Product Estimation and Risk Assessment. The paper focuses on the need of product estimation and assessment. The paper discusses a Case Based Reasoning and Assertion technique using a questionnaire. Researchers have compiled a list of questions that have been discussed in several researches. A simple questionnaire format has been discussed here which has been obtained from related literature reviews. We have chosen here one of the most popular algorithm for knowledge mining and classification. Several Decision Tree algorithms have been discussed here on the basis on splitting criteria and pruning methodologies. Data is trained using Decision Tree algorithm and used to obtain the results.

Keywords: Decision Tree Algorithm, Knowledge Mining, Rule Induction, Project Estimation, Risk Assessment, Critical Factors.

INTRODUCTION

Software project development is a complex process with high variance on both methodologies and objectives. Effective project planning, decision making and successful delivery rely on the date and information available on project. A project is a complex, non-routine, and one time effort limited by time, budget, resources and performance specifications design to meet customer needs [2]. Software project development industry is one of the largest manufacturing industries in the world, with $350 billion in off-the-shelf software sold every year and over $100 billion in customized code [4].

The research on the problem of software project result estimation began in 1974 by Keider SP. According to SD Time's report, only 35% of the software projects could be categorized as completed and remaining 65% projects was either categorized as outright failure or challenged [7]. The latter two stages of software projects development have plagued the industry for years.

Project estimation with respect to result is an investment, that is, there is cost associated with identifying critically affecting factors and establishing plans to mitigate those factors. Based on the definition of some literatures, projects on the basis of result can be classified in two categories [5]:

1. Successful – the project completed on time and within budget, with all the features and function originally specified.
2. Unsuccessful – the project is completed and operational, but over-budget, over the time estimate, and with fewer features, and functions than originally specified. This also includes the projects cancelled before completion or never implemented.

The research here aims at a framework for Risk assessment and project estimation. Result estimation is crucial for software development projects. It is used for project planning and control purposes during the project execution. The objectives are as follows:

1. Gathering data as per described questionnaire for different software projects.
2. Conversion of the available date into knowledge by training the date.
3. Rule induction for project estimation based on several Decision Tree algorithms.

The rule induction process can be implemented through any of the available Rule Set Generation technique. We have taken here into consideration the Decision Tree inducers. This is a powerful and popular tool for classification and result estimation. Decision trees can handle both nominal and numeric input attributes and are capable of handling datasets that may have missing values.

Framework for Project Result Estimation

The purpose of this paper is to be able to convert the subjective outputs into values based on numeric quantification. When the available or gathered date is trained

and converted into values, with the help of the data training algorithms, the output is the rule induced to be applied further.

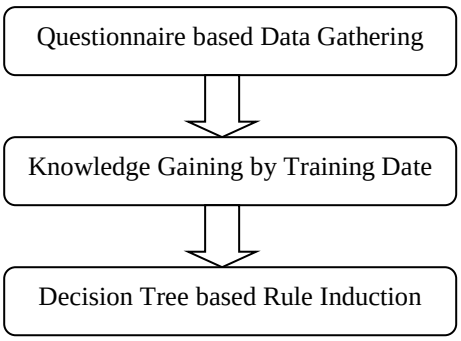

Figure 1

The collection method is crucial in obtaining reliable data. Once data is collected and understood by knowledge management system, it can be managed to systematically benefit other projects. A number of authors have suggested a list of questionnaire after an in depth analysis of critical factors. A detailed literature review of 'Project Management Practices' in Communications of IBIMA (Volume 1, 2008) by Iman Attarzadeh and Siew Hock Ow enabled us to prepare our questionnaire comprising of critically affecting factors. The appendix of this paper contains the details of the Project Management Questionnaire used in this research framework.
As a part of data gathering process to train the available data, we managed to collect data for six different projects. These included - RMP for Carnegie Mellon University, University of Minnesota RMP and live project's industrial data.

Knowledge mining and rule induction is necessary so as to devise a result from the trained data. The induction process takes input as the trained data along with the associated output as per questionnaire. As discussed above, we are implementing the decision tree rule induction based algorithms for this process of non trivial discovery of implicit information which was previously unknown but potentially interesting from trained data.

Decision Tree Algorithms based Knowledge Mining

A number of decision tree algorithm based induction methods have been used here and results were analyzed. Knowledge mining is the application of specific algorithms for the extraction of nuggets of information. Those could be in the form of patterns, correlations, estimations or rules from data.
A decision tree is a classifier which is most popular and yet powerful for result estimation purposes. All decision tree based inducer algorithm automatically construct a decision tree from a given dataset. The attractiveness of decision tree is due to the fact that, in contrast to neural networks, decision tree represents rules. Rules can readily be expressed so that humans can understand them or even can directly apply them. Decision tree is flow chart like structure where each node is either [18]:

1. Root node: a node with no incoming node.
2. Leaf node: also the decision node. It shows the value of target attribute.
3. Internal node: specifies some test to be carried out on a single attribute-value, with one branch and sub-tree for each possible outcome.

Such type of rule induction is a typical inductive approach to learn knowledge on classification. A decision tree can be used to classify an example by starting at the root of the tree and moving through it until a leaf node, which given the classification of the instance. Certain levels of strengths and weaknesses have been reported in literature for this type of classification tool.

Strengths of decision tree based rule induction:

- This type of induction is self explanatory. Since tree can be converted to rules, hence they are more comprehensible.
- Data training is easy as they can handle both numeric and nominal values as input attributes.
- Decision trees are capable of handling datasets that may have missing values and provide clear indication of which fields are most important for classification.

A few weaknesses have also been noticed and reported. Computationally, decision trees can be expensive to train. Also, the greedy characteristic of this type of classification method leads to its over-sensitivity to the training set and to noise in data.

Decision Tree Rule Induction Algorithms

Followings are some of the decision tree based rules generation algorithms:

ID3 – developed in early 1980s by J. Ross Quinlan, is Iterative Dichotomiser 3:

- Adopts a greedy non-backtracking approach where trees are constructed in a top-down recursive manner.
- It is considered to be very simple decision tree algorithm as it does not apply any pruning procedure.
- This uses a divide-and-conquer approach but cannot handle missing values [11].

CART – developed by a group of statistician in 1984, Classification and Regression Trees, describes the generation of binary trees [15].

- It also adopts the same approaches as of ID3 constructing trees using divide and conquer approach.
- Major characteristic is that it constructs binary trees, that is, each node has exactly two outgoing edges.
- An important feature of CART is that it can generate Regression trees. The leaves of such trees represent a real number instead of a class.

C4.5 – Quinlan presented C4.5 in 1993, a successor of ID3.

- It can handle numeric attributes and can induce rules from a training set which has missing values.
- It uses Gain Ratio as splitting criteria. The splitting stops when the number of instances to be split is below a certain threshold.
- C4.5 too adopts a greedy, non-backtracking approach constructing trees in top-down recursive manner [17].

EXPERIMENTS AND RESULTS

The research aims at a framework for risk assessment and product estimation on the basis of a rule induction approach using decision tree algorithms. As discussed in the framework section of this paper and as per the format given in the appendix, the project data was gathered and subsequently trained to gain the numeric quantification of the information gathered. The numeric values provide an exact basis for classification using Decision tree based algorithms. Tree based rules generated using numeric values are easier to implement and train further.

The results using recorded outcomes and data (numerical data) using above discussed algorithms is as follows:

Project ID	ID3 Output	C 4.5 Output	CART Output	Average	Results
P01	2.413596	2.397082	2.427093	2.412590333	Success
P02	2.298346	2.289981	2.299136	2.295821	Success
P03	2.394186	2.385185	2.399216	2.392862333	Failure
P04	2.627864	2.615679	2.609137	2.61756	Failure
P05	2.186754	2.179613	2.189342	2.185236333	Success
P06	2.487891	2.479236	2.498234	2.488453667	Success

Figure 2

The following figure shows the Data import process of the Rapid Miner 5 for the generation of rules from the mapped numeric values to their respective recorded results.

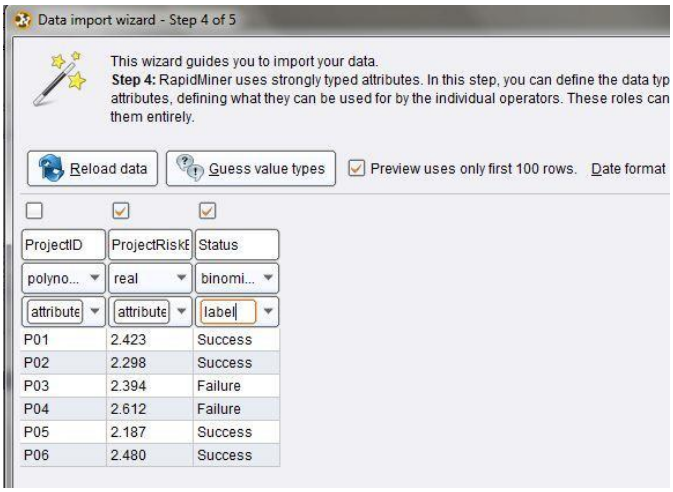

Figure 3

Once the data has been imported with the obtained average values, the rules are then generated using the decision tree rule induction technique. The rules formed were obtained when the values were mapped with their recorded project results and applied with decision tree rule inducers. The obtained rules are as follows:

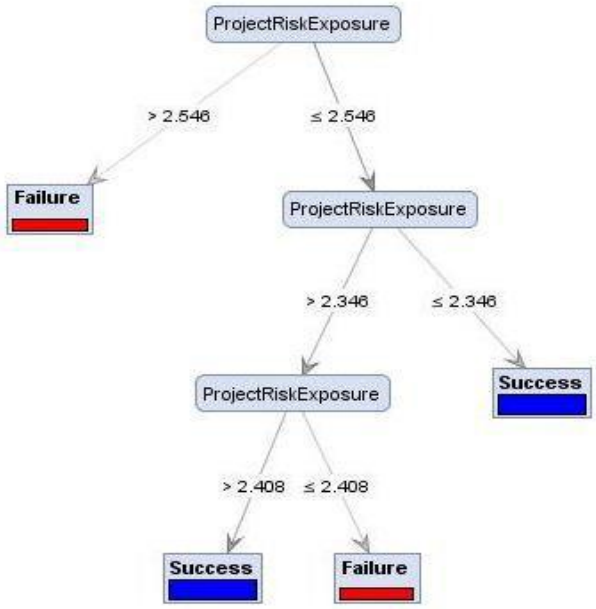

Figure 4

There has been a clear text annotation of the generated rules as well. It is one of the major advantages of C4.5 which can handle and train numeric data. Text annotation of the values gives an easy hand towards decision making and product estimation.

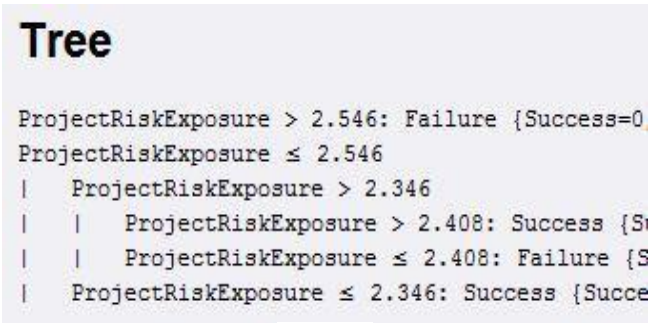

Figure 5

Comparison between Algorithms – ID3 vs. C4.5 vs. CART

ID3 algorithm selects the best attribute based on the concept of Entropy and Information Gain for developing the tree.

C4.5 algorithm acts similar to ID3 but improves a few of ID3 behaviors [18]:

- A possibility to use continuous data.
- Ability to use attributes with different weights.
- Using and training unknown /missing values.
- Pruning the tree after being created.

The C4.5 algorithm differs in several respects from CART, for example [12]:

- CART uses the Gini diversity index for classifying tests, while C4.5 uses criteria based on the information.
- CART looks for alternative tests that approximate the results when tested attribute has an missing or unknown value.

CONCLUSION

The framework aims at studying the relationship between project result and data and thus inducing a rule set which gives an approach for product estimation.

Decision trees simply train data so that it can be presented quickly enough to a non-specialist audience without getting lost in difficult to understand mathematical formulations.

In this article, we wanted to focus on the key elements of their construction from a set of data, and then we presented the algorithm ID3 and C4.5 that respond to these specifications. And we did compare ID3, C4.5 and CART, which led us to confirm that the most powerful and preferred method in machine learning is certainly C4.5.

C4.5 algorithm, as discussed above, gave the output decision tree rule taking Information Gain as the splitting criteria and a set value of confidence of 0.25.

A Road Ahead

The basic requirement for enhancing the rules is the availability of more data. The process of classification and product result estimation can be made much more reliable when a large amount of data is trained and rules are obtained.

APPENDIX A

A detailed literature review of 'Project Management Practices' in Communications of IBIMA (Volume 1, 2008) by Iman and Siew enabled us to prepare our questionnaire comprising of critically affecting factors [3].
Critically affecting factors considered in this approach are:

1. Requirements are clearly defined and requirement changes are handled precisely/correctly.
2. User involvement is adequate and continuous.
3. Management skills of project leader and effective project management methodology.
4. Accuracy of time incurred estimation.
5. Accuracy of cost invested and estimation.
6. Appropriate and adequate allocation of resources, both technical and personnel.

APPENDIX B

Project Management Questionnaire is as follows:
Section A: Team Profile

1. Project Title
2. Team Members.

Section B: Team Rating
Please indicate the importance of the following six critical factors that contributed to the success or resulted in the failure of your project.

1	2	3	4	5
Very Poor	Poor	Average	Good	Very Good

Table 1

Project assessment/estimation criteria:

No.	Critical Factor	Team Member Rating
1	Requirement clearly defined and changes handled accurately.	
2	User involvement is adequate and continuous.	
3	Management skills of leader and methodology.	
4	Time estimation accuracy.	
5	Cost estimation accuracy.	
6	Appropriate and adequate allocation of resources.	

Table 2

Acknowledgement
We would sincerely like to thank our IT industry peers who helped us and responded to our queries related to the live and under development software project. Sincere thanks to the Project leader at a HCL Technologies' beverages and a pharmaceutical related (offshore development center) IT project. We would also like to thank the Project leader at Capgemini India Limited's healthcare related IT project. Sincere thanks to Carnegie Mellon University and University of Minnesota for sharing RMP related data and responding to the queries.

REFERENCES

[1] Xiaohong Shan, GuoRui Jiang, Tiyan Huang, Beijing University of Technology, China "A Framework of estimating software project success potential based on association rule mining", IEEE, 2009.

[2] Tharwon Arnuphaptrairong "Software Project Risks: Evidence from the Literature Survey", International MultiConference of Engineers and Computer Scientists 2011 Vol I, IMECS March 16-18, 2011, Hong Kong.

[3] Iman Attarzadeh, Siew Hock Ow, Department of Software Engineering, Faculty of Computer Science & Information Technology, University of Malaya, Kuala Lumpur, Malaysia "Project Management Practices".

[4] Willy Herroelen and Roel Leus "Software Project Scheduling Under Uncertainties" European Journal of Operational Research (Pages 289-306), September 2005 (Vol. 165, Issue II).

[5] AH Yousef, A Gamal, A Warda, M Mahmoud, "Software Projects success factors identification using Data Mining", IEEE 2006, pp.447-453.

[6] Ronal P. Higuera, Yacov Y. Haimes, "Software Risk Management", Technical Report, SEI-96-TR-012, June, 1996.

[7] Subhas C. Misra, Vinod Kumar, Uma Kumar, "Different techniques for risk management in software engineering: a review", Carleton university, ASAC, 2006.

[8] Abdullah Al Murad Chowdhary and Shamsul Arefeen, "Software Project management: importance and Practices", IJCIT, Vol. 02 Issue 01, 2011.

[9] T. Addision and S. Vallabh, S. "Controlling Software Project Risks – An Empirical Study of Method used by Experience Project Managers", Proceeding of SAICSIT 2002, pp.128-140.

[10] Lior Rokach and Oded Maimon, "Decision trees", Department of Industrial Engineering, Tel-Aviv University.

[11] Ding-An Chiang, Wei Chen, Yi Fan Wang, Lain Jinn Hwang, "Rules Generation form the Decision Trees", Department of Information Engineering, Tamkang University.

[12] Andrew W. Moore, "Decision Trees", School of Computer Science, Carnegie Mellon University.

[13] J. R. Quinlan, "Generating Production Rules from Decision Trees", Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, USA.

[14] Murthy S. K., Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey. Data Mining and Knowledge Discovery, 2(4):345-389, 1998.

[15] Alsabti K., Ranka S. and Singh V., CLOUDS: A Decision Tree Classifier for Large Datasets, Conference on Knowledge Discovery and Data Mining (KDD-98), August 1998.

[16] Breiman L., Friedman J., Olshen R., and Stone C.. Classification and Regression Trees. Wadsworth Int. Group, 1984.

[17] Apte, C, Thomas J. Watson, "Data mining: an industrial research perspective", Computational Science and Engineering, IEEE 1997 Volume: 4, Issue: 2, page(s): 6 – 9.

[18] Kearns M. and Mansour Y., On the boosting ability of top-down decision tree learning algorithms. Journal of Computer and Systems Sciences, 58(1): 109-128, 1999.

[19] Lim X., Loh W.Y., and Shih X., A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms Machine Learning 40:203-228, 2000.

[20] Phillips, J. Project management professional study guide. 2nd Edition, McGraw Hill, California, 2006.

[21] Brock, S. A Balanced Approach to IT Project Management, Proceedings of SAICSIT, ACM 2003, pp. 2 –10.