Using Least Squares Support Vector Machines to the Product Cost Estimation

Shi-Gan Deng[1] and Tsung-Han Yeh[2*]

[1]Department of Power Vehicle and Systems Engineering, Chung-Cheng Institute of Technology, National Defense University
[2] School of National Defense Science, Chung-Cheng Institute of Technology, National Defense University

ABSTRACT

This research makes the first attempt to apply a novel machine learning method, the least squares support vector machines (LS-SVM), to solving product cost estimation problems in the product life cycle. Four real product cost estimation problems, proposed in previous studies, are used and the estimation performance of LS-SVM model evaluated. These cases include estimations of the costs of carbon steel pipe material, steel pipe bending, pressure vessel manufacturing, and pump purchasing. The performance of numerous cost estimation models, including regression analysis, neural networks, and support vector regression, established in the previous articles, are compared with that of the LS-SVM model. The test results verified that the LS-SVM model can provide more accurate estimation performance and outperforms other methods. The results of this analysis can serve as a useful reference for product cost planning and control in industries.

Keywords: product cost estimation, least squares support vector machines, neural networks, regression analysis, support vector regression.

運用最小平方支援向量機於產品成本預測之研究

鄧世剛[1]　葉宗翰[2*]

[1] 國防大學理工學院動力暨系統工程學系
[2] 國防大學理工學院國防科學研究所

摘要

本研究首次嘗試運用一個新穎的機器學習方法 – 最小平方支援向量機，針對產品生命壽期成本預測問題進行探討。在案例探討部分係使用過去學者所提出之四個實際產品成本預測案例，進而驗證最小平方支援向量機之可行性以及準確性。四個案例分別包含碳鋼管材料成本預測、鋼管製造成本預測、高壓管製造成本預測以及幫浦採購成本預測等問題。同時使用過去研究提出之類神經網路、迴歸分析以及支援向量迴歸預測結果，與最小平方支援向量之進行預測績效比較。最終透過測試結果顯示，最小平方支援向量機能相較於其他方法提供更準確的預測能力。本研究之結果可提供未來產業作為進行成本規劃與控管之參考依據。

關鍵詞：產品成本預測，最小平方支援向量機，類神經網路，迴歸分析法，支援向量迴歸

文稿收件日期 99.3.24;文稿修正後接受日期 99.11.11;*通訊作者
Manuscript received March 24, 2010; revised November 11, 2010; * Corresponding author

I. INTRODUCTION

An accurate and rapidly cost estimation model is critical for the life cycle of industrial products in the early design and production phases. A sufficiently accurate product cost estimation model can help achieve good decision making. Industries compete based primarily on product quality, product cost, and on-time product delivery [1-4]. The ability to give rapid price quotations and have a renewable estimation model is also important in highly changeable competitive environments [5]. Numerous studies indicate that, although the costs in the early design and plan phases account for a small proportion of the total cost, they affect 70-80% of the total cost, and the early phases have a significant impact on the overall life cycle cost [1, 4, 6]. As a result, cost control and planning must be considered during the early phases of the life cycle.

Cost estimation is the estimation of costs incurred in producing a product, are made before actual production begin, and predict how much products will cost; it is one of the main inputs for economic evaluation [7]. The main stages of product life cycle cost are the design, manufacturing, marketing and after sale, and disposal and recycling stages. The design stage includes engineering design, drawing, and design modification costs. Material and production costs are contained in the manufacturing stage [8]. Traditionally, the production industry is based on its project life cycle on the design-to-performance manufacturing concept, which focused on the airframe performance as the main consideration in the airframe life cycle. However, this approach was extremely costly and wasteful. Design-to-cost (DTC), forms an integral part of a cost-oriented planning and controlling policy that focuses on the relationship between design decisions and the resulting costs of manufacturing, supports, and operations. The DTC is based on limiting costs and making trade-offs among product performance, manufacturing schedule, and life-cycle costs, with the ultimate goal of achieving project objectives [9-10]. A prominent example of successful application of DTC is the NASA mission to Mars. In 1976, NASA's two Viking-Mars Lander missions cost $3 billion to develop. In 1997, the Pathfinder-Rover mission cost just $175 million, an incredible 94% reduction [11]. This significant difference could be attributed to the change in organizational policy from a design-to-performance orientation to a design-to-cost orientation. Following this successful case, the concept of the DTC planning concept became a global industrial trend, replacing the highly wasteful practices of design-to-performance methods. Before implement the DTC cost planning concept, the accurate cost estimation method is essential.

Previously, typical cost estimation methods included human decision making and statistical regression analysis techniques. The former includes top down and bottom up cost planning methods [11]. These two methods mostly involve estimation based on human experience; in both methods, the projected costs are inevitably less than the actual costs, which leads to many projects not reaching successful completion. Statistical regression analysis techniques consist of one or more functions or cost estimation relationships, which are developed by applying regression analysis to historical information about projects [10, 12]. These traditional cost estimation methods can be improved for accurate estimation performance. Some researchers are attempting to use novel machine learning techniques to overcome the drawbacks of existing systems. The most popular method of machine learning is called neural networks (NN). Numerous studies have verified that NN outperform regression analysis method when facing highly complex and nonlinear problems. In 1995, the support vector machines (SVM) was developed by Vapnik. The SVM solution is a global optimal, and fewer parameters are determined in the modeling process. These advantages overcome the drawbacks of neural networks.

In the past research, we have used back-propagation support vector regression (SVR), neural networks (BPN) and regression analysis methods to the airframe structural parts design cost estimation problem, and verified the feasibility and accuracy of machine learning methods. And in this research, we make the first

attempt to use a novel machine learning method —the least squares support vector machines (LS-SVM) — to solve the cost estimation modeling problem.

The main purpose of this research was to apply least squares support vector machines (LS-SVM) to the product cost estimation problem. We consider four real production cost estimation case studies proposed in previous papers, and establishing an LS-SVM model based on fair cost estimation parameters and condition. The results show that LS-SVM outperforms other methods, and provides better accuracy. Thus, this study provides a better solution to the problem of cost estimation.

The structure of this paper is organized as follows: Section 1 introduces the fundamental concepts of project cost planning and estimation. Section 2 presents a literature review of cost estimation methods and related applications. Section 3 introduces the LS-SVM theory and modeling procedure. Section 4 presents four case studies for production cost estimation and analysis results. Finally, Section 5 provides conclusions and suggestions.

II. COST ESTIMATION METHODS REVIEW

Cost estimation methods can be divided into qualitative and quantitative methods. Qualitative methods include those based on human judgment and heuristic rules. Quantitative methods include statistical parametric and machine learning methods [13]. This research focuses on the introduction of a quantitative cost estimation method.

2.1 Statistical Regression Analysis

Statistical regression analysis has been widely used for cost estimation in many fields since the 1970s. The main advantage of regression analysis is that it can describe relationships between variables and target values. The variables generally express product features that influence the final cost. The functions are called cost estimation relationships, and are used to establish cost estimation models through statistical regression analysis methodology [14]. Cost estimation software programs, such as PRICE H, SEER H, and COCOMO-II, have been developed for specific industries. All of these cost estimation software programs are based on statistical regression analysis [10, 12]. Statistical regression analysis methodology includes some function types. For simple prediction problems, a one-order regression model is appropriate. As the complexity of feature variables and problem characteristics increases, more regression analysis forms are developed to solve the different problems. A second-order regression analysis model consists of terms of squared variables and interactive terms between the variables.

Although regression analysis is popular and useful for many cost estimation problems, it has some drawbacks. The regression analysis method is unsuitable for accounting for large numbers of complex variables [15]. The type of cost relationship between variables must be supposed *a priori* and the number of input variables is limited [14]. Cost estimation is always difficult in the early product development phase of the life cycle when only conceptual information is known. The major difficulty of parametric methods is that the cost driver relationships must be known to build an accurate estimation model [6, 16]. Wang [4] proposes machine learning as an advanced cost estimation method. Machine learning methods can provide an accurate and flexible estimation model with less data. When cost factors change, modifying a few parameters can renew the estimating model, rather than being forced to rebuild it. Regression analysis affords no general approach to help the cost estimator determine the estimating model that well fits historical databases for a specific problem. Numerous studies have used the neural networks to solve cost estimation problems.

2.2 Neural Networks

The neural network method is based on imitating the neuron transfer frameworks present in the human brain, especially those involved in the training and learning phases, and is also known as the artificial neural network (ANN) method. The most popular approach is the back-propagation neural network (BPN). BPN compares the target and network output values and then minimizes the error using the gradient steepest descent method. The parameter

combination decision is the most significant issue in BPN modeling. The estimating performance will be influenced directly by the parameter decision. The main parameters include the number of hidden layers and neurons, transfer functions between layers, stop operation criteria and operation iterations. Some studies indicated that trial and error is the general parameter decision method for artificial neural networks (ANN) [1, 4, 6]. The main advantages of neural networks are that they can be used to construct complex nonlinear function estimation models and do not impose any limit on the number of features.

Neural networks technique is widely applied to solving cost estimation problems to overcome the drawbacks of the regression model. Numerous researchers have conducted studies on the cost estimation problem. The following literature shows the related applications for cost estimation where neural networks are applied.

- McKim [17] applied the NN method to establish a cost estimation model for purchasing pumps, using the same cost data and comparing the results with three common methods proposed by Bieledeld and Rucklos [18].
- Creese and Li [19] adopted the neural network to establish cost estimation models for timber bridge projects and compared the performance with a linear regression analysis method.
- Garza and Rouhana [15] focus on estimating the cost of carbon steel pipe material. The NN method is used to build the estimation model, and its performance is compared with the traditional regression model. The material cost data set is reported by Sigurdsen [20].
- Zhang and Fuh [1] established a feature based BPN model for estimating the costs involved in the early design phase of packaging products.
- Shtub and Versano [21] approach the steel pipe bending process cost estimation problem, and apply NN and a regression model. In the NN modeling process, the trial and error parameter decision method is used. The results show the NN estimations outperform the traditional regression analysis method.
- Günaydın and Dŏgan [22] investigated the utility of the neural network methodology to overcome cost estimation problems in the early phases of building design.
- Cavalieri et al. [23] compared the results of parametric regression and ANN for estimating the unitary manufacturing costs of a new type of brake disks produced by the automotive industry.
- Wang [24] applied ANN as the cost estimation method in a collaboration manufacturing environment and the results produce the most accurate and flexible cost response for improved decision making.
- Wang [4] propose a cost estimation model using back-propagation neural networks (BPN) with feature based models to dramatically simplify the complex conventional cost estimation procedures and requested computation parameters.
- Verlinden et al. [5] apply artificial neural networks (ANN) and multiple regression analysis (MLR) to building a cost estimation model for sheet metal components and provide the rapid price quotations customers expect. The results show that neural networks provide better performance, but are still considered a black box.
- Ciurana et al. [25] developed a cost estimating model for vertical high speed machining centers based on machining features, and used multiple regression analysis and ANN as the modeling methods. Finally, they examined and compared the estimation performance of these methods determines the method that provides best accuracy.
- Liu et al. [26] apply regression tree models, artificial neural networks (ANN), and support vector regression (SVR) non-parametric models for estimating life cycle cost under different environments. In case studies, two popular real cost estimation cases are used to verify model performance. The cost data is reported by Smith and Mason [14]. The results show that support vector regression (SVR) and ANN outperform the regression model.
- Deng et al. [27] focus on the design cost estimation of airframe wing-box main structural parts and established a model using back-propagation neural networks

(BPN) and second order regression methods. In the BPN modeling process, differing combinations of training functions and numbers of hidden layer nodes were tried. The Levenberg-Marquardt (LM) training function provides better accuracy through the trial and error parameter selection process. The testing results shown that BPN outperforms regression analysis.

- Duran et al. [28] used artificial neural networks (ANN) to develop a cost estimating model for the early design phase of shell and tube heat exchangers. In ANN modeling process, many training and testing samples size combination are tried to search the best samples size.
- Chang et al. [29] integrated the case-based reasoning (CBR) and artificial neural networks (ANN) as a product unit cost estimation model for mobile phone industry. The results show this model can provide accurate performance.

Although the NN technique has been widely applied to cost estimation problems, it can be improved. When designing a neural network model, it is necessary to set numerous parameters, including the numbers of neurons in the input, hidden, and output layers; the activation function; and the training and learning functions. Trial and error is generally the parameter selection method for neural networks [1, 4, 6, 30]. In addition, the neural network solution is not globally optimal solution; therefore, neural networks lack stable performance [30-31]. Verlinden et al. [5] indicated that the support vector machines (SVM) method can overcome the drawbacks of BPN. The SVM method used in the Lagrange multiplier optimal programming method transformed the original problem into a convex problem. The SVM solution is unique and global; the modeling process requires fewer parameters and will save time and cost over trial and error parameter selection.

2.3 Support Vector Machines

The support vector machines (SVM) was developed by Vapnik [32]. Most different of the traditional neural network models which implement the empirical risk minimization principle, SVM implement the structural risk minimization principle (SRM) which search to minimize an upper bound of the generalization error rather than minimize the training error. This principle is based on the fact that the generalization error is bounded by the sum of the training error and a confidence interval term that depends on the Vapnik–Chervonenkis (VC) dimension. Based on this principle, SVM achieve an optimum network structure by striking a right balance between the empirical error and the VC-confidence interval. This finally results in better generalization performance than other neural network models. Another advantage of SVM is that the training of SVM is equivalent to solving a linearly constrained quadratic programming. That means the solution of SVM is unique, optimal and absent from local optimal solution [31-32].

Support vector regression (SVR) was developed by Vapnik in 1997. SVR introduces the ε-loss function concept to SVM, extending it to solving function estimation problems. SVR has been applied to many problems. Tay and Cao [31] apply SVR to solving the time series financial forecasting problem, and compare it with BPN. Kim [30] uses SVM, BPN and case-based reasoning (CBR) methods to forecasting daily stock price change. Pai and Lin [33] propose to forecast stock prices using a hybrid ARIMA and SVM model. Deng and Yeh [34] compare SVR, BPN and regression analysis method to solving airframe wing-box structure parts design cost estimation problem. The results show SVR can achieve more accurate performance and easier for modeling. Above these studies indicate that SVM method outperform other methods, achieve stable solution, and easier to modeling than BPN.

III. LEAST SQUARES SUPPORT VECTOR MACHINES

The Least Squares Support Vector Machines (LS-SVM) was proposed by Suykens in 2002 [35]; it modifies the least squares loss function and introduces it to support vector machines. LS-SVM is based on equality constraints and a sum square error cost function, as opposed to earlier approaches that utilize inequality constraints and solve complex convex optimization problems. The LS-SVM

reformulation simplifies the problem and solution by adopting a linear system; the solution follows from a linear KKT system rather than quadratic problems that are difficult to compute. Thus, LS-SVM is easier to optimize and has a shorter computing time. Another important difference between SVR and LS-SVM is that the LS-SVM calculation considers the training errors of all training samples. LS-SVM is useful for applications where nearly all of the training samples can affect the training phases. Fig.1 compares between SVR and LS-SVM.

In recent years, LS-SVM has been applied to solve the classification and function estimation problems. The problem should identify the target (output) and influenced (input) variables before using LS-SVM modeling procedure. The LS-SVM can adapt in various data situation, like few or large sample size, lots input variables, and complex nonlinear problems. The recently application of LS-SVM, includes electric load forecasting [36], financial credit scoring classification [37], and electroencephalogram (EEG) signal classification [38].

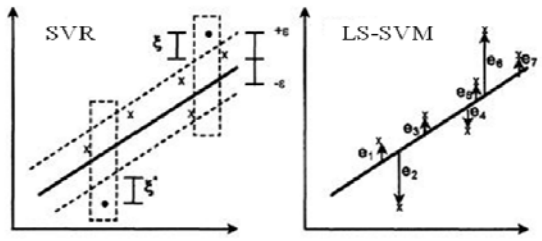

Fig.1. Comparison between SVR and LS-SVM.

3.1 LS-SVM Basic Theory

The LS-SVM theory [35] is based on the assumption that the dataset $S = \{(x_1, y_1),\ldots, (x_n, y_n)\}$, which processes a nonlinear function and a decision function, can be written as shown in Eq.(1). In Eq.(1), w denotes the weight vector; Φ represents the nonlinear function that maps the input space to a high-dimension feature space and performs linear regression; and b is the bias term. Fig.2 illustrates the high-dimension nonlinear transfer function.

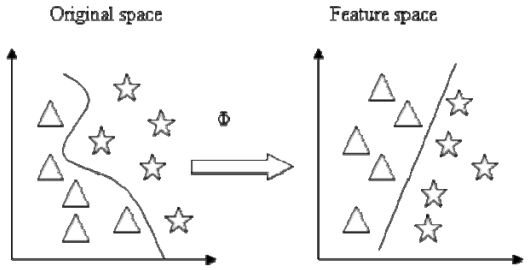

Fig.2. Nonlinear transfer function.

$$f(x) = \Phi(x) \cdot w + b \tag{1}$$

For the function estimation problem, the SRM principle is introduced, and the optimization problem is used to formulate the R function Eq.(2), where C denotes the regularization constant and e_i represents the training data error.

$$\text{Minimize: } R(w,e,b) = \frac{1}{2}\|w\|^2 + \frac{1}{2}C\sum_{i=1}^{n} e_i^2 \tag{2}$$
$$\text{s.t. } y_i = [\Phi(x_i) \cdot w] + b + e_i \quad i = 1,\ldots.n$$

To derive solutions w and e, the Lagrange Multiplier optimal programming method is applied to solve (Eq. 2); the method considers objective and constraint terms simultaneously. The Lagrange function L is shown as Eq.(3).

$$L(w,e,b;\alpha) = \frac{1}{2}\|w\|^2 + \frac{1}{2}C\sum_{i=1}^{n} e_i^2 - \sum_{i=1}^{n}\alpha_i\{(\Phi(x_i)\cdot w) + b + e_i - y_i\} \tag{3}$$

In (Eq.3), $\alpha_i \geqq 0$ called Lagrange multipliers, which can be either positive or negative due to the following equality constraints, from based on the Karush Kuhn-Tucher's (KKT) conditions and which may obtain the extreme value in the saddle point, the conditions for optimality are given by Eq.(4). Eq.(4) can be expressed as the solution to the following set of linear equations Eq.(5).

$$\begin{aligned}
&\partial_w L = w - \sum_{i=1}^{n}\alpha_i\Phi(x_i) = 0\\
&\partial_b L = \sum_{i=1}^{n}\alpha_i = 0\\
&\partial_{e_i} L = C\cdot e_i - \alpha_i = 0\\
&\partial_{\alpha_i} L = (\Phi(x_i)\cdot w) + b + e_i - y_i = 0
\end{aligned} \tag{4}$$

$$\begin{bmatrix} I & 0 & 0 & -Z^T \\ 0 & 0 & 0 & -1_v^T \\ 0 & 0 & CI & -I \\ Z & 1_v & I & 0 \end{bmatrix} \begin{bmatrix} w \\ b \\ e \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ y \end{bmatrix} \quad (5)$$

In Eq.(5), $Z = [\Phi(x_1)^T;\ldots;\ \Phi(x_n)^T]$, $y = [y_1;\ldots;y_n]$, $1_v = [1;\ldots;1]$, $\alpha = [\alpha_1;\ldots;\ \alpha_n]$, and $e = [e_1;\ldots;e_n]$. The solution is also given by Eq.(6).

$$\begin{bmatrix} 0 & 1_v^T \\ 1_v & ZZ^T + C^{-1}I \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix} \quad (6)$$

In order to simplify the solving process, let $\Omega = ZZ^T + C^{-1}I$, where α and b are the solution to Eq.(7) and Eq.(8).

$$\alpha = (y - b1_v)\Omega^{-1} \quad (7)$$

$$b = (1_v^T \Omega^{-1} 1_v)^{-1} 1_v^T \Omega^{-1} y \quad (8)$$

The resulting LS-SVM model for function estimation is represented as Eq.(9).

$$f(x) = \sum_{i=1}^{n} \alpha_i K(x, x_i) + b \quad (9)$$

In Eq.(9), the dot product $K(x \bullet x_i)$ is known as the kernel functions. Kernel functions enable the dot product to be computed in a high-dimension feature space using low-dimension space data input without the transfer function Φ and must satisfy the condition specified by Mercer [33]. This study employed the radial basis function (RBF), a common function that is useful in function estimation problems. RBF kernel function is represented in Eq.(10), where γ denotes the kernel functions parameter [39].

$$K(x_i, x_j) = \exp\left(-\gamma \left\| x_i - x_j \right\|^2\right) \quad (10)$$

3.2 LS-SVM Modeling Procedure

The optimal parameter combination selection is the most significant topic when establishing the LS-SVM estimation model, because it can significantly affect performance; consequently, this research applied the grid search algorithm with a k-fold cross-validation method [40] to obtain the optimal parameter combination. When adopting LS-SVM with the RBF kernel function, the parameter combination (C, γ) should be established, where C denotes the regularization parameter, namely, the trade-off between minimizing the training error and minimizing the complexity of the LS-SVM model; γ denotes the RBF kernel parameter and represents the change in the RBF kernel function.

The LS-SVM modeling procedure is shown in Fig.3. The following provides a detailed description of the LS-SVM modeling procedure. First, divide all data into training and testing data sets. The training data set is used to build the LS-SVM model, and the testing data set verifies the LS-SVM model performance. Second, search the optimal parameter combination using the grid search algorithm with the cross-validation method. Separate the training data into grid training and testing data. This research applied the 10-fold cross-validation method, dividing the training into 10 aliquot parts. The grid training data contains nine aliquot parts and the other is the grid testing data. Train the LS-SVM model with the grid training data and the initial parameter combination (C, γ). Test the LS-SVM model with the grid testing data. Reiterate the process 10 times, collecting and calculating the average error. Replace the new parameter combination and repeat the process until it approaches the stopping criteria. Finally, we obtain the optimal parameter combination (C, γ) with minimized error. Finally, adopt the optimal parameter combination to build the LS-SVM model. Substitute the testing data set into the LS-SVM model to obtain the estimation values. Use performance criteria to calculate the error between the actual and estimation values; use the testing performance to verify the estimation performance of the LS-SVM model.

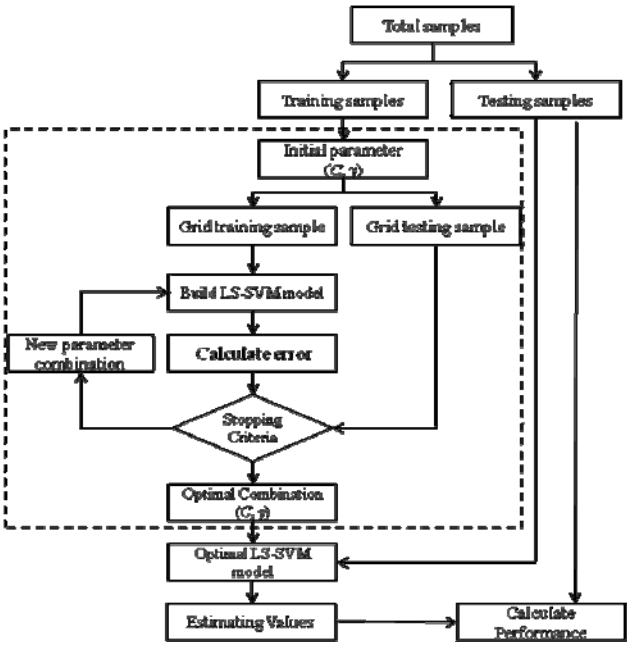

Fig.3. LS-SVM modeling procedure.

3.3 Estimation Performance Criteria

This research applied mean absolute percentage error (MAPE), mean squared error (MSE) and coefficient of determination (R^2) as criteria for assessing estimation performance. The estimated values approached the actual values more closely than those of MAPE or MSE. The coefficient of determination denotes the degree of matching between the actual and estimated values. The closer R^2 is to 1, the more accurate the performance. The performance criteria calculation formulas were derived from Eq.(11-13), where n denotes the sample number, s_i represents the actual value, and s_i' is the estimated value.

$$MSE = \frac{1}{n}\sqrt{\sum_{i-1}^{n}\left(s_i - s_i'\right)^2} \tag{11}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{s_i - s_i'}{s_i}\right| \times 100\% \tag{12}$$

$$R^2 = \frac{\sum_{i=1}^{n}(s_i' - \bar{s})^2}{\sum_{i=1}^{n}(s_i - \bar{s})^2} \tag{13}$$

IV. PRODUCT COST ESTIMATION CASE STUDIES

In this section, we verify the feasibility and accuracy of LS-SVM method, and use four product related cost estimation cases proposed by previous research. Case 1 involves estimating the material cost of carbon steel pipe; case 2 proposes estimating the bending process cost of the steel pipe; case 3 proposes estimating the manufacturing cost of pressure vessels; and case 4 proposes estimating the purchase cost of the pump. For all cases, we establish the LS-SVM model and consider the same data set, cost drivers, cost function, and definition, which are reported from previous studies. LS-SVM estimation performance is compared with other models, proposed in former studies. Finally, we verify the LS-SVM method can provide better performance than other methods and make it easier to build a model.

4.1 Case 1: Carbon Steel Pipe Material Cost Estimation

In case 1, we apply LS-SVM to build a material cost estimation model for the carbon steel pipe. The original case definition and data sets are reported by Sigurdsen [20], and Garza and Rouhana [15]; introduce neural networks to construct the estimation model. The main estimation objective is the material cost of the carbon steel pipe. The cost drivers (independent variables) of material cost contain the pipe diameter, the number of elbows, and the flange rate of pipe. The entire data set contains 16 samples, randomly divided into 10 training and six testing samples.

Case 1 Cost estimation function:

Material cost of carbon steel pipe = f (D, N_E, F_R)

Case 1 Parameter definition:

- D: diameter of pipe (unit: inch)
- N_E: number of elbows
- F_R: flange rating (unit: psi)

In the LS-SVM training phase of case 1, the parameter setting can be shown by the following description. The RBF kernel function was selected as the nonlinear mapping function. The optimal parameter combination (C, γ) of the kernel and regularization was calculated through a grid search algorithm. The LS-SVM training

model was established with the parameter combination ($C = 3.3534$, $\gamma = 0.3662$). Table 1 shows the parameter setting of the LS-SVM training model.

The case 1 estimation performance (MSE, MAPE, R^2) of LS-SVM model was (2.1551, 8.54%, 0.9912). The estimation performance of NN was (3.1109, 9.54%, 0.9873) which reported by Garza and Rouhana [15]. The results show that both LS-SVM and NN method can accurately estimate the carbon pipe material cost. In the NN modeling process, more parameter combinations must be determined, and it takes more time. On the other hand, LS-SVM modeling process decides only two parameters, saving time. Table 2 shows the estimation results and performance of LS-SVM and NN.

Table.1 Case 1: LS-SVM estimation model parameter setting

LS-SVM training model	Kernel	C	γ
	RBF	3.3534	0.3662

4.2 Case 2: Steel Pipe Bending Process Cost Estimation

In case 2, we use LS-SVM to solve the steel pipe bending process cost estimation problem. The data set of 36 samples is original reported by Shtub and Versano [21]. The main estimation objective is the bending process cost. The independent variables include the external and internal diameter of the pipe, the number of axes in space, the number of bends, and the distance between the lending location and the end of the pipe. In this case, we apply LS-SVM to build the process cost estimation model with 36 samples and verify its performance with the same samples. The cost estimation function and parameter definition summary of case 2 are:

Case 2 Cost estimation function:

Bending process cost = $f(d_1, d_2, K_f, Z_r, L_g)$

Case 2 Parameter definition:

- d_1 : the external diameter of the pipe
- d_2 : the internal diameter of the pipe
- K_f : the number of axes in space in which the pipe is bent
- Z_r : the number of bends
- L_g : the distance between the bending location and the end of the pipe

In the LS-SVM training phase, the optimal parameter combination was $C = 21.1588$, $\gamma = 2.846$, and the RBF kernel function served as the nonlinear transfer function. Table 3 shows the parameter setting of the LS-SVM model. The case 2 estimation performance (MSE, MAPE, R^2) of the LS-SVM model was (101.1606, 4.09%, 0.9889); the REG estimation performance (3639.0294, 24.6%, 0.602); and NN was (1367.5294, 16.09%, 0.8504). The results show that the mean absolute percentage error of LS-SVM is superior to that of the NN model by about 12%, and the REG model by about 20%. Table 4 shows the estimation results and performance for regression, NN, and LS-SVM.

Table.3 Case 2: LS-SVM estimation model parameter setting

LS-SVM training model	Kernel	C	γ
	RBF	3.3534	0.3662

4.3 Case 3: Pressure Vessels Manufacturing Cost Estimation

In case 3, we focus on the pressure vessel manufacturing cost estimation of a new chemical product. The original data set of 20 samples, reported by Smith and Mason [14], and Liu et al. [26], use ANN and SVR to build an estimation model and evaluate estimation performance. In this case, we apply LS-SVM to build the manufacturing cost estimation model with the 20 samples, and verify its performance with the same samples. The cost function objective is the manufacturing cost of pressure vessels, and the cost drivers include the height, diameter, and thickness of the pressure vessels. The cost estimation function and parameter definition summary of case 3 are:

Case 3 Cost estimation function:

Manufacturing cost = f (H, D, T)

Case 3 Parameter definition:

- H : height of vessels (unit: mm)
- D : diameter of vessels (unit: mm)
- T : thickness of vessels (unit: mm)

The LS-SVM modeling process of case 3 used the RBF kernel function, and the optimal parameter combination was $C = 3969.1614$, $\gamma = 115.4697$. Table 5 shows the parameter setting of the LS-SVM training model. The estimation performance (MAPE, R^2) of LS-SVM was (8.083%, 0.9924), it significantly outperforms ANN (17.73%, 0.9506) and SVR

(24.123%, 0.969). Table 6 shows the estimation results and performance of ANN, SVR, and LS-SVM.

Table.5 Case 3: LS-SVM estimation model parameter setting

LS-SVM training model	Kernel	C	γ
	RBF	3.3534	0.3662

4.4 Case 4: Pump Purchase Cost Estimation

In case 4, we consider the pump purchase cost estimation problem. The original data set of 23 samples, reported by Bielefeld and Rucklos [18], and McKim [17], use the NN method to establish estimation model. The cost function objective is the purchase cost of pumps, and the cost drivers include the flow rate in gallons per minute and the head, measured in feet. The cost estimation function and parameter definition summary of case 4 are:

Case 4 Cost estimation function:

Purchase cost of pumps = f (g, h)

Case 4 Parameter definition:

- g : the flow rate in gallons per minute (unit: gpm)
- h : the head measured in feet (unit: ft)

The LS-SVM training phase of case 4, the setting of related parameters is as flow. The RBF kernel function was selected for the nonlinear transfer function and the optimal parameter combination was $C = 52.8353$, $\gamma = 0.7909$. Table 7 shows the parameter setting of the LS-SVM training model. The estimation performance (MSE, MAPE, R^2) of LS-SVM was (497399, 1.52%, 0.9786), which outperforms NN (2491131, 3.65%, 0.8928). The results show that although NN can approach good performance, it spent more time deciding parameters by the trial-and error method, and could not ensure the solution quality. The solution of LS-SVM was uniquely optimal, and afforded a stable solution. Table 8 shows the estimation results and performance of LS-SVM and NN.

Table.7 Case 4: LS-SVM estimation model parameter setting

LS-SVM training model	Kernel	C	γ
	RBF	3.3534	0.3662

V. CONCLUSION

An accurate, rapid, and robust cost estimation model is important for industrial products in their design or production stages. This research applied a novel machine learning method – the least squares support vector machines (LS-SVM) to solving industry product cost estimation problems. In case studies, estimations of the product costs of four real products, presented in previous studies, were used in our work. These cases include material, process, manufacturing, and purchase cost estimation problems. The estimation results and performance show that LS-SVM can provide more accurate estimations. The LS-SVM solution was globally optimal and fewer parameters needed to be determined when establishing the LS-SVM model; the solution is stable. The NN method also provides accurate performance, more parameters must be determined and locally optimal solutions must be avoided. The LS-SVM method is easier to optimize and has a shorter computing time than SVR. The LS-SVM calculation considers the training errors of all training samples, whereas the SVR method considers only significant samples when building a model, and some samples are discarded. Finally, this research suggests that the use of LS-SVM will approach accurate estimation performance for industry product cost control and planning.

In the real world, cost databases are highly susceptible and sensitive to outliers and abnormal data. The quality of cost data directly influences the accuracy of cost estimation models. In future work, we will make attempt to combine LS-SVM and data mining techniques (data processing and transformation) to solve the abnormal and outlier data situation in cost databases, and apply real product cost estimation case. Furthermore, the efficiency and speed of LS-SVMs parameter selection will be enhanced, and a more robust algorithm will be used in the next step, such as heuristic algorithm or another optimal method. Finally, we will strive to provide a more accurate, robust, and rapid cost estimation model for industries to improve their competitiveness.

REFERENCES

[1] Zhang, Y. F., and Fuh, J. Y. H., "A neural network approach for early cost estimation of packaging products," Computers and Industrial Engineering, Vol. 34, No. 2, pp. 433-450, 1998.

[2] Jung, J. Y., "Manufacturing cost estimation for machined parts based on manufacturing features," Journal of Intelligent Manufacturing, Vol. 13, pp. 227-238, 2002.

[3] Hundal, M. S., "Product Life Cycle: The Ultimate Aim in Product Development," Engineering Design in Integrated Product Development, EDIProD'2002, pp 57-64, 2002.

[4] Wang, H. S., "Application of BPN with feature-based models on cost estimation of plastic injection products," Computers and Industrial Engineering, Vol. 53, pp. 79-94, 2007.

[5] Verlinden, B., Duflou, J. R., Collin, P., and Cattrysse, D., "Cost estimation for sheet metal parts using multiple regression and artificial neural networks: A case study," International Journal of Production Economics, Vol. 111, pp. 484-492, 2008.

[6] Bode, J., "Decision support with neural networks in the management of research and development: Concepts and application to cost estimation," Information and Management, Vol. 34, pp. 33-40, 1997.

[7] Creese, R. C., Adihan, M., and Pabla, B. S., Estimating and Costing for the Metal Manufacturing Industries, Marcel Dekker, 1992.

[8] Perera, H. S. C., Nagarur, N., and Tabucanon, M. T., "Component part standardization: A way to reduce the life-cycle costs of products," International Journal of Production Economics, Vol. 60-61, pp. 109-116, 1999.

[9] Michaels, J. V., and Wood, W. P., Design To Cost, John Wiley and Sons, 1989.

[10] Stewart, R. D., and Wyskida, R. M., Cost Estimator's Reference Manual 2nd Edition, John Wiley and Sons, 1995

[11] Mantel, S. J., Meredith, J. R., Shafer, S. M., and Sutton, M. M., Core Concepts: Project Management in Practice 2nd Edition, Wiley, 2005.

[12] International Society of Parametric Analysis (ISPA), Parametric Estimating Handbook. 4th edition, 2008.

[13] Layer, A., Brinke, E. T., Houten, F. V., Kals, H., and Haasis, S., "Recent and future trends in cost estimation," Computer Integrated Manufacturing, Vol. 15, No. 6, pp. 499-510, 2002.

[14] Smith, A. E., and Mason, A. K., "Cost Estimation predictive modeling: regression versus neural network," The Engineering Economist, Vol. 42, No. 2, pp. 137-161, 1997.

[15] Garza, J., and Rouhana, K., "Neural Networks versus Parameter-Based Application in Cost Estimating," Cost Engineering, Vol. 37, No. 2, pp. 14-18, 1995.

[16] Bode, J., "Neural networks for cost estimation: simulations and pilot application," International Journal of Production Research, Vol. 38, No. 6, pp. 1231-1254, 2000.

[17] McKim, R. A., "Neural Network Application to Cost Engineering," Cost Engineering, Vol. 35, No. 7, pp. 31-35, 1993.

[18] Bielefeld, J. R., and Rucklos, G. D., "Cost Scaling Factors: How accurate are they?," Cost Engineering, Vol. 34, No. 10, pp. 15-20, 1992.

[19] Creese, R. C., and Li, L., "Cost estimation of timber bridges using neural network," Cost Engineering, Vol. 37, No. 5, pp. 17-22, 1995.

[20] Sigurdsen, A., "CERA: An Integrated Cost Estimating Program," Cost Engineering, Vol. 34, No. 6, pp. 25-30, 1992.

[21] Shtub, A., and Versano, R., "Estimating the cost of steel pipe bending, a comparison between neural networks and regression analysis," International Journal of Production Economics, Vol. 62, pp. 201-207, 1999

[22] Günaydın, M. H., and Dŏgan, Z. S., "A neural network approach for early cost estimation of structural systems of buildings," International Journal of Project Management, Vol. 22, pp. 595-602, 2004.

[23] Cavalieri, S., Maccarrone, P., and Pinto, R.,

"Parametric vs. neural network model for the estimation of production costs: A case study in the automotive industry," International Journal of Production Economics, Vol. 91, pp. 165-177, 2004.

[24] Wang, Q., "Artificial neural networks as cost engineering methods in a collaborative manufacturing environment," International Journal of Production Economics, Vol. 109, pp. 53-64, 2007.

[25] Ciurana, J., Quintana, G., and Garcia-Romeu, M. L., "Estimating the cost of vertical high-speed machining centers, a comparison between multiple regression analysis and the neural networks approach," International Journal of Production Economics, Vol. 115, No. 1, pp. 171-178, 2008.

[26] Liu, H., Gopalkrishnan, V., Kim, T. H. Q., and Ng, W. K., "Regression models for estimating product life cycle cost," Journal of Intelligent Manufacturing, Vol. 20, pp. 401-408, 2009.

[27] Deng, S., Chin, H., and Yeh, T. H., "Using Artificial Neural Networks to Airframe Wing Structural Design Cost Estimation," Journal of Chung Cheng Institute of Technology, Vol. 38, No.1, pp. 97-106, 2009.

[28] Duran, O., Rodriguez, N., and Consalter, L. A., "Neural networks for cost estimation of shell and tube heat exchangers," Expert Systems with Applications, Vol. 36, 7435-7440, 2009.

[29] Chang, P. C., Lin, J. J., and Dzan, W. Y., "Forecasting of manufacturing cost in mobile phone products by case-based reasoning and artificial neural network models," Journal of Intelligent Manufacturing, 2010 (in press).

[30] Kim, K. J., "Financial time series forecasting using support vector machines," Neurocomputing, Vol. 55, pp. 307-319, 2003.

[31] Tay, E. H., and Cao, L., "Application of support vector machines in financial time series forecasting," Omega, Vol. 29, pp. 309-317, 2001.

[32] Vapnik, V., The nature of statistical learning theory, Springer, 1995.

[33] Pai, P. F., and Lin, C. S., "A hybrid ARIMA and support vector machines model in stock price forecasting," Omega, Vol. 33, pp. 497-505, 2005.

[34] Deng, S., and Yeh, T. H., "Applying Machine Learning Methods to the Airframe Structural Design Cost Estimation – A Case Study of Wing-Box," 19th Annual International INCOSE Symposium, Singapore, 2009.

[35] Suykens, J. A. K., Gestel, T. V., Brabanter, J. D., and Vandewalle, J., Least square support vector machines, World Scientific, 2002.

[36] Espinoza, M., Suykens, J. A. K., Belmans, R., and Moor, B. D., "Electric Load Forecasting – Using Kernel-based Modeling for Nonlinear System Identification," IEEE Control Systems Magazine, pp. 43-57, 2007.

[37] Yu, L., Lai, K. K., and Zhou, L., "Least squares support vector machines ensemble models for credit scoring," Expert Systems with Applications, Vol. 37, pp. 127-133, 2010.

[38] Übeylï, E. D., "Least squares support vector machine employing model-based methods coefficients for analysis of EEG signals," Expert Systems with Applications, Vol. 37, pp. 233-239, 2010.

[39] Smola, A. J., and Schölkopf, B., Learning with Kernels – Support Vector Machines, Regularization, Optimization, and Beyond, MIT Press, 2002.

[40] Duan, K., Keerthi, S. S., and Poo, A. N., "Evaluation of simple performance measures for tuning SVM hyperparameters," Neurocomputing, Vol. 50, pp. 41-59. 2003.

Table.2 Case 1: estimation results and performance

Jobs	Parameter			Actual cost	Estimation cost	
	D	N_E	F_R		NN	LS-SVM
Training sample						
2	20	14	150	43.2	42.96	42.9145
4	2	14	250	1.9	3.08	1.2496
6	8	16	150	11.7	11.44	11.5972
8	18	8	200	26.1	27.49	26.2526
10	24	12	100	50.2	47.35	50.6268
11	16	12	200	28.4	29.49	28.7336
13	6	12	150	6.5	6.17	6.0237
14	24	8	300	42.3	42.71	42.5321
15	12	4	300	10.8	8.4	10.8218
16	20	8	100	28.9	31.4	29.3184
Testing sample						
1	20	14	250	46.1	43.05	43.0913
3	20	14	100	42.1	42.73	42.074
5	12	12	100	16.8	17.71	17.391
7	16	12	100	26.3	28.96	27.3057
9	4	4	300	2.5	1.7	3.3518
12	20	12	250	41.3	40.65	39.9615
Estimation Performance						
MSE					3.1109	2.1551
MAPE					9.54%	8.54%
R^2					0.9873	0.9912
Original data reported by Sigurdsen [20] NN estimation results reported by Garza and Rouhana [15]						

Table.4 Case 2: estimation results and performance

Number	Parameter					Actual cost	Estimation cost		
	d_1	d2	K_f	Z_r	L_g		REG	NN	LS-SVM
1	16	13	5	3	1	239.5	149.3	157	203.4
2	16	13	3	2	1	106.2	99.3	124	111.2
3	18	16	5	4	2	408	312.4	419.4	396.6
4	16	14	1	1	2	156.6	163.8	150.8	156.0
5	10	8	5	5	2	433.5	301	365.4	424.3
6	20	17	3	3	1	109	151.2	109.3	115.0
7	18	15	1	1	2	157.2	151	140.9	160.2
8	11	9	3	3	1	101.6	137.1	117.8	111.3
9	20	18	4	2	1	170.4	139.3	154	172.2
10	12.7	11.3	6	3	2	210	311.1	187.1	211.4
11	11	9	2	1	2	148.6	163.5	163.3	148.2
12	30	28	4	1	2	289.4	231.7	304.5	286.6
13	10	9	3	2	1	166	101.4	97.5	155.8
14	25	23	5	2	2	306	273.3	275	305.8
15	11	8	4	3	1	127.2	132.3	156.5	120.9
16	14	12	2	1	2	153.8	172.6	159.7	153.2
17	18	16	6	4	2	425	326.2	358.7	418.4
18	28	26	2	2	1	118.5	143.7	96	122.9
19	11	8	5	2	2	162.5	232.6	170.3	168.5
20	14	12	5	3	1	112.7	175.6	195.9	145.4
21	25.4	22.9	6	3	2	412	300.8	377.1	402.1
22	16	13	4	2	1	170.8	107.2	153.9	171.0
23	11	9	4	2	2	200.1	223.8	175.4	196.3
24	16	14	2	1	1	88	67.7	120.9	92.1
25	18	16	3	2	1	113.9	122.6	131.9	110.5
26	11	9	3	1	2	151	178.5	172.2	156.1
27	19.1	16.6	5	3	2	207.8	290.8	276.2	223.2
28	15	13	5	4	1	145.1	212	163.1	151.9
29	22	20	2	1	2	204	193	144.4	202.8
30	15	12	4	1	2	182.4	186.3	163.5	181.8
31	15	13	5	3	2	221.3	283.9	241.8	221.8
32	12.7	11.3	4	2	2	199.3	240.3	175.6	194.3
33	15	13	6	2	1	237.8	131.2	183.1	236.7
34	14	12	5	5	1	170.6	248.2	159.2	171.0
35	12.7	10.9	2	1	1	84	60.7	93.9	79.2
36	10	9	2	1	1	76.1	65.1	71	87.6
Estimation Performance									
MSE							3639.0294	1367.5294	101.1606
MAPE							24.60%	16.09%	4.09%
R^2							0.602	0.8504	0.9889
Original data reported by Shtub and Versano [21] REG and NN estimation results reported by Shtub and Versano [21]									

Table.6 Case 3: estimation results and performance

Number	Parameter			Actual cost	Estimation cost		
	H	D	T		ANN	SVR	LS-SVM
1	1200	1066	10	10754	15944	25691	11518
2	4500	1526	15	18172	15487	16530	18555
3	6500	1500	16	23605	16985	26291	22890
4	12250	1200	12	23956	28541	7715	21241
5	21800	1050	12	28400	34375	20537	31522
6	23300	900	14	31400	32454	36962	32304
7	26700	1500	15	42200	54456	59962	51396
8	12100	3000	11	47970	62048	58425	53851
9	17500	2400	12	48000	56067	58746	51237
10	26500	1348	14	51000	40812	48682	45862
11	28300	1800	14	53900	62303	67280	60251
12	14700	2400	10	54600	40911	39871	46025
13	26600	1500	15	58040	51667	59187	51215
14	24800	2500	13	61790	71618	75949	70945
15	25000	2100	14	61800	61490	69409	61458
16	24700	2000	16	67460	63878	75281	62557
17	29500	2250	13	80400	75610	77765	74405
18	21900	3150	12	85750	70185	76709	82680
19	32300	5100	17	207800	238675	202340	207006
20	53500	3000	29	240000	210030	245012	240080
Estimation Performance							
MSE					158894273	99782151	24481266
MAPE					17.73%	24.123%	8.083%
R^2					0.9506	0.969	0.9924
Original data reported by Smith and Mason [14] ANN and SVR estimation results reported by Liu et al. [26]							

Table 8. Case 4: estimation results and performance

Number	Parameter		Actual cost	Estimation cost	
	g	h		NN	LS-SVM
1	100	700	25584	25828	24683
2	100	1200	28296	28645	28355
3	100	2000	32160	31023	32093
4	200	700	25584	25953	25684
5	200	1200	28296	29307	26371
6	200	2000	32160	34121	31986
7	300	700	27924	26508	27459
8	300	1200	30936	29798	31235
9	300	2000	36660	36160	36326
10	300	2400	39240	37914	38738
11	400	700	28908	27856	29099
12	400	1200	31860	30446	32312
13	400	2000	36660	37123	36721
14	400	2400	39240	39687	38983
15	500	700	28908	30422	28657
16	500	1200	31860	31606	32202
17	500	2000	36660	37285	36908
18	500	2400	39240	40203	39038
19	600	700	31668	34400	32158
20	600	1200	34008	33580	33788
21	600	2000	39240	37029	38106
22	700	700	43668	39222	42097
23	700	1200	34008	36460	35261
Estimation Performance					
MSE				2491131	33994
MAPE				3.65%	0.41%
R^2				0.8928	0.9985
Original data reported by Bielefeld and Rucklos [18] NN estimation result reported by McKim [17]					