Institute of Internet and Intelligent Technologies
Vilnius Gediminas Technical University
Saulėtekio al. 11, 10223 Vilnius, Lithuania
http://www.isarc2008.vgtu.lt/

The 25th International Symposium on Automation and Robotics in Construction

June 26–29, 2008

ISARC-2008

DYNAMIC PREDICTION OF PROJECT SUCCESS USING EVOLUTIONARY SUPPORT VECTOR MACHINE INFERENCE MODEL

Min-Yuan Cheng
Department of Construction Engineering, National Taiwan University Of Science And Technology
#43,Sec.4,Keelung Rd.,Taipei,106,Taiwan,R.O.C
myc@mail.ntust.edu.tw

Yu-Wei Wu
Department of Construction Engineering, National Taiwan University Of Science And Technology
#43,Sec.4,Keelung Rd.,Taipei,106,Taiwan,R.O.C
D9305503@mail.ntust.edu.tw

ABSTRACT

The purpose of construction management is to successfully accomplish projects, which requires a continuous monitoring and control procedure. To dynamically predict project success, this research proposes an Evolutionary Support Vector Machine Inference Model (ESIM). ESIM is developed based on a hybrid approach that fuses support vector machine (SVM) and fast messy genetic algorithm (fmGA). SVM is primarily concerned with learning and curve fitting; and fmGA with optimization. Furthermore, the model integrates the process of continuous assessment of project performance (CAPP) to dynamically select factors that influence project success. CAPP was developed to identify continuous variables that have the ability for predicting project outcome. Training and test patterns are collected from CAPP database that contains 46 construction projects. These projects are real data collected by Russell from the 16 representative Construction Industry Institute (CII) member companies. K-means clustering was employed to conduct an unsupervised clustering to extract similar cases for comparison. Results show that ESIM can successfully predict the project success.

KEYWORDS

Project management, Predictions, Support vector machines, fast messy genetic algorithm

1. INSTRUCTIONS

In the construction industry, construction project success infers that certain expectations of participants, including owners, planners, designers, architects, contractors, and operators, are fulfilled. Once a construction project has been bid, the prime contract is typically subdivided into multiple subcontracts. Large numbers of participants are, therefore, involved in the project planning and implementation phases. Expectations can only be met by conducting a comprehensive analysis of participants [1]. The measurements of project success in the construction industry are cost, schedule, performance, and safety. Hughes [2] developed a Construction Project Success Survey instrument to identify important success metrics before the start of a project, and to evaluate the level of success achieved at project completion. The measuring metrics include objective (such as

cost, schedule, performance, and safety) and subjective considerations. Griffith [3] developed an objective metrics that comprised of four variables: budget achievement, schedule achievement, design capacity, and plant utilization. The authors discovered that despite the complexities involved in measuring project success, a measurement can be developed based on objective project performance.

The construction industry is replete with myriad uncertainties that make management exceedingly complex. Factors for success, therefore, vary from project to project. Although human experts can often achieve a satisfactory project outcome, shortfalls nearly always occur due to managers failing to take all relevant factors into consideration and lacking access to all relevant information.

Various scientific and engineering fields have been paying increasing attention in recent years to fusing different artificial intelligence (AI) paradigms. A number of studies have demonstrated that performances achieved by fusing different AI techniques are better than those achieved by employing a single conventional technique [4]. Two tools, the fast messy genetic algorithms (fmGA) and support vector machine (SVM) have been successfully applied to solve various problems in construction management. Considering the characteristics and merits of each, this paper combines the two to propose an Evolutionary Support Vector Machine Inference Model (ESIM). In the ESIM, the SVM is primarily employed to address learning and curve fitting, while fmGA addresses optimization. This model was developed to achieve the fittest C and gamma parameters with minimal prediction error.

An appreciation of critical factors is crucial to assess the requirements of project success and to achieve successfully project objectives. Statistical methods represent a basic approach to identify significant factors from historical data or questionnaire results. However, the dynamic nature of critical factors means that changes in project conditions must be monitored continuously. The Construction Industry Institute [5] cooperated with the University of Wisconsin at Madison to develop a prediction software tool, named Continuous Assessment of Project Performance (CAPP) [6], which allows managers to identify significant factors continuously and dynamically.

In this study, CAPP software is employed to determine significant factors for project success and AI approaches are used to assess project success. Project managers can use the model to predict the degree of success of a new project, allowing managers to enhance their effective control over projects and prevent problems. The remaining sections of this paper include Section 2: a introduction of AI approaches which comprehend K-mean clustering and Evolutionary Support Vector Machine Inference Model with fmGA, and SVM involved; Section 3: significant factors for project success are determined using CAPP software and AI approaches apply to project success prediction; Section 4: conclusions are described.

2. ARTIFICIAL INTELLIGENCE APPROACHES

2.1. K-means Clustering

Many algorithms are able to identify specific domains. K-means clustering is a simple and fast approach to data clustering that starts with k centroids (seeds), which are usually generated randomly. Each data set (sample) is assigned to the cluster with closer centroid of the Euclidean distance measurement. It is customary to set a threshold on iteration numbers to prevent excessive calculation times. After a number of iteration steps, every clustering feature can be determined. As desired number of clusters can be set as a limitation for target convergence, perfect convergence cannot be guaranteed. K-means usually converges in practical applications, especially in pattern recognition problems. K-means clustering is widely and commonly employed owing to its simplicity, although it does present some inherent drawbacks such as a fixed setting for the optimal solution or time consumption.

2.2. Fast Messy Genetic Algorithms (fmGA)

The fmGA, developed by Goldberg et al. [7], can find efficiently optimal solutions for large-scale permutation problems. The fmGA-based approach is known for its flexibility in allowing hybridization with other methodologies to obtain better solutions.

2.3. Support Vector Machines (SVM)

The theory that underlies support vector machines represents a new statistical technique that has drawn much attention in recent years. This learning theory may be seen as an alternative training technique for polynomial, radial basis function and multi-layer percept classifiers. SVM are based on the structural risk minimization (SRM) induction principle, which aims to restrict the generalization error (rather than the mean square error) to certain defined bounds. In many applications, SVM have been shown to deliver higher performance than traditional learning machines and have been introduced as powerful tools to solve classification and regression problems.

2.4. Evolutionary Support Vector Machine Inference Model (ESIM)

Support vector machines and fast messy genetic algorithms represent recently developed AI paradigms. SVM were first suggested by Vapnik [8] and have recently been applied to a range of problems that include pattern recognition, bioinformatics, and text categorization. An SVM classifies data with different class labels by determining a set of support vectors that are members of the set of training inputs that outline a hyper plane in a feature space. It provides a generic mechanism that fits the hyper plane surface to the training data using a kernel function. The user may select a kernel function (e.g. linear, polynomial, or sigmoid) for the SVM during the training process, which identifies support vectors along the function surface. Using SVM presents users with the problem of how to set optimal kernel parameters. Therefore, obtaining SVM parameters must occur simultaneously. Proper parameter settings can improve SVM prediction accuracy, with parameters that should be optimized including penalty parameter C and kernel function parameters such as the gamma of the radial basis function (RBF) kernel. In designing an SVM, one must choose a kernel function, set kernel parameters and determine a soft margin constant C (penalty parameter). The Grid algorithm is an alternative to finding the best C and gamma when using the RBF kernel function. However, this method is time consuming and does not perform well [9]. Fast messy genetic algorithms were developed by Goldberg et al. in 1993. Unlike the well-known simple genetic algorithm (sGA), which uses fixed length strings to represent possible solutions, fmGA applies messy chromosomes to form strings of various lengths. Its ability to identify efficiently optimal solutions for large-scale permutation problems gives fmGA the potential to generate SVM parameters C and gamma simultaneously. Considering the characteristics and merits of each, this paper combines the two to propose an Evolutionary Support Vector Machine Inference Model (ESIM). In the ESIM, the SVM is primarily employed to address learning and curve fitting, while fmGA addresses optimization. This model was developed to achieve the fittest C and gamma parameters with minimal prediction error. The structure of ESIM is shown in Fig. 1.

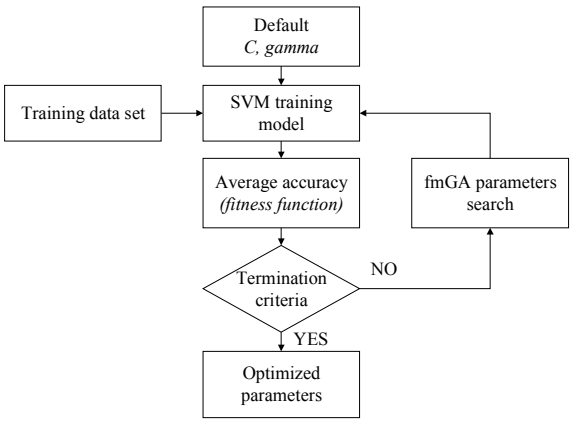

Figure 1. Structure of ESIM

3. PROJECT SUCCESS PREDICTION MODEL

Specific processes and methods used to implement ESIM are summarized in Fig. 2. Referring to Fig. 2, the blocks on the left hand side are the procedures used to implement the model. The blocks on the right hand side are detailed methods and attributes concerned with execution of the tasks on the left hand side.

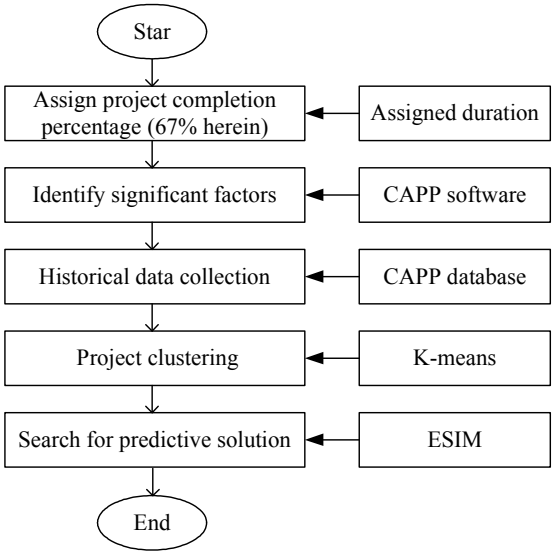

Figure 2. Model application process

3.1. Assign Project Completion Percentage

Using CAPP software, 54 historical construction projects were collected from 17 CII member companies and analyzed using 76 variables. Current project progress and the level of significance of each factor should be identified first using CAPP software. Significant factors vary during project stages. To identify factors, a completion percentage should be selected for this analysis. For purposes of research in this paper, project progress is set at 67% complete.

3.2. Identify significant factors

A threshold level of significance should be selected to identify factors of greatest significance. CAPP recommends that an attached alpha below 0.1 identifies a referenced factor. In this paper, a threshold for the alpha was set at less than 0.025 in order to reduce the number of identified factors. According to project performance, CAPP defined the four degrees for project success of "successful", "on time or on budget", "less-than-successful", and "disastrous" [6]. Basing on this definition, this paper assigned four quantitative values for project success linearly (see Table 1).

Table 1. Quantitative Project performance

Project performance	Value
Successful	1.000
On time or on budget	0.667
Less-than-successful	0.333
Disastrous	0.000

Sequentially, CAPP can be employed to calculate significant factors. Factors can be analyzed using CAPP software (see Fig. 3).

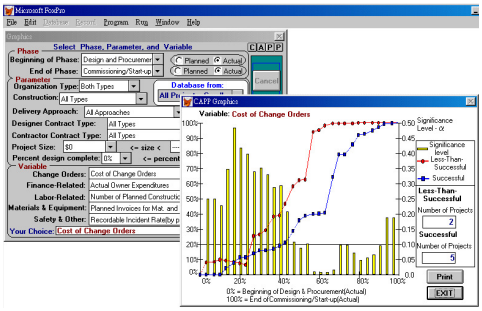

Figure 3. CAPP Graphics for Cost of Change Orders

Histogram in CAPP Graphics shows level of significance, denoting high effectiveness at low quantity. With project progress set at 67%, the value of histogram is about 0.02 (below the threshold 0.025), indicating that the factor "cost of change orders" is identified as a significant factor in this study. Eleven factors significant to project success were identified in total (see Table 2).

Table 2. Time-dependent factors identified by CAPP

Factors	Column I.D. in CAPP	Analyzed Significant Level
1. Actual design % complete	C5_16	0.01
2. Actual owner expenditures	C3_10	0.01
3. Invoiced construction costs	C2_14	0.02
4. Designer planned effort hours	C2_13	0.01
5. Actual invoices for material and equipment	C3_28	0.01
6. Paid construction costs	C3_14	0.01
7. Cost of owner project commitments	C2_24	0.01
8. Recordable incident rate (by period)	C2_38	0.01
9. Cost of change orders	C2_17	0.02

Factors	Column I.D. in CAPP	Analyzed Significant Level
10. Quantity of change orders	C3_17	0.01
11. Actual overtime work	C3_41	0.02

3.3. Historical data collection

Forty-six of the 54 valid projects in the CAPP database met the criterion that all eleven factor values are non-null. Forty-two of the 46 were selected for training (see reference [6]), leaving four valid projects available for testing (see Fig. 4).

No.	Inputs											Output
	C5_16	C3_10	C2_14	C2_13	C3_28	C3_14	C2_24	C2_38	C2_17	C3_17	C3_41	
1.	0.000	0.000	0.118	0.150	0.154	0.135	0.000	0.000	0.251	0.456	0.961	0.000
2.	0.074	0.841	0.657	0.079	0.622	0.000	0.000	0.249	0.000	0.000	0.000	1.000
3.	0.000	0.277	0.166	0.000	0.000	0.000	0.000	0.941	0.138	0.200	0.635	0.333
4.	0.000	0.807	0.585	0.000	0.000	0.000	0.000	0.000	0.081	0.211	0.000	0.667

Figure 4. Testing data from the CAPP database

3.4. Project Clustering

With CAPP's kind permission, 56 projects in CAPP database were employed in this study. Forty-six projects fulfilled our criteria and were treated as raw data for project success learning. Of the 46 data sets, 42 were treated as training data and 4 were assigned as testing data for ESIM learning. Model accuracy varies in correspondence with the dynamic factors of influence on project success. Testing RMSE was 0.1781, respectively. Detailed training results are shown in Table 3. While result trends are positive, they are not categorized well to determine project success, identify Less-than-successful projects (project performance=0.333), or determine on-time or on-budget (project performance=0.667) projects. Additional strategies should be employed to overcome such deficiencies.

Table 3. Results for Project Success Assessment without Prepared Data Clustering

Testing Case	Predicted Output	Desired Output	Training RMSE
1	0.0978	0.0000	0.1781
2	1.0527	1.0000	
3	0.6347	0.3330	
4	0.8199	0.6670	

K-means clustering is a multi-variable analysis data clustering method that aggregates similar data and identifies discrepancies between clustered categories. CAPP database data used in this study were gathered from different construction companies and vary in terms of project attributes (e.g., type of construction, cost, procurement approaches, etc.) To improve assessment accuracy, K-means clustering was used prior to ESIM learning to collate training data sets that were most similar to the assessment target. SPSS, a commercial statistics software package, was the tool used to conduct K-means clustering analysis for this purpose. After the number of clusters been set, each cluster center iterated toward the fittest location by Euclidean distance measurement. The number of clusters was chosen as 2 to represent positive and negative quality. The four testing data (CS1, CS2, CS3, and CS4) were treated as clustering targets respectively. For CS2, K-means clustering was employed for the 42 training data and CS2. The clustering results are shown in Table 4, in which the CS2 is attached to cluster 2, where there are 17 data sets in this cluster. Similarly, there are 25 training cases for CS1, 26 for CS3, and 17 for CS4. In other words, for each new project assessment, K-means clustering was applied to the assembly of the 42 training projects as well as the new one with 2 sets of clusters having been set. Thus, SPSS generated 2 cluster centers. Finally, data sets in which the new project had been clustered were treated as training data (part of 42 training projects, without the new one) for sequential ESIM learning to assess new project performance. The reason for setting 2 sets of clusters was to avoid having only a small number of projects for ESIM learning. Therefore, if the data pool is large enough in other studies, the selected number of clusters could be increased. In summary, time-dependent factors were not the only factors that

changed dynamically with CAPP analysis. Training data sets also varied for different project performance assessment targets with K-means clustering.

Table 4. Results of K-means Clustering

Variable	Initial Cluster Centers		Final Cluster Centers	
	Cluster		Cluster	
	1	2	1	2
C5_16	1.000	0.000	0.087	0.122
C3_10	0.000	1.000	0.088	0.472
C2_14	0.021	0.327	0.055	0.379
C2_13	0.233	0.000	0.102	0.090
C3_28	0.021	0.675	0.028	0.292
C3_14	0.027	0.419	0.077	0.171
C2_24	0.000	0.847	0.026	0.283
C2_38	0.864	0.000	0.077	0.033
C2_17	0.145	0.000	0.068	0.180
C3_17	0.000	0.000	0.104	0.161
C3_41	0.021	0.000	0.100	0.028

Notations:
1. Convergence achieved due to no or minimal distance change. The maximum distance by which any center has changed is 0.000. The current iteration is 3. The minimum distance between initial centers is 2.053.
2. There were 43 valid cases. Of which, 26 cases were in cluster 1 and 17 cases were in cluster 2. No cases were missing.

Table 5 Comparisons for K-means Clustering of Performance Assessment Results

	Testing Case	Predicted Output	Desired Output	Training RMSE
Without K-means Clustering	1	0.0978	0.0000	0.1781
	2	1.0527	1.0000	
	3	0.6347	0.3330	
	4	0.8199	0.6670	
With K-means Clustering	1, CS1	0.0083	0.0000	0.0071
	2, CS2	0.9932	1.0000	
	3, CS3	0.3237	0.3330	
	4, CS4	0.6678	0.6670	

3.5. Search for predictive solution

After K-means clustering analysis, ESIM project performance learning for a particular case can follow sequentially. Results of RMSE are listed in Table 5, with results (not using prepared data clustering) shown in Table 3. Results show that K-means clustering does indeed improve project performance assessment. Therefore the project success assessment processes have been demonstrated as representing a reasonable, feasible, and effective approach.

4. CONCLUSION

This paper proposes a model for assessing project success using AI approaches that employ fast messy genetic algorithm, support vector machine, and K-means clustering. The two commercial software packages used include CAPP for project access and SPSS for data clustering. The results achieved in this paper can be summarized as follows:

1. Using CII's copyrighted CAPP software, the time-dependent factors that dynamically influence project performance can be managed in order to achieve precise project success assessment.
2. Although data in the CAPP database are representative of typical construction projects, their features vary widely. Extracting similar historical cases using K-means clustering can improve prediction accuracy. This study performs clustering using SPSS software.
3. The uncertain information and complex mapping in project performance assessment are conducted using ESIM. ESIM uses SVM to perform input-output mapping and fmGA to achieve global optimization. As its feasibility for project performance assessment has been demonstrated, therefore ESIM is proposed herein.
4. Project assessment helps managers to make strategies in a time efficient manner and take correct actions to achieve final project success. With the proposed model, dynamic project performance assessment can be achieved using CAPP, SPSS, and ESIM.

5. ACKNOWLEDGMENT

The authors would like to thank the Construction Industry Institute's kind permission to use, analyze, show, and extract data from their CAPP database.

REFERENCES

[1] Sanvido, V., Grobler, F., Parfitt, K., Guvenis, M., and Coyle, M. (1992). "Critical success factors for construction projects." Journal of Construction Engineering and Management, 118(1) 94-111.

[2] Hughes, S. W., Tippett, D. D. and Thomas, W. K. (2004). "Measuring project success in the construction industry." Engineering Management Journal, 16(3) 31-37.

[3] Griffith, A. F., Gibson, G. E., Jr., Hamilton, M. R., Tortora, A. L., and Wilson, C. T. (1999). "Project success index for capital facility construction projects." Journal of Performance of Constructed Facilities, 13(1) 39-45.

[4] Yang, J.B. and N.J. Yau, "Integrating case-based reasoning and expert system techniques for solving experience-oriented problems," Journal of the Chinese Institute of Engineers, vol. 23(1), pp. 83–95, 2000.

[5] CII (1996). Predictive tools: closing the performance gap, Research Summary, RS107-1, The Construction Industry Institute, Austin, Texas.

[6] Russell, J. S., Jaselskis, E. J., and Lawrence, S. P. (1997). "Continuous Assessment of Project Performance." Journal of Construction Engineering and Management, ASCE, 123(1), 64–71.

[7] D.E. Goldberg, K. Deb, H. Kaegupta, G. Harik, "Rapid, accurate optimization of difficult problems using fast messy genetic algorithms," Proceedings of the Fifth International Conference on Genetic Algorithms, pp. 56– 64, 1993

[8] C.-L. Huang, C.-J. Wang. "A GA-based feature selection and parameters optimization for support vector machines," Expert Systems with Applications, vol. 31, pp. 231-240, 2006.

[9] Hsu, C. W., Lin, C. J. "A simple decomposition method for support vector machine," Machine Learning, vol. 46(1–3), pp. 219–314, 2002.