International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 24 (2016) pp. 11700-11711

Analysis of Complex PCB Design Data Using Data Mining Approach to Better Estimate the Work Hours

Tzu-Liang (Bill) Tseng[1], Johnny C. Ho[2], Sungjoo Lee[3] and Yongjin Kwon*[3]

[1]Department of Industrial, Manufacturing and Systems Engineering, The University of Texas at El Paso, El Paso, TX 79968, USA.

[2]Turner College of Business, Columbus State University, Columbus, GA 31907

[3]Department of Industrial Engineering, Ajou University, Suwon443-749, Republic of Korea.

**Corresponding Author's*

Abstract

This paper deals with the real-world problem of handling and utilizing a big set of data that have been collected over many years. The company, which designs and produces customer-ordered PCBs (printed circuit boards), places a bid to the customer in accordance with the estimation model that predicts the total working hours involved in the design and fabrication of PCBs. With the estimation accuracy of less than 30%, the company was unable to effectively compete with the competitors in terms of securing the customer order with the reasonable profit guaranteed. In today's competitive environment, such practice can no longer be tolerated. Hence, it was decided to reevaluate the estimation model and reformulate the prediction equations. The complex nature of data set precludes the use of conventional statistical approaches, and we found that the use of data mining techniques can effectively handle the problem. With the significantly improved estimation accuracy, the proposed data mining method can enhance the company's ability to effectively compete in the market place.

INTRODUCTION

The development of an electronics product, specifically printed circuit board assemblies (PCAs), is divided into three phases of realization: design, resource planning, and manufacturing. Task elements starting with functional specifications development through to design approval as the *design phase* is considered. In the *resource planning phase*, the component parts and bare board are ordered and the process planning for the board is accomplished. The *manufacturing phase* includes the bare board fabrication, component placement, as well as testing and customer verification. The total cycle time is the time required for all three phases—i.e., the time from the design concept to product completion through the manufacturing phase. Figure 1 generally depicts the packaging hierarchy of PCB, while Figure 2 shows various modules embedded in the PCB. In the design of printed circuit board assemblies (PCBAs), the layout stage of the design process has a major impact on manufacturing cost, lead-time, and various other product aspects such as reliability, performance, maintainability, and service life. Time-to-market is a critical factor for the new product introduction, and the elimination of manufacturing problems by means of repeated design iterations is no longer an acceptable option. The continuing demand for smaller, more complex products, and the consequent use of smaller, finer lead pitch components, new packaging technologies, and increasing population densities, makes the achievement of a "right first time" layout design increasingly difficult. Consequently, a complex PCB design lenders itself to a low yield during the fabrication processes. In this context, the PCB design and manufacturing company A-Teck (a pseudonym for the well-established company) has been collecting the data over several decades that require to complete any customer-ordered PCB designs from the initial design activities to the manufacture and including the fabrication of related paper works. The company developed a specific software to track all design and fabrication activities and the engineers are to enter the data along the required processes. There are over 83 variables in the data and the entire set of collected data is huge in size. The data are used to estimate the time to finish the custom-ordered PCB. In turn, the estimated time is used to calculate the cost involved in the work, and the cost is submitted to the customer. Usually, any qualified suppliers can compete in the bidding process, and the cost factor is very important to secure the customer order. Due to this reason, the company has been collecting the data for many years.

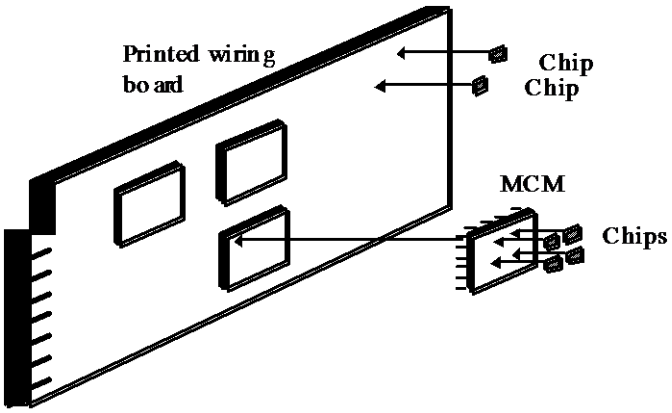

Figure 1. The packaging hierarchy of PCB

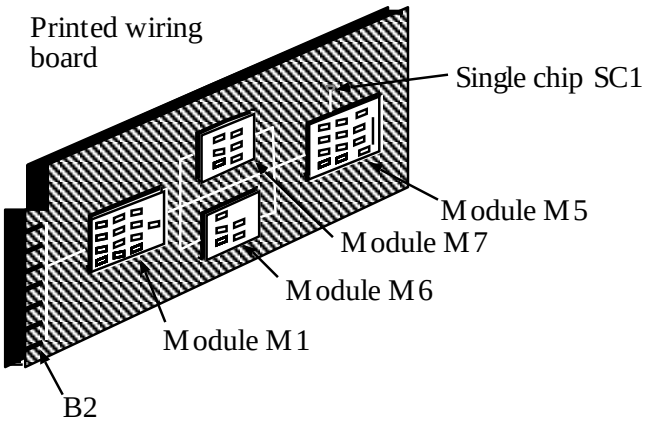

Figure 2. An example of modules embedded in the PCB

However, despite the best effort of the company, the cost estimation formulae produce only 30% accuracy in terms of predicting the working hours involved in any new design. This poor accuracy has been a major hindrance in terms of securing customer orders, as well as the reduced profit margins of the company. In this aspect, the A-Tech decides to reevaluate the equations and considers the application of new techniques, including the data mining approach. Therefore, the primary objective of this paper is to develop a data mining based methodology and apply it to model the relationship of outputs (i.e., the time to generate both Schematic Layout and Design Layout) to inputs (e.g., the number of connections, leads, layers, and board area, etc.) and evaluate the accuracy. The format of this paper is as follows. First, approaches to predictive models are reviewed with the conclusion that data mining is an emerging technique that can be applied in the forecasting area. Then, the Kohonen Neural Networks and Decision Tree are introduced, followed by the comparison of data mining and statistical analysis approaches in forecasting. The real data are analyzed and, finally, the research findings are discussed in the conclusion section.

REVIEW OF DATA MINING AND STATISTICAL ANALYSIS APPROACHES

Data Mining

Data mining is the process of finding trends and patterns in large databases. It is used to extract information and knowledge from vast amount of data (Chung et al., 2002 [1]). Traditional on-line transaction processing systems allow storing data into databases in an efficient way, but they fail to analyze the data. It is very hard to extract information due to lack of tools and techniques to turn data into information and knowledge (Sörensen et al., 2003 [2]). The data mining process is often characterized as a multi-stage iterative process involving data selection, data cleaning, and application of data mining algorithms, evaluation, and so forth. The process of data mining involves several steps like cleaning the data, preprocessing the data, as well as mining the data. These steps are explained in Figure 3.

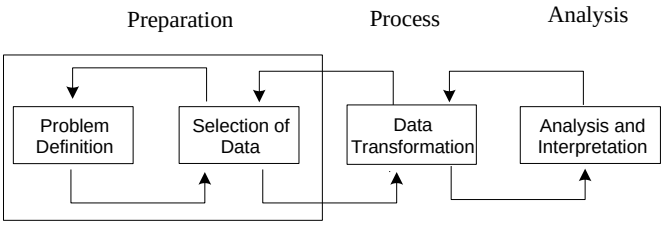

Figure 3. Data mining process

1) Problem definition and selection of input data: The first step in data mining is to specify the input data that are to be mined and analyzed. Numerous data analysis tools can help to select the specific data from a variety of data types and sources to which preprocessing and mining functions can be applied. Irrelevant attributes are omitted and relevant records in a dataset are selected.

2) Data transformation: After the input data are specified, the input data are transformed using data mining functions. Preprocessing such as discretization is used to mine the data effectively. Aggregation of the fields can be done and null values from the dataset can be removed using these preprocessing functions. Cleaning and preprocessing of the data involves several steps like feature extraction, and construction. Formatting the data is also helpful for the representation. In data mining, around 80% of the resources are spent on cleaning and preprocessing the data. For the actual implementation of the mining operation, the input data require to be cleaned as it is the source for the extraction of trends and patterns from the vast amount of information (Wei et al., 2003 [3]).

3) Data analysis and interpretation of the results: Transformed data are mined using one or more functions. Data mining tools are used for mining the input data using functions such as associations, classifications, and neural clustering. The end results are analyzed effectively by using the visualization effect in the data-mining tool (e.g., IBM Intelligent Miner). These results are used iteratively for mining because iterative mining causes the effective pattern extraction from the data set.

4) Data mining technologies and tools: Data mining technologies provide the next generation decision support software and services for the business and science applications. Some examples are the Internet, e-commerce, direct marketing, healthcare, genetics, CRM, telecommunications, utilities, financial services. There are many tools designed for these tasks of data mining. The choice of selecting the tool depends upon the kind of application. For this study, the IBM Intelligent Miner for Data is selected. It is a suite of statistical, preprocessing, and mining functions that can be used for large database. It also provides visualization effects for viewing and interpreting mining results. Data mining is an iterative process that typically involves selecting input data, transforming, running a mining function and interpreting the results. The Intelligent Miner assists with all the steps in this process. It provides a wide range of algorithms for clustering, classification,

association, prediction, and other related functions. Intelligent miner functions can be applied independently, iteratively or in combination. It uses many functions and mathematical models to discover the patterns in the data. In this paper, there are two techniques used for the data mining process. The data clustering algorithm (Kohonen neural network) and the data classification algorithm, which classify the data into classes using decision tree approach. These techniques are introduced next.

5) Kohonen Neural Networks: A recent work adopted neural networks to elucidate the ability to learn complex relationships between parameters and responses, usually for process and quality control (Heaton, 2005 [4]). These models are frequently used to identify optimal process settings. An approximated model can be constructed using a neural network. Although statistical regression methods and neural network method both can effectively correct the dimensional measurements of geometric features on a part profile, Chang et al. [5] indicated that neural network methods can be a very powerful alternative for precision measurement using computer vision system. Neural networks have been successfully applied to diverse areas such as speech synthesis and pattern recognition (Hinton et al. 2012 [6] [7] [8]). Once trained, a neural network can be evaluated very quickly, particularly during the optimization phase. In the recent review of neural network applications in manufacturing, Zhang et al. [9] cited such diverse venues as milling, metal cutting, injection modeling, arc welding and spray painting. Details regarding further applications can be found in [10]. Neural networks are formed by processing parallel units called neurons, which closely resemble the structure of a human neurological system. The elementary processors are interconnected so that knowledge pertaining to the relationship between input and output parameters are stored in the weights of the connections between them. Each neuron except the first layer contains the weighted sum of previous input neuron by an exponential function. This function allows neural networks to be generalized with a wide range of application. Neural networks can be categorized into network structures such as the multilayer perceptron, the feedback model of Hopfield and Tank [11], the adaptive resonance technique (ART) networks, the Kohonen network, and the learning methods such as back-propagation (BP). The ability to learn is one of the main advantages that makes the neural networks so attractive. They also have the capability of performing parallel processing and possess a significant fault tolerance. Since the BP neural network can be used to approximately realize continuous mapping [12], we can adopt the BP neural network owing to its ability to map the complex relationship between input data and corresponding outputs. Furthermore, the neural network is a system composed of many simple processing elements operating in parallel whose function is determined by network structure, connection strengths, and the processing performed at computing elements or nodes (Chris.S, 1996 [13]). In neural networks, knowledge is acquired by the network through a learning process. Neural networks are used for searching and querying the databases. It has the ability to learn during the process of search. These advantages make it suitable for data mining. The *self-organizing maps* (SOM) proposed by Teuvo Kohonen, referred to as Kohonen networks, are extremely popular topological neural networks. Input data vectors can be perceived as points in an *input vector space*. Classification tasks are mainly to project input data into a space of lower dimension (the *output space*) in which the classification is easier to decide (heaton, 2003 [14]). The intelligent miner searches the mining base for characteristics that most frequently occur in common, and groups the related records accordingly. Neural clustering employs Kohonen Feature Map neural network. Kohonen feature maps use a process called self-organization to group similar input records together. Kohonen neural networks are used to identify clusters in the input data, which are used for the analysis.

Moreover, the Kohonen neural network is also known as self-organizing map (SOM) and a type of artificial neural network that is trained using unsupervised learning to produce a low-dimensional discretized representation of the input space of the training samples. The Kohonen model provides a topological mapping. It places a fixed number of input patterns from the input layer into a higher-dimensional output or Kohonen layer. Training in the Kohonen network begins with the winner's neighbourhood of a fairly large size. Then, as training proceeds, the neighbourhood size gradually decreases. The Kohonen neural network operates in two modes: training and mapping. Training builds the map using input examples, while mapping automatically classifies a new input vector. The SOM consists of components called neurons. Associated each neuron is a weight vector of the same dimension as the input data vector and a position in the map space. It is described a mapping from a higher-dimensional input space to a lower-dimensional map space. The process to place a vector onto the map is to find the neurons with the closest weight vector to the data space vector. The Kohonen Learning Algorithm is illustrated below:

1. *Randomize the map's nodes' weight vectors*
2. *Grab an input vector D(t)*
3. *Traverse each node in the map*
    - *Use the Euclidean distance formula to find the similarity between the input vector and the map's node's weight vector*
    - *Track the node that produces the smallest distance (this node is the best matching unit, BMU)*
4. *Update the nodes in the neighborhood of the BMU (including the BMU itself) by pulling them closer to the input vector*
    - $Wv(s+1)=Wv(s)+\theta(u,v,s)\cdot\alpha(s)(D(t)-Wv(s))$
5. *Increase s and repeat from step 2 while* $s < \lambda$

Note that s is the current iteration, λ is the iteration limit, t is the index of the target input data vector in the input data set D, is a target input data vector, v is the index of the node in the map, Wv is the current weight vector of node *v,* u is the index of the best matching unit (BMU) in the map, θ(u,v,s) is a restraint due to distance from BMU, usually called the

neighborhood function, α(s) is a learning restraint due to iteration progress.

6) Decision Tree: Data classification is the process, which finds the common properties among a set of objects in a database and classifies them into different classes, according to a classification model. The classification first analyzes the training data and develops an accurate description or a model for each class using the features available in the data (Ming-Syan et al., 1996 [15-16]). A decision-tree based classification method constructs decision trees, which is a simple way of representing the knowledge. They classify examples into a finite number of classes. Nodes in a decision tree are labeled with attribute names and edges are labeled with different classes (Sörensen et al., 2003 [2]). The tree structure is constructed following a set of decision rules applied sequentially (S. Ren., 2003 [17]). There are two phases of the decision tree generation: the growth phase and the pruning phase. The growth phase involves inducting a decision tree from the training data. The pruning phase generalizes the decision tree that was generated in the pruning phase in order to avoid over fitting (Kweku-Muata et al., 2003 [18]). Decision tree is an important knowledge structure that can result from data mining activities. The induction of decision tree is done using a supervised knowledge discovery process in which prior knowledge regarding classes in the database is used to guide the discovery. The Intelligent Miner generates decision trees, using the tree-induction algorithm, which provides an easy-to-understand description of the underlying distribution of the data. They are also used in association with other mining algorithms for better understanding and classification of the rules.

Statistical Analysis

To investigate the relationships among a group of variables, it is useful to create a model for those variables. One of the most widely accepted processes of finding a mathematical model that best fits the data is a regression analysis. In this paper, the traditional multiple regression techniques, logistic regression and fuzzy linear regression are introduced. Each of these techniques is described below.

1) Traditional multiple regression analysis: In traditional multiple regression, the dependent variable is a function of independent variables. The random error term is added to make the model probabilistic rather than deterministic, indicating the amount of variance in the dependent variable not accounted or by the linear combination of the independent variables. The value of the coefficient determines the contribution of the independent variable [19]. Moreover, regression analysis is used in determining the best-fit model for describing the relationship. In the usual conventional model, deviations between the observed values and the estimated values are supposed to be due to measurement errors or random variations. Therefore, the statistical techniques are applied for estimation and inference in regression analysis. But sometimes the deviations are due to the imprecise observed data or the indefiniteness of the system structure.

2) Logistic regression analysis: The logistic regression approach fits a model specially designed for a dichotomous dependent, qualitative variable at two levels [19]. Applications occur frequently in the social sciences, in which people or the actions of people often fall into one of two categories. The general logistic model is not a linear function of the parameters as traditional regression. Swensen (1997) [20] used logistic regression to determine whether there were associations between job attributes of force and repetitiveness. In order to test the hypotheses of no association between exposure categories, student tests were calculated to determine if there was a statistically significant difference in the overall means between low- and high-force jobs. Logistic regression techniques were used to evaluate the association between exposure variables and personal characteristics (i.e., sex, age, year on the job, etc.) in the development of the model.

3) Fuzzy linear regression analysis: The ambiguity or fuzziness of human's subjective judgment is influential and often difficult to address in predictive modeling. Therefore, effective modeling of these systems is challenging. Alternatively, the concept of fuzzy set theory seems to be applicable for modeling such systems [21-24]. Fuzzy linear regression models are introduced via the concept of possibility. In fuzzy linear regression models, deviations between the observed values and the estimated values are assumed to depend on the fuzziness of the system parameters, in contrast to the traditional linear regression analysis in which deviations are expected to be the result of observation errors [25-27]. Since the parameters of the fuzzy linear regression model are fuzzy numbers, the model can accommodate the distortion introduced by the linearization.

Table 1 compares the data mining techniques and the statistical analysis approaches in forecasting methods. As it implies, each case demands a careful selection of the analysis method. Each method also needs to be tailored in order to best suit the given data sets.

Table 1. Comparison of data mining and statistical analysis approaches in forecasting

	(1) Kohonen Neural Networks	(2) Decision Tree	(3) Traditional Multiple Regression Analysis	(4) Logistic Regression Analysis	(5) Fuzzy Linear Regression Analysis
Purpose	Reduce dimensions of data through the use of self-organize neural networks	Classify data into a finite number of classes and provide an easy-to-understand description of the underlying distribution of the data.	Predict the unknown value of a variable from the known value of two or more variables.	Obtain odds in the presence of more than one explanatory variable.	Develop vague model using the formalization of uncertainty rather than numerical intervals using fuzzy intervals.
Data Type	Unknown input data	Known data	Dependent and Independent variables	Discrete	Fuzzy data
Assumption Required	None	None	Normal Distribution Constant Variance	No outliers and inter-correlations	Crisp input and output data
Solution Approach	$Wv(s+1)=Wv(s)+\theta(u,v,s)\cdot\alpha(s)(D(t)-Wv(s))$	Rule induction Classification tree	$Y=\beta 0+\beta 1x+\beta 2x^2+\varepsilon$	Logistic function	Fuzzy regression Interval analysis Least-squares of errors
Mechanism of Defective Identification	Competitive learning	Pruning	ANOVA	Deviance and likelihood ratio tests Pseudo-R^2s Hosmer-Lemeshow test	Fuzzy coefficient
Perspective (Paradigm)	Population	Individual	Population	Population	Population

Note: Data mining includes (1) & (2) while statistical analysis includes (3) – (5)

CURRENT ESTIMATION MODEL AND DATA STRUCTURE

Each time the PCB design team receives a customer order and starts the design activities, they collect the data that are needed to complete each design. There are a total of five stages in the data entry [28-29]. Those include (1) Schematic Capture, (2) Design Layout, (3) Fabrication Documentation, (4) Assembly Documentation, and (5) Create Library Data. Depending on the complexity of design, each step demands a certain number of hours to complete. The time to complete each design step is referred to as 'the actual hours'. As design progresses, the designers log in the actual hours they spend on the design in Diary Times Entry software (see Figure 4). The Project Administration software (see "COMPUTE ACTUALS" button in Figure 5) then automatically calculates the total hours that actually took to finish the whole design process. In those five areas of consideration, there are over 83 variables to count. Some examples are given as follows. The 'Complexity_1' equals lead density of 0 to 9 square inch and includes Design Styles or Types of Simple Power Supplies, Extender Boards, Simple Analog Designs, and Low Density Backplanes. 'Complexity_4' equals lead density of 30 to 40 leads per square inch and includes Design Styles or Types of Moderate Density RF and Micro Strip Designs, Dense Digital with Multiple Buss Structures, and Dense Digital or Analog Hybrid Designs. The variable 'Job_Type' includes New_Layout, Re_Layout, Assembly_Clone (basically a copy), Documentation_Only, Schematic_Only, Permanent_Hold, Manufacturing_Clean_Up, and Assembly_Cuts. The variable 'Tech_Type' includes the BUM (Build Up Material), Flex, MCM, Rigid, Rigid Flex, Sequentially Laminated, and Other. The variable 'Design_Std' includes 12 different variables including Prototype, Commercial (Obsolete - less stringent than Mil), Mil-Std-275 (Obsolete), Mil-Std-2000 (Obsolete), IPC 6012 (Obsolete), Fast Thru-Put, IPC 6012 Class 2 (Group all "IPC" together), IPC 6012 Class 3 (Group all "IPC" together), IPC 6013 Class 3 (Group all "IPC" together), IPC 6018 Class 2 (Group all "IPC" together), IPC 6018 Class 3 (Group all "IPC" together), and Mil-P-55110 (Obsolete). The variable 'Board_Complexity' includes Simple, Average, Complex, Very Complext, and Ultra Complex. The variable 'Board_Style' includes Undefined, Interconnect, Power Supply, RF, Digital, Analog, and Combination. The variable 'Design_Standard' includes Prototype (Fast Through Put) and Commercial (Production Type). Among those stages, there are two most important activities that directly impact the working hours. They represent the Schematic Capture and the Design Layout. In Schematic Capture, a two-dimensional, symbolic illustration of electric components and the connections between the components are generated. This shows all the components in a PCB design and displays how they interconnect to each other. It can be a whole new design or a modified version of existing design. Depending on the design characteristics, complexity, and the types of design, the completion hours vary greatly. After completion of the Schematic Capture, the Layout Design is initiated. In this stage, the number of layers are determined within the PCB. Each layer in PCB acts as the electrical path ways, and it could be a single or multiple layers depending on the functions and the complexity of the design. The more the number of layers in a single PCB is, the higher the cost of fabrication and the time to complete the design are required. Our initial data analysis revealed that the variables belonged

to the two stages; namely (1) the Schematic Capture and (2) the Layout Design, take the most time in completion of the new customer-ordered PCB design. The other variable in remaining three stages are highly correlated to the first two stages. The other three stages; namely (3) the Fabrication Documentation, (4) the Assembly Documentation, and (5) the Create Library Data, are merely the subsequent works from the first two stages. They are highly dependent on the complexity of the first two stages. Hence, it has been determined that by accurately analyzing the first stages, the overall cost estimation can be significantly improved. Based on this reason, the study focuses on the most important variables in **the first two stages**.

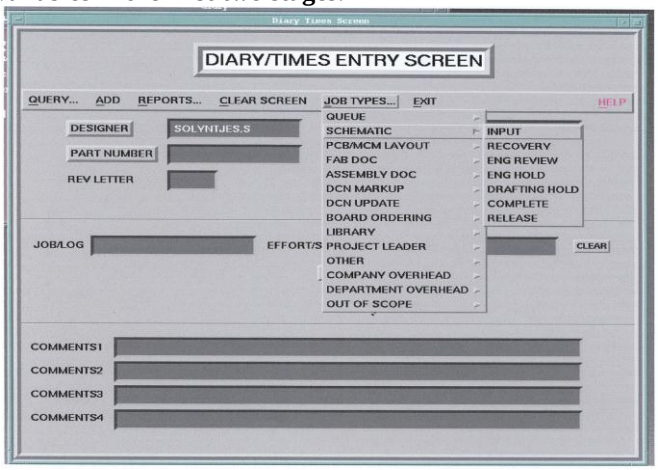

Figure 4. User interface of the Diary Times Entry software

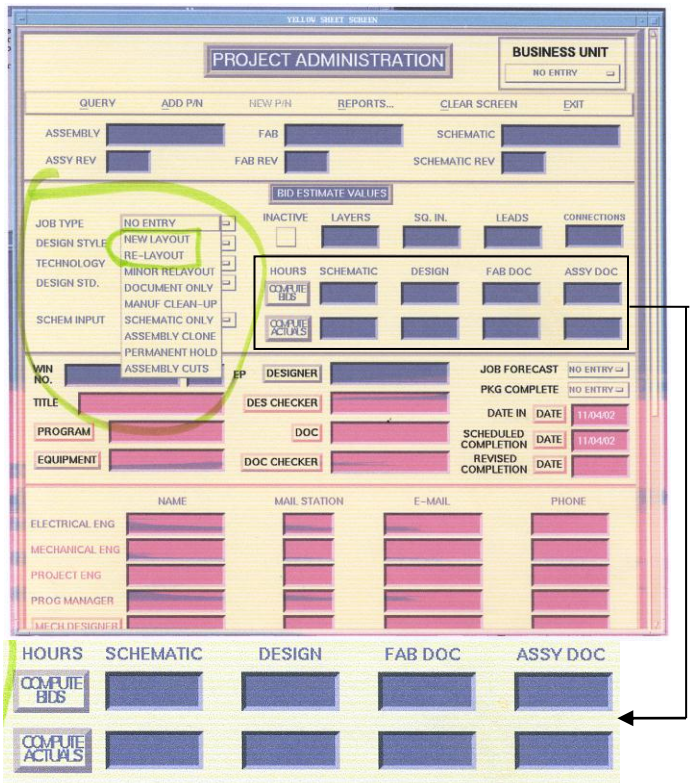

Figure 5. User interface screen of the Project Administration software and an enlarged view of COMPUTE buttons

Currently, the serious problem has been the inaccuracy of the estimation model that was developed many years ago. The model contains a set of regression equations with no more than five independent variables. Designers are required to input design variables at the beginning of the design cycle (i.e., board style, the number of connections and leads, board areas, and the number of layers) and the estimated time is calculated for four different tasks. When a new part symbol has to be included in a schematic, the hours that take to generate the symbols are also added. While actual design processes involve many variables measured and recorded by the designers, many of them were not considered in the model. Figure 6 shows the wide fluctuations in the collected data. The blue line across the graph shows the bidding hours (the estimation from the model), while the data shows too much spread to be accurately predicted. Figure 7 shows the same problem between the actual data versus predicted values (denoted as 'bid'). By looking at the graphs, one can immediately identify the current problem.

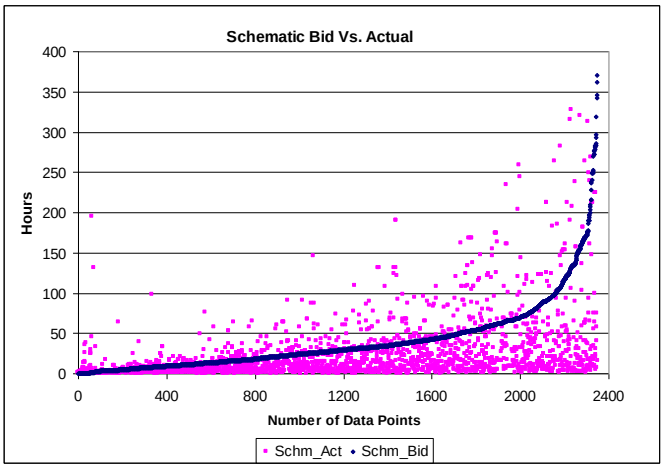

Figure 6. Schematic actual hours versus bidding data

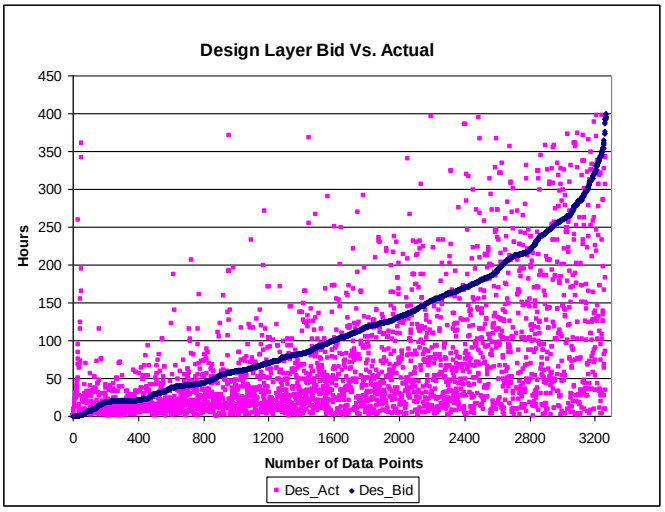

Figure 7. Design actual hours versus bidding data

In its current form, the PCB Design Work Estimating Program requires the following input from the user. The Confidence Level defines how confident the Printed Circuit Design Department is that the job can actually be completed in the estimated number of hours. So, by default, the actual number of hours to complete the design task will be met 50% of the time (it represents more likely average values). For some budgetary constraints if you need a greater level of confidence that the actual design time would be less than the estimate, you would raise the 'Confidence Level'. The Task to Estimate includes the following: Design Layout & Schematic Capture, Design Layout without Schematic (Schematic provided by Engineer), Schematic Only, Fabrication Documentation Only, Assembly Documentation Only, and Create Library Data. The Design Standard includes: Prototype PCB, Commercial PCB, Military Specification 275 PCB, and Military Specification 2000 PCB. In the Board Style, the following variables are entered: Interconnect Board, Power Supply Board, RF Board, Digital Board, Analog Board, and Combinational Board. The Number of Connections category includes the following: Board Area (in square inches), Number of Leads, and Number of Layers. The work estimating formulae are developed as follows. In Schematic Capture, two most important variables are selected: (1) the Layers and (2) the Connections. Even though, the leads and the connections are highly interrelated, only the number of connections are used in the formula. In Design Layout formula, it is quite the opposite. This time, the number of Leads are used instead of Connections. Those equations are developed using a regression analysis technique, however, no interaction effects are considered. The actual equations are provided below.

1) Current Schematic Layout estimation formula:

a) Interconnect Board = (1.3716 * Layers) + (20.2071 * Connections / 1000)

b) Power Supply Board= (2.6998 * Layers) + (22.4299 * Connections / 1000)

c) RF Board = (2.0211 * Layers) + (17.3892 * Connections / 1000)

d) Digital Board = (1.6571 * Layers) + (14.9531 * Connections / 1000)

e) Analog Board = (-2.7155 * Layers) + (56.9927 * Connections / 1000)

f) Combinational Board = (0.5401 * Layers) + (38.4504 * Connections / 1000)

2) Current Design Layout estimation formula:

a) Interconnect Board = (9.8243 * Layers) + (7.3954 * Leads / 1000)

b) Power Supply Board = (15.4635 * Layers) + (44.4044 * Leads / 1000)

c) RF Board = (27.3564 * Layers) + (-2.0232 * Area) + (60.3404 * Leads / 1000)

d) Digital Board = (14.8596 * Layers) + (-1.6101 * Area) + (40.3118 * Leads / 1000)

e) Analog Board = (69.5196 * Leads / 1000)

f) Combinational Board = (19.2603 * Layers) + (8.344 * Leads / 1000)

4. SOLUTIONS APPROACH AND RESULTS

The purpose of this section is to create a more accurate estimation model for the PCB design team. The improved model is expected to help the company becoming more competitive in their business. It is common for a today's large size data (often several gigabytes or more) that there would be missing, inconsistent, and noisy data. It is therefore necessary to preprocess the data before the analysis [21]. The data contains many missing entries and outliers. Also, very old data entries (more than 10-years old) were not considered for the analysis. It was decided with the engineers. In recent years, CAD technologies have advanced dramatically, hence the work efficiency has improved along the way. Therefore, the old data sets are no longer considered valid to be used in the estimation. In addition, the design hours exceeded 400 hours were considered as outliers. The missing entries were deleted from the data set. In our analysis, the 'Layers, Comp_Leads, SQ_Inches, Connections' variables are selected from the initial data set out of 83 variables. At the same time, the Board_Style that classifies the types of board into 0 to 6 categories is considered as the most important. From those information, the data mining equations are derived. Other methods mentioned in the previous section were considered unsuitable for this case, due to the complex nature of the data set. Therefore, only the data mining techniques were considered suitable for the analysis. Upon the analysis, the Schematic Capture and the Design Layout accuracy have been improved significantly, as compared to the original regression equations with the average accuracy of only 30%. The improved accuracy ratio has been illustrated in Tables 2 and 3.

Table 2. The accuracy of the "Schematic Capture" estimation

Board Style	CR	Total Records	Rules	No. of Records for this Rule	Accuracy Ratio
1	88	642	IF (LAYERS ≠ 16) and (LAYERS ≠ 14) and (LAYERS ≠ 11) and (LAYERS ≠ 13) and (LAYERS ≠ 12) and (LAYERS ≠ 9) ***Then*** value<=15	624	92
			IF (LAYERS = 16) or (LAYERS = 14) or (LAYERS = 11) or (LAYERS = 13) or (LAYERS = 12) or (LAYERS = 9) **AND** SQ_INCHES ≥ 59.45 ***Then*** 50<=Value<150	18	60
2	71	73	IF COMP_LEADS<335 Then value<=10	44	98
			IF COMP_LEADS ≥ 335 **AND** (LAYERS = 3) or (LAYERS = 8) or (LAYERS = 14) or (LAYERS = 10) ***Then*** 25<=Value<100	29	60
3	72	164	IF CONNECTIONS 355 **Then** value<=10	139	91
			IF CONNECTIONS ≥ 355 **AND** (LAYERS = 9) or (LAYERS = 3) **Then** 10<Value<=25	25	76
4	72	275	IF (LAYERS = 2) or (LAYERS = 4) or (LAYERS = 1) or (LAYERS = 0) or (LAYERS = 3) or (LAYERS = 7) or (LAYERS = "") **Then** Value<=10	154	91
			IF (LAYERS ≠ 2) and (LAYERS ≠ 4) and (LAYERS ≠ 1) and (LAYERS ≠ 0) and (LAYERS ≠ 3) and (LAYERS ≠ 7) and (LAYERS ≠ "") and 1. SQ_INCHES ≥ 23.07 **AND** COMP_LEADS<2,714 **AND** CONNECTIONS ≥ 1,508.5 then value>10 and <50 2. (LAYERS ≠ 12) **and** (LAYERS ≠ 10) **and** (LAYERS ≠ 11) **and** (LAYERS ≠ 9) **AND** SQ_INCHES<23.07 **Then** value<=10	19 102	74 72
5	88	171	IF CONNECTIONS<1,051.5 **Then** value<=25	163	94
			IF CONNECTIONS ≥ 1,051.5 and (LAYERS = 9) or (LAYERS = 10) **Then** 25<value<=50	8	90
6	70	70	IF COMP_LEADS ≥ 596.5 and COMP_LEADS 2,417.5 **Then** 15<value<=25	5	100
			IF COMP_LEADS ≥ 3,545 **and** (LAYERS ≠ 8) **and** (LAYERS ≠ 9) **and** (LAYERS ≠ 2) **and** (LAYERS ≠ 4) **and** (LAYERS ≠ 11) **AND** SQ_INCHES ≥ 40.995 **AND** CONNECTIONS ≥ 2,975.5 **Then** 15<value<50	9	76
			IF COMP_LEADS ≥ 2,417.5 **AND** (LAYERS ≠ 8) **and** (LAYERS ≠ 9) **and** (LAYERS ≠ 2) and (LAYERS ≠ 4) **and** (LAYERS ≠ 11) and SQ_INCHES 40.995 **AND** CONNECTIONS 3,254.5 **Then** 50<=value<200	56	52

NOTE: Field CR implies Classification Ratio of that particular model.

Table 3. The accuracy of the "Design Layout" estimation

Board Style	CR	Total Records	Rules	No. of Records for this Rule	Accuracy Ratio
1	77	574	IF ((LAYERS ≠ 12) and (LAYERS ≠ 16) and (LAYERS ≠ 11) **and** (LAYERS ≠ 10) and (LAYERS ≠ 14) **and** (LAYERS ≠ 8) **and** (LAYERS ≠ 13) **and** (LAYERS ≠ 18) **and** (LAYERS ≠ 20)) **Then** value<=50	524	85
			IF ((LAYERS = 12) or (LAYERS = 16) or (LAYERS = 11) or (LAYERS = 10) or (LAYERS = 14) or (LAYERS = 8) or (LAYERS = 13) or (LAYERS = 18) or (LAYERS = 20)) **AND** COMP_LEADS ≥ 2,230.5 **Then** 100<value<=500	50	60
2	68	86	IF ((LAYERS = 6) or (LAYERS = 8) or (LAYERS = 14) or (LAYERS = 10) or (LAYERS = 7)) **and** CONNECTIONS< 193 **Then** value<=50	6	100
			IF ((LAYERS = 6) or (LAYERS = 8) or (LAYERS = 14) or (LAYERS = 10) or (LAYERS = 7)) **AND** 193<=CONNECTIONS <1614.5 **AND** COMP_LEADS ≥ 1,434 **Then** 100<value<=200	5	100
			IF ((LAYERS = 6) or (LAYERS = 8) or (LAYERS = 14) or (LAYERS = 10) or (LAYERS = 7)) **AND** CONNECTIONS ≥ 1,164.5 **AND** SQ_INCHES<27.93 **Then** value<=50	7	86
			IF (LAYERS ≠ 6) **and** (LAYERS ≠ 8) **and** (LAYERS ≠ 14) **and** (LAYERS ≠ 10) **and** (LAYERS ≠ 7) **Then** value<=50	58	80
			IF ((LAYERS = 6) or (LAYERS = 8) or (LAYERS = 14) or (LAYERS = 10) or (LAYERS = 7)) **and** 193<=CONNECTIONS< 1,164.5 **and** COMP_LEADS 1,434 **and** LAYERS=3 **Then** 50<value<=100	10	60
3	73	292	IF ((LAYERS ≠ 2) **and** (LAYERS ≠ 4) **and** (LAYERS ≠ 3) **and** (LAYERS ≠ 9) **and** (LAYERS ≠ 1) and (LAYERS ≠ 14) and (LAYERS ≠ "")) **and** CONNECTIONS<3,029.5 **and** COMP_LEADS ≥ 2,687.5 **Then** value<=50	6	100
			IF ((LAYERS = 2) or (LAYERS = 4) or (LAYERS = 3) or (LAYERS = 9) or (LAYERS = 1) or (LAYERS = 14) or (LAYERS = "")) **Then** value<=50	269	78
			IF ((LAYERS ≠ 2) **and** (LAYERS ≠ 4) **and** (LAYERS ≠ 3) and (LAYERS ≠ 9) **and** (LAYERS ≠ 1) **and** (LAYERS ≠ 14) **and** (LAYERS ≠ "")) **and** CONNECTIONS ≥ 3,029.5 **Then** 100<value<=500	17	71
4	72	232	IF CONNECTIONS<1,627 **and** ((LAYERS = 2) or (LAYERS = 1) or (LAYERS = 4) or (LAYERS = 0) or (LAYERS = 11) or (LAYERS = "")) **Then** value<=50	132	90
			IF ((LAYERS ≠ 2) **and** (LAYERS ≠ 1) **and** (LAYERS ≠ 4) **and** (LAYERS ≠ 0) **and** (LAYERS ≠ 11) **and** (LAYERS ≠ "")) **and** (1,043<CONNECTIONS <1,206.5) **and** CONNECTIONS ≥ 1,043 **Then** 50<value<=100	21	72
			IF CONNECTIONS ≥ 1,627 **Then** 100<value<=500	79	71
5	84	121	IF ((LAYERS ≠ 9) **and** (LAYERS ≠ 6) **and** (LAYERS ≠ 12) **and** (LAYERS ≠ 4) **and** (LAYERS ≠ 2)) **and** CONNECTIONS ≥ 418.5 **and** SQ_INCHES ≥ 51.645 **Then** 100<value<=200	4	100
			IF CONNECTIONS 418.5 **Then** value<=50	104	96

			IF ((LAYERS = 9) or (LAYERS = 6) or (LAYERS = 12) or (LAYERS = 4) or (LAYERS = 2)) **and** (CONNECTIONS ≥ 418.5) **and** SQ_INCHES ≥ 51.645 **Then** 50<value<=100	13	85
6	68	139	IF CONNECTIONS ≥ 1,695 **and** SQ_INCHES< 7.01 **and** COMP_LEADS 1,832.5 **Then** 700<value<=1000	6	100
			IF CONNECTIONS ≥ 1,695 **and** SQ_INCHES 7.01 **and** COMP_LEADS ≥ 1,832.5 **Then** 50<value<=100	6	100
			IF 1695<CONNECTIONS<3279 **and** SQ_INCHES ≥ 7.01 **and** 2,840 <COMP_LEADS <2970 **Then** 50<value<=100	8	100
			IF ((LAYERS ≠ 2) **and** (LAYERS ≠ 4) **and** (LAYERS ≠ 6) **and** (LAYERS ≠ 3) **and** (LAYERS ≠ 14) **and** (LAYERS ≠ 9) **and** (LAYERS ≠ 0) **and** (LAYERS ≠ "")) **and** (1676.5<CONNECTIONS<1695) **and** (SQ_INCHES<55.225) **and** COMP_LEADS ≥ 1,932.5 **Then** value<=50	7	86
			IF 1695<CONNECTIONS<3279 **and** SQ_INCHES ≥ 7.01 **and** COMP_LEADS ≥ 2,840 **Then** 100<value<=500	112	76

NOTE: Field CR implies Classification Ratio of that particular model.

Based on computational results from Table 2 and Table 3, we could conclude that the decision rules in "Schematic Capture" (i.e., 1,395 in total) and "Design Layout" (i.e., 2,841 in total) categories have been derived respectively. In general, all inductived rules are corresponding to each board style. Per derived rules, the decision maker could observe that these rules include simple and complex rules based on number of "conditions" in the rule. For instance, if number of conditions is less than or equal to "5" then it is defined as a simple rule and vice versa. Moreover, number of logic junctions (e.g., "AND" & "OR") in each rule could be observed as well. For instance, A(5) means 5 "AND" is the rule and the similar situation is applied to "OR." Through the aforementioned represtations, the decision maker could easily fathom the complexity of the rule. At last, the aggregated accuracy under each category and each board type is also derived. Due to complex interaction effects among data variables, the accuracy of the data mining approach exceeds over 80% (Note: 83% for the category of "Schematic Capture" and the category of "Design Layout.") while the statistical method only yields 10% to 5% of prediction accuracy. Using our proposed method, it is more likely that the company can better reply to the customer request with an improved cost estimation. This will in turn attribute to the enhanced opportunity of the company in terms of securing customer orders with a higher level of profit.

Table 4. Summary based on "Schematic Capture" and Design Layout estimations

Rules in Schematic Capture	Simple Rule (S) v.s. Complex Rule (C)	# of AND v.s. # of OR	Agregated Accuracy	**Rules in Design Layout**	Simple Rule (S) v.s. Complex Rule (C)	# of AND v.s. # of OR	Agregated Accuracy
SC-1	C, C	A(5), O(5)A(1)	0.911	**DL-1**	C, C	A(8), O(7)A(1)	0.828
SC-2	S, S	Null, A(1)O(3)	0.829	**DL-2**	C, C, C, S, C	O(4)A(1), O(4)A(2), O(4)A(2), A(4), O(4)A(3)	0.807
SC-3	S, S	Null, A(1)O(1)	0.887	**DL-3**	C, C, C	A(8), O(6), A(7)	0.780
SC-4	C, C, C	O(6), A(8)+A(10)	0.828	**DL-4**	C, C, S	A(1)O(5), A(7), Null	0.819
SC-5	S, S	Null, A(1)O(1)	0.938	**DL-5**	C, S, C	A(6), Null, O(4)A(2)	0.950
SC-6	S, C, C	A(1), A(7), A(7)	0.585	**DL-6**	S, S, S, C, S	A(2), A(2), A(2), A(10), A(2)	0.800

Remark:
(1) The threshold value to determine either a "Simple Rule" or a "Complex Rule" is based on number of "conditions" in the rule; If it is less than or equal to "5" then it is a Simple Rule and vice versa.
(2) A(N) represents "N" logic junctions called "AND" in the rule. For instance, A(5) means 5 "AND" is the rule. The similar situation is applied to O(N).
(3) There are two sub-rules in SC-4, therefore, number of "AND" is depicted as A(8)+A(10)

CONCLUSIONS AND DISCUSSION

In this study, we analyzed the complex data set that spans many years of collection period. It can be reasonably assumed that, since we are living in the age of information and communication technologies, the collection of huge data sets and the subsequent use of them would be accelerating. In the past, such acts were too costly and there were no adequate means of collecting a vast array of data variables. With the data in-hands, the company was having a great trouble in terms of very low prediction accuracy of the estimation model [30]. Even though the estimation model was good from the perspectives of conventional statistical methods, it was not adequate for the company in order to effectively compete in the market place [31-32]. By this context, this study reconstructed the estimation model using the data mining approach. It clearly exceeds the conventional method of regression equations, and illustrates a much improved estimation accuracy. The complex array of data set precludes the other methods. It can be foreseeable that in the future, such trends will even accelerate and the use of advanced data analysis method will become even more important.

ACKNOWLEDGEMENT

This work was supported by the Ajou University Research Fund and the National Science Foundation (DUE-TUES-1246050). The authors would like to express their sincere gratitude towards the financial support.

REFERENCES

[1] Chung, H. Michael, Fred Gey, and Selwyn Piramuthu. "Data Mining and Information Retrieval." *System Sciences, 2002. HICSS. Proceedings of the 35th Annual Hawaii International Conference on*. IEEE, 2002.

[2] Sörensen, Kenneth, and Gerrit K. Janssens. "Data mining with genetic algorithms on binary trees." *European Journal of Operational Research* 151.2 (2003): 253-264.

[3] Wei, Chih-Ping, Selwyn Piramuthu, and Michael J. Shaw. "Knowledge discovery and data mining." *Handbook on Knowledge Management*. Springer Berlin Heidelberg, 2003. 157-189.

[4] Heaton, Jeff. "Understanding the Kohonen Neural Network. Introduction to Neural Networks with Java. Heaton Research." (2005).

[5] Chang, Li-Yen. "Analysis of freeway accident frequencies: negative binomial regression versus artificial neural network." *Safety science* 43.8 (2005): 541-557.

[6] Hinton, Geoffrey, et al. "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups." *IEEE Signal Processing Magazine* 29.6 (2012): 82-97.

[7] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." *Nature* 521.7553 (2015): 436-444.

[8] Sun, Wenqing, et al. "Enhancing deep convolutional neural network scheme for breast cancer diagnosis with unlabeled data." *Computerized Medical Imaging and Graphics* (2016).

[9] Zhang, H-C., and S. H. Huang. "Applications of neural networks in manufacturing: a state-of-the-art survey." *THE INTERNATIONAL JOURNAL OF PRODUCTION RESEARCH* 33.3 (1995): 705-728.

[10] Dagli, Cihan H., ed. *Artificial neural networks for intelligent manufacturing*. Springer Science & Business Media, 2012.

[11] Gan, XuSheng, Jin-Liang Chen, and Hai-Tao Zhao. "Prediction of Aircraft Collision Unsafe Event Based on Hopfield Neural Network Model." *2015 International Symposium on Computers & Informatics*. Atlantis Press, 2015.

[12] Hopfield, John J., and David W. Tank. "“Neural” computation of decisions in optimization problems." *Biological cybernetics* 52.3 (1985): 141-152.

[13] Yin, Fei, et al. "Back Propagation neural network modeling for warpage prediction and optimization of plastic products during injection molding." *Materials & design* 32.4 (2011): 1844-1850.

[14] Rost, Burkhard, and Chris Sander. "Bridging the protein sequence-structure gap by structure predictions." *Annual review of biophysics and biomolecular structure* 25.1 (1996): 113-136.

[15] Heaton, Jeff. "Programming neural networks in Java." *bttp://www. heatonresearch. com* (2004).

[16] Lin, Keng-Pei, and Ming-Syan Chen. "Releasing the svm classifier with privacy-preservation." *2008 Eighth IEEE International Conference on Data Mining*. IEEE, 2008.

[17] Ren, Shijin. "Phenol mechanism of toxic action classification and prediction: a decision tree approach." *Toxicology letters* 144.3 (2003): 313-323.

[18] Kweku, Muata, and Bryson Osei. "EvaIuation of decision trees: a multi huntencriteria approach." (2003).

[19] Menard, Scott. *Applied logistic regression analysis*. No. 106. Sage, 2002.

[20] Swensen, Stephen J., et al. "The probability of malignancy in solitary pulmonary nodules: application to small radiologically indeterminate nodules." *Archives of internal medicine* 157.8 (1997): 849-855.

[21] Han, Jiawei, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[22] Misra, Janardan, and Indranil Saha. "Artificial neural networks in hardware: A survey of two decades of progress." Neurocomputing 74.1 (2010): 239-255.

[23] Sundaravaradan, Naren, et al. "Data mining approaches for life cycle assessment." *IEEE ISSST* 11 (2011).

[24] Tsai, Tsung-Nan. "Development of a soldering quality classifier system using a hybrid data mining approach." *Expert Systems with Applications* 39.5 (2012): 5727-5738.

[25] Chang, Pei-Chann, Chin-Yuan Fan, and Jyun-Jie Lin. "Monthly electricity demand forecasting based on a weighted evolving fuzzy neural network

approach." *International Journal of Electrical Power & Energy Systems* 33.1 (2011): 17-27.

[26] Kuo, R. J., S. Y. Hong, and Y. C. Huang. "Integration of particle swarm optimization-based fuzzy neural network and artificial neural network for supplier selection." *Applied Mathematical Modelling* 34.12 (2010): 3976-3990.

[27] Basu, Jayanta Kumar, Debnath Bhattacharyya, and Tai-hoon Kim. "Use of artificial neural network in pattern recognition." International Journal of Software Engineering and Its Applications 4.2 (2010).

[28] Kuo, Chung-Jen, Chen-Fu Chien, and Jan-Daw Chen. "Manufacturing intelligence to exploit the value of production and tool data to reduce cycle time." *IEEE Transactions on Automation Science and Engineering* 8.1 (2011): 103-111.

[29] Köksal, Gülser, İnci Batmaz, and Murat Caner Testik. "A review of data mining applications for quality improvement in manufacturing industry." *Expert systems with Applications* 38.10 (2011): 13448-13467.

[30] Mouelhi-Chibani, Wiem, and Henri Pierreval. "Training a neural network to select dispatching rules in real time." *Computers & Industrial Engineering*58.2 (2010): 249-256.

[31] Liu, Miao, et al. "Comparison of random forest, support vector machine and back propagation neural network for electronic tongue data classification: Application to the recognition of orange beverage and Chinese vinegar." *Sensors and Actuators B: Chemical* 177 (2013): 970-980.

[32] Hunter, David, et al. "Selection of proper neural network sizes and architectures—a comparative study." *IEEE Transactions on Industrial Informatics* 8.2 (2012): 228-240.