A HYBRID INTELLIGENT SYSTEM FOR DATA ANALYSIS AND VISUALIZATION

by

WANG SHIR LI

Thesis submitted in fulfillment of the requirements

for the degree of

Master of Science

March 2007

ACKNOWLEDGEMENTS

I would like to express my heartfelt gratitude to both of my supervisors, Dr Zalina Abdul Aziz and Dr. Lim Chee Peng for their supports and motivations. Their guidance and incisive advice have inspired me to generate fruitful approaches in achieving the objective in this research. Without their efforts, I would not able to bring this research to a completion.

My gratitude is extended to my beloved father, brother and sister that support and encourage me all along without hesitation. I treasure dearly their encouragement and moral support.

Last but not least, I would like to express my sincere gratitude to my friends, Koay Fong Thai, Ng Theam Foo, Lim Say Yarn, Lim Wei Lee, Goh Wei Chen, Tan Yean Ching, Tang Yeng Hok, Pua Ling Lia and all the ex-QCI friends, for their unlimited support and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

List of Abbreviations

ANN	Artificial Neural Network
ANOVA	Analysis of Variance
CCD	Central Composite Design
DOE	Design of Experiment
EDM	Electrical Discharge Machining
GA	Genetic Algorithm
MLP	Multilayer Perceptron
MSE	Mean Squared Error
PCA	Principal Component Analysis
PID	Proportional-Integral-Derivative
RBF	Radial Basis Function
RSM	Response Surface Methodology
SMT	Surface Mount Technology
SOM	Self-Organizing Map
SVM	Support Vector Machines
WEDM	Wire Electrical Discharge Machining

SATU SISTEM CERDIK HIBRID UNTUK ANALISIS DAN PENGLIHATAN DATA

ABSTRAK

Satu sistem hibrid yang menggabungkan rangkaian neural perceptron berbilang lapisan (MLP) dan ruas-bulatan (circle-segments) telah dibangunkan dalam penyelidikan ini. Rangkain MLP yang bersifat seperti “kotak hitam” memberi ramalan tanpa sebarang penjelasan. Tanpa dilengkapi dengan peralatan yang bersifat gambaran, pengguna menghadapi masalah untuk mengestrak maklumat dan memahami penyelesaian yang diberikan oleh rangkaian tersebut. Oleh yang demikian, ruas-bulatan digunakan untuk memberi sekaitan yang bersifat gambaran di antara data input-output, dan seterusnya digunakan untuk membezakan input yang tidak penting.

Keberkesanan sistem hibrid tersebut diuji dengan kajian kes yang sebenar dan kajian kes benchmark. Kajian kes dalam bidang pemodelan dan ramalan merangkumi Friedman#1, proses pemesinan nyahcas elektrik dawai (wire EDM), sistem kawalan cakera keras, and pengawalan PID. Kes kajian dalam bidang pengelasan terdiri daripada kes benchmark Iris dan Wine, dan akhir sekali strok diagnosis.

Keberkesanan sistem hibrid tersebut dalam masalah pemodelan and ramalan dibandingkan dengan metodologi permukaan sambutan (RSM). Sistem hibrid tersebut mencapai ketepatan sekurang-kurangnya 11% berbanding dengan RSM. Dalam masalah pengelasan, sistem hibrid tersebut mencapai tetepatan yang setanding atau lebih baik daripada MLP yang bergabung dengan analisis komponen utama (PCA) dan MLP tanpa bergabung dengan sebarang kaedah penyarian sifat. Sistem MLP-ruas bulatan berkesan dalam penganalisisan dan penglihatan data, dan ini terbukti dengan keupayaan ramalan dan penglihatannya yang lebih baik. Sistem hibrid ini memainkan peranan penting dalam analisis and penglihatan data, terutamanya dalam bidang pemodelan dan pengelasan.

A HYBRID INTELLIGENT SYSTEM FOR DATA ANALYSIS AND VISUALIZATION

ABSTRACT

In this research, a hybrid system consisting of the multilayer perceptron (MLP) neural network and the circle-segments method for data analysis and visualization is designed and developed. Acting as a black box, the MLP network normally gives a prediction without providing a facility for users to visualize the solution. As such, this research proposes to hybrid the circle-segments method with the MLP network, whereby the circle-segments method is used in two different ways, i.e., to provide visual correlation between the input-output data samples to users and, thus, to allow users to eliminate insignificant inputs from the input data set.

The effectiveness of the proposed MLP-circle segments system is evaluated using a number of benchmark and real case studies. For process modelling and prediction problems, the case studies investigated include the Friedman#1 benchmark problem, wire electrical discharge machining (EDM) process, disk drive read control system, and PID controller tuning. For data classification problems, the case studies investigated include the Iris, and Wine benchmark problems, as well as a real medical problem pertaining to acute stroke diagnosis.

In process modelling and prediction problems, the performances of the hybrid system are compared with those from the response surface methodology (RSM). The proposed system achieves an improvement of at least 11% in term of accuracy as compared with the results from the RSM. In data classification problems, the results are compared with those from MLP coupled with the principal component analysis (PCA) as well as MLP without any feature selection method. It is found that the accuracy of proposed system is as good as, if not better than, MLP coupled with the PCA and MLP without any feature selection method. Based on the results obtained, the proposed MLP-circle segments system demonstrates better prediction and visualization

abilities, thus justifying its potentials as a useful and effective system for data analysis and visualization. The proposed MLP-circle segments system is useful in data analysis and visualization, especially in the domain of process modelling and prediction, as well as data classification problems.

CHAPTER 1

INTRODUCTION

1.1 Preliminaries

Intelligent data exploration and analysis in process modelling and prediction as well as classification has attracted a lot of interests in a variety of domains, e.g. pharmaceutical, biochemistry, food research, mechanical engineering, manufacturing technology, and medical diagnosis (Conforti, 1999; Lerner *et al*., 1994; Kavzoglu & Mather, 2000; Sasikala & Kumaravel, 2005; Peh *et al*., 2000; Lim *et al*., 2003; Dutta *et al*., 2004; Lou & Nakai, 2001; Bourquin *et al*., 1998; Spedding & Wang, 1997). Scientists and/or engineers are interested to model the processes they are interested in because the task allows prediction and classification to be made about new inputs and also the result of the changes quickly.

Artificial neural networks (ANNs) have gained popularity in solving data exploration and analysis problems in various fields (as mentioned above), owing to its ability to work like human brain. Among the various types of ANN models, the Multi-Layer Perceptron (MLP) is one of the widely used model. It is considered as a model-free approach in building a learning system for data analysis and prediction. As published in the literature, the MLP has been successfully applied to a variety of tasks, including modelling and prediction in injection moulding process, stencil-printing process, the mechanical properties of steels, heated catalytic converter, atmospheric plasma spray process, drug dissolution profiles, wire electrical discharge machining process and the thermal inactivation of bacteria (Yarlagadda & Khong, 2001; Yang *et al*., 2005; Sterjovski *et al*., 2005; Akcayol & Cinar, 2005; Guessasma *et al*., 2003; Peh *et al*., 2000; Spedding & Wang, 1995; Lou & Nakai, 2001). The MLP is generally easy to use and good in approximating any input/output map (Barletta & Grisario, 2006).

The MLP network learns from data and builds a predictive model even there is lack of information regarding the underlying model that generates the data samples. Generally, the MLP network (as well as a lot of other types of ANN) acts as a black-box that provides a solution without any explanation. As a consequence, a new problem emerges, that is the MLP network does not have an effective way for domain users to visualize the resulting solution. This problem becomes more serious when it involves many attributes which cannot be projected into two-dimensional or three-dimensional plots. The user thus cannot gain much information from the solution.

For data exploration and analysis to be more effective, it is important to include the user in the process since humans have flexibility, creativity, and common sense in analyzing data. Indeed, data visualization is increasing of importance in exploring and analyzing multidimensional data. With the help of data visualization, the user is allowed to assess, identify, compare, verify, and understand the possible hypothesis in a data set. Data visualization thus exploits the abilities of the user in extracting information from the data set.

Data visualization is intuitive and straightforward. In addition, data visualization provides an overview which displays all the relevant attributes in a glance. An overview of a large data set enables one to identify patterns or relations, which are hardly achievable when all the attributes are not presented in the same space. Data visualization can also deal with missing attributes by allocating certain characteristics, i.e., colour, to represent the missing attributes. Furthermore, data visualization is able to display properties of data that have complex relation or possibly pattern that are not obtainable from the resulting prediction of ANN. As a result, the role of data visualization is important in helping the user understand phenomena of interest in data.

The information discovery based on data visualization in a data set is not a black box based on some searching algorithms that return information about the data but rather an interactive process involving a human (Fayyad *et al*., 2002). With the help of data visualization, human beings look for patterns and relationships in data, and thus, extract information from it. Therefore, the interpretation based on data visualization is rather subjective. Data visualization is used as an assisting tool to help humans understand the underlying data based on their own interpretation.

The circle-segments method is chosen among various data visualization techniques because the method is not limited by the number of attributes involved and can be used to display numeric data. Unlike graph-based category, i.e., bar chart and *x-y* plot, the circle-segments method allows the representation of the whole data set in a plot even the number of attributes exceeds two. The circle-segments is used to map each data values to a coloured pixel and present the data values that belong to one dimension in a separate sub-window. The related dimension is placed to each other to form a circle. Such structure allows easier detection of dependencies and correlation between the dimensions represented in the sub-window (Keim, 2000). Besides that, the circle-segments method is not affected by the overlapping problem which occurs in the implementation of parallel-coordinates.

In the previous work, the circle-segments method was used to display the history of a stock data (Ankerst *et al*., 1996). In this research, the circle-segments method is used in a different way, whereby it is used to look for possible relationship between the inputs and output(s). It is used to provide visual relation in the input-output data samples, and to identify the domain input features in a classification problem. The use of the circle-segments method depends on the problem domain, which can be categorized as methods 1 and 2. More explanations on these two methods are described in sections 3.3.1 and 3.3.2.

1.2 Problem Statement

Most of the processes in the real world involve multiple inputs and outputs, or known as factors and responses, respectively, according to the statistical concept. Outputs are usually variables that one is interested to investigate. On the other hand, inputs are variables that cause some effects toward the output if some changes have been made on it. For example, an engineer is interested in the effects of cutting speed, tool geometry, and cutting angle on the life of a machine tool. In this case, the life of a machine tool is known as the output (response), while the cutting speed, tool geometry, and cutting angle are knows as the inputs (factors).

The underlying relationship between the inputs and outputs (factors and responses) is often unknown. A lot of effort has been spent on developing methods that are able to model the relationship. Thus, process modelling and prediction is important in many industrial applications, and this task cannot be characterized by "hit or miss". Various methods have been applied to solve this problem, as the solution provides a way on understanding how the process works.

In general, there are two common strategies of experimentation frequently practiced by engineers and scientists, i.e., the *best-guess* approach and the *one-factor-at-a-time* approach. The *best-guess* approach depends greatly on the technical or theoretical knowledge of the process or system under investigation, as well as the experimenter's practical experience. If the initial best-guess of the combination of factors does not produce the desired or expected results, the experimenter needs to take another best-guess. This method is not a good practice because it seems like the "hit or miss" game. Furthermore, the *best-guess* process may keep on continuing if the desired or expected results do not appear. As for the consequences, a lot of times, efforts, materials are wasted.

As for the *one-factor-at-a-time* approach, it consists of selecting a baseline sets of levels for each factor, then varying each factor over its range with the other factors held constant at the baseline level until an improvement is observed. The disadvantage of this method is it neglects any possible interaction between the factors. An interaction is the failure of one factor to produce same effect on the response at different levels of another factor. This method is most frequently practiced in biotechnology, e.g. for improving fermentation condition (Dutta *et al.*, 2004)

The DOE (design of experiment) is one of the most powerful statistical techniques for improving quality and increasing productivity. The DOE is a process of planning the experiment so that appropriate data which can be analyzed by statistical methods are collected, resulting in valid and objective conclusions. Through DOE, changes are intentionally introduced into the process or system in order to observe their effect on the performance characteristics of the system or process. The introduced changes are known as *factor(s)* while the performance characteristics are known as *response(s)*. A statistical approach is the most efficient method for optimizing these changes. Otherwise, improper engineering experimentation may end up wasting time, money, manpower, material, machines and other inputs that are needed to complete the experiment. There are various types of DOE that are applicable such as 2^k factorial design, 3^k factorial design, the response surface methodology (RSM), the Taguchi method, and other types of designs (Montgomery, 2001).

The MLP network and the response surface methodology (RSM) are among the widely used methods in process modelling and prediction. Both methods build their own predictive models through a set of experimental data. However, both these methods have their own limitations. The RSM is used to fit polynomial problems, but the maximum order of polynomials that it is able to form is only two (second-order). For

problems which are non-linear, the RSM is not able to perform adequately and this limitation is known as the order problem. Besides that, the RSM suffers from another limitation, known as the dimension problem. The RSM can only provide two-dimensional (contour plot) or three-dimensional (response surface plot) graphs for visualization purposes; thus it is limited to display two or three factors and responses involved in the study. If a problem involves more three factors and responses, one can visualize the fitted response against two factors only while fixing the remaining factors at a certain level. The problem becomes worse if multi-responses are involved. In other words, the RSM is not able to project high dimensional data into a low dimensional geometry, such as two-dimensional or three-dimensional graphs for data visualization.

As opposed to the RSM, the MLP network is not limited by the order problem. With the provided data, the MLP network is able to learn and form the relationship between the pairs of inputs and outputs, and thus provides the solutions. However, a user can only accept or reject the provided solution without gaining much information from the underlying data. It is difficult for the user to make the next move if he/she is not able to get an overall picture of the data. As a result, integrating the MLP network with a data visualization facility is needed so that the user can visualize the solution. Indeed, data visualization provides a qualitative overview of a large and complex data set, summarizes data, and assists identifying regions of interest and appropriate features for more focused analysis.

1.3 Motivation

Data visualization can be used to explore and search for the possible hypothesis within a data set. During data exploration, the user is searching for patterns

or relationships and is attempting to arrive at some hypothesis. In a confirmatory visualization, the user has a hypothesis that need to be tested. For example, a number of features are predetermined through a data visualization technique. Then analytic tools, e.g. the MLP network, are used to confirm or refute the hypothesis.

The MLP network is powerful in modelling data, while data visualization is useful in helping the user to understand the phenomena interest in data. Both these methods should be integrated to make the resulting solution convincing since such a system has the visualization ability that allows the user to access, identify, and understand the possible relationships with a data set. Thus, the system provides useful information for the user to make a decision based on visualization.

From the above explanation, the primary motivation of this research is, therefore, to research into a hybrid system consisting of ANN and data visualization, and to apply the proposed hybrid system to modelling and predication, as well as classification tasks. The resulting system has practical benefits for scientists/engineers who are interested to understand or gain more information from their processes or products. Gaining better understanding enables them to have better control, hence improve the quality of processes or products.

1.4 Research Objectives

The main aim of this research is to combine the strengths of the MLP network and data visualization into a hybrid system for data analysis and visualization. The proposed system not only has the ability to approximate the relationship in a data set, but also the ability to provide visualization of the predicted solution. The specific research objectives include

- To develop a hybrid system that integrates the MLP network and the circle-segments method;
- To evaluate the applicability of the proposed system to modelling and prediction tasks;
- To demonstrate the use of the proposed data visualization technique as a feature selection tool for the MLP network in data classification tasks;

1.5 Research Scope and Methodology

Figure 1.1 shows a flowchart of overall work stages involved in this research. The goal of this research is to develop an integration of the MLP network and data visualization. The motivation is to provide visualization ability to the ANN so that the opaqueness of the network solution can be reduced. With the visualization facility, the user can extract information rather than depends on the numerical solution blindly. To evaluate the effectiveness of such integration, a series of experiments on modelling and prediction, as well as classification comprising data obtained from public domains, simulation, and real applications are carried out.

In this research, the MLP network is used particularly to predict and classify the data. On the other hand, the circle-segments method is used to visualize the phenomena of interest in the data. The circle-segments method is coupled with the MLP network to perform in two different ways: to provide visual correlation between the attributes, and to eliminate the insignificant attributes from the original data set.

In handling modelling and prediction tasks, the proposed MLP-circle segment system is not only used to model and predict multi-factor single-response process, but

also multi-factor multi-response process. The effectiveness of the proposed system is compared with the RSM in terms of prediction and visualization abilities.

In addition, the integration proposed system is used to tackle data classification problems, in which unimportant input features in a data set can be pruned by identifying patterns from the circle-segments method. The performance of the MLP network coupled with the circle-segments is compared with the MLP network coupled with the PCA and the MLP network without any feature selection method. Table 1.1 shows the summary of the case studies that have been carried out. Based on the Table 1.1, there are two ways of using the circle-segments method, which can be known as methods 1 and 2. More explanations on these two methods are explained in sections 3.3.1 and 3.3.2.

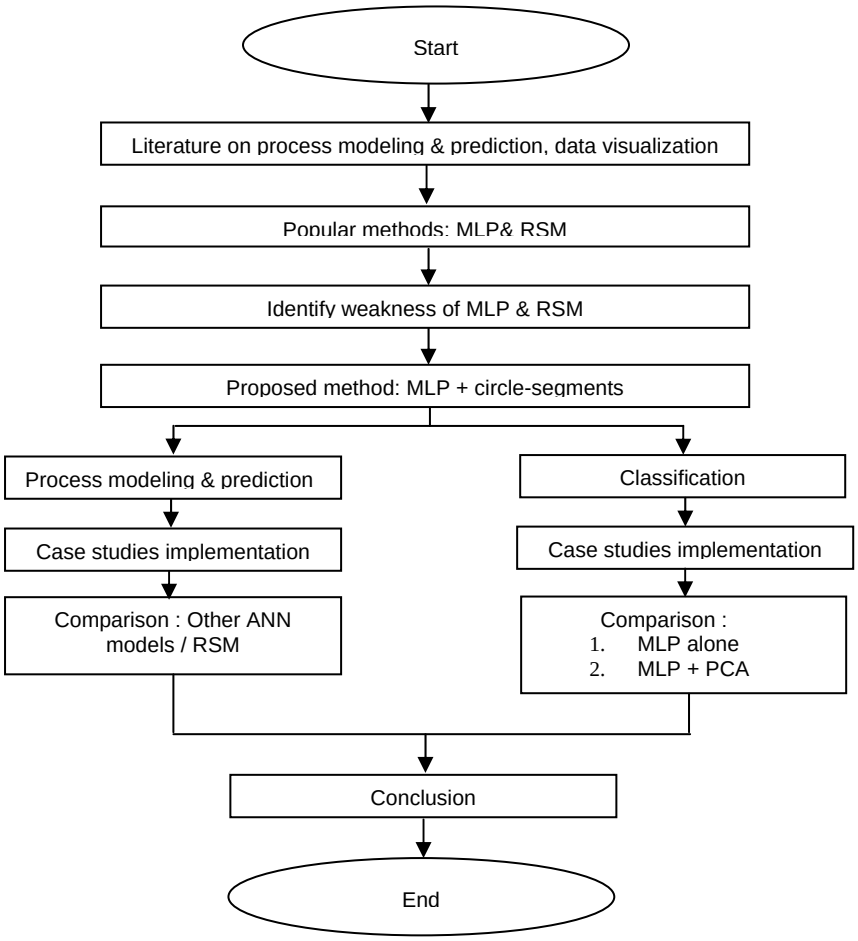

Figure 1.1 – Flow chart of overall work stages involved in the research.

Table 1.1 - Summary of the case studies.

No.	Case Studies	Problem Types	Proposed Method	Comparison Method
1.	Friedman#1 problem (Benchmark problem)	Approximation	MLP + circle-segments (Method1)	1. Bench 2. Simple 3. GRNNFA 4. NBAG
2.	Wire EDM process (Obtained from a journal)	Process Modelling and Prediction	MLP + circle-segments (Method 1)	RSM
3.	Controller for disk drive read system (Simulation data)	Process Modelling and Prediction	MLP + circle-segments (Method 1)	RSM
4.	PID controller (Simulation data)	Process Modelling and Prediction	MLP + circle-segments (Method 1)	RSM
5.	Iris data (Benchmark problem)	Classification	MLP + circle-segments (Method 2)	1. MLP + PCA 2. MLP without any feature selection method
6.	Wine data (Benchmark problem)	Classification	MLP + circle-segments (Method 2)	1. MLP + PCA 2. MLP without any feature selection method
7.	Acute stroke diagnosis (Real medical data)	Classification	MLP + circle-segments (Method 2)	1. MLP + PCA 2. MLP without any feature selection method

1.6 Organization of Thesis

In this chapter, the reasons that drive the research are presented. In chapter 2, a comprehensive review on ANN and data visualization is presented. The strengths and weaknesses of both approaches are then described.

In chapter 3, the techniques used in the development of the hybrid system, namely the MLP network and the circle-segment method, are explained in detail. In addition, the methods used to determine the network parameters and their effects are discussed.

In chapter 4, the applicability of the proposed hybrid system to several modelling and prediction problems is evaluated. The problems studied consist of benchmark problems as well as industrial applications, i.e., Friedman#1, wire electrical discharge machining, and controller tuning. The results are compared with those from the RSM.

In chapter 5, classification problems from benchmark data and real medical data are selected to evaluate the performance of the proposed system. The benchmark problems include the Iris and Wine data, and the medical problem involves acute stroke diagnosis. The results of the proposed system are compared with those from the MLP network coupled with the PCA and the MLP network without any feature selection method.

In chapter 6, conclusions and contributions of this research are presented. Some areas for future work are also suggested in the chapter.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

The collection and analysis of data is fundamentally is about trying to find a potential relationship in a set of data. A major difficulty with the data is that they are subjected to uncertainty. Therefore, the task of data analysis and prediction becomes a challenge. Machine learning and statistical techniques are commonly deployed to solve the task. Machine learning is the application of computer program in performing some classes of task through learning from experience, with its performance improves with experience (Mitchell, 1997). Statistic is the field of study concerned with the collection, analysis, and interpretation of uncertain data in a numerical form (Navidi, 2006). The MLP network and the RSM are examples of machine learning and statistical techniques, respectively, that can be used in data analysis and prediction tasks.

The following sections present a review on process modelling and prediction in various fields. Comparison is made between the ANN and RSM. It is important to involve user in the data exploration and analysis regardless the types of method used (e.g., ANN and RSM). To achieve this goal, data visualization is an important requirement that needs to be induced into the process. A review on data visualization is also presented in this chapter.

2.2 Process Modelling and Prediction

Many industrial processes involve inputs and outputs, or known as factors and responses according to the statistical concept. A lot of efforts have been carried out to establish the relationships between inputs and outputs (factors and responses). Responses are the dependent variables in a process that scientists/engineers are interested to observe the changes. Factors are the independent variables that cause the changes of responses once some changes are made on them. To establish the relationship between factors-responses in process modelling and prediction has become a challenging task.

In most processes, the underlying relationship between factors-responses is unknown or unclear. In such situation, scientists/engineers need to manipulate or control the processes according to their requirements. In this regard, ANNs have been used extensively in process modelling and prediction. One of the main advantages of ANNs is that it is applicable even when the underlying relationship between factors-responses is unclear, as shown by numerous publications as follows.

In Barletta & Grisario (2006), it is stated that the underlying relationship between factors and responses of the laser cleaning process is unclear. Laser cleaning is a process which selectively vaporizes the unwanted or undesired layers clinging onto the workpiece surface by using controlled irradiation emitted by a focused laser beam. Though some applied examples of laser removal of organic residuals, contaminants, and oxide layers have already existed, the detailed relationship between the laser parameters and cleaning performance were still not particularly clear. In the investigation by Barletta & Grisario (2006), the process parameters such as laser power, scan speed, defocus length, and number of passes were selected as the factors that affected the paint stripping factor (PFS). Two neural network models, i.e., MLP

and Radial Basis Function (RBF) were applied. The results were compared with those from the regression model, and it was found that the MLP network achieved a higher degree of accuracy.

Surface mount technology (SMT) has become the dominant development in the assembly of sophisticated electronic devices. The stencil-printing method is used to deposit the solder paste onto the printed circuit boards (PCBs) in the SMT process. However, it is yet to have an appropriate method for selecting and planning the control variables of the stencil-printing process (Yang *et al*., 2005). An integration of the MLP and DOE was used to model the defect problem in the solder paste stencil printing process. A fractional factorial design, which is one of the DOE methods, was used to collect the data. In the case study, eight important factors were identified for predicting the deposited paste volumes. The results showed that the MLP was effective in solving the problem.

The MLP was used to model the deposition yield in the atmospheric plasma spray process (Guessasma *et al*., 2003). The spray process was considered as a complex problem with parameter interdependencies and property correlation. Based on the case study, the predictive result by the MLP showed good agreement with the experimental result.

According to Yarlagadda & Khong (2001), the moulding parameters in injection moulding remained an uncertainty without the assistance of domain experts. Experts often referred to previous mould designs that were similar to the current design and used those successful moulding parameters as guidelines. Therefore, the trial and error method was still the practice to determine the optimum injection moulding process parameters. To overcome this problem, the MLP was used to predict the injection moulding process parameters (Yarlagadda & Khong, 2001).

The MLP was used for modelling a burner heated catalytic converter during cold start in a four stroke, spark ignition engine (Akcayol & Cinar, 2005). A catalytic converter was used to reduce pollutant emissions from internal combustion engines under normal operating condition. It was difficult to model the catalytic converter performance of the engine during cold start, because it involved complicated heat transfer and process, and chemical reactions at both the catalytic converter and exhaust pipe. The work done by Akcayol & Cinar (2005) demonstrated the use of the MLP for the prediction of catalyst temperature, hydrocarbon emission, and carbon monoxide emission. The results showed that the MLP could achieve high accuracy of prediction. Therefore, the MLP was useful in predicting the performance of the catalytic converter.

In machining parts, surface quality is one of the most specified customer requirements, and the main indicator of the surface quality is surface roughness. Based on Özel & Karpat (2005), there are various machining parameters that affect the surface roughness. However, those effects had not been adequately quantified. Özel & Karpat (2005) had developed the MLP for predicting the surface roughness and tool flank wear in finished hard turning process. As opposed to the regression model, it was found that the MLP provided better prediction capabilities for both responses.

On the other hand an integration of the DOE, MLP, and GA was developed for modelling surface roughness in end milling mold surfaces of a plastic part (Oktem *et al*., 2006). Cutting parameters such as cutting speed, cutting feed, axial-radial depth of cut, and machining tolerances were selected as the factors that would affect the surface roughness. The DOE was utilized in carrying out the experimental measurements, while the GA coupled with MLP was used to find optimum cutting parameters leading to minimum surface roughness. Based on the case study, a good

agreement was observed between the predicted values by the MLP and the experimental values.

The use of MLP was demonstrated in the prediction of steel properties (Sterjovski *et al*., 2005). Three MLP models were used for prediction

i. the impact toughness of quenched and tempered pressure vessel steel exposed to multiple postweld heat treatment cycles.
ii. the hardness of the simulated heat affected zone in pipeline and tap fitting steels after in-service welding.
iii. the hot ductility and hot strength of various microalloyed steels over the temperature range for strand or slab straightening in the continuous casting process.

The results of the case study showed that the three MLP models could successfully predict the properties of steels.

The above examples show that ANNs can be applied to process modelling and prediction. The developed ANN models can be very helpful in determining the appropriate operational parameters. Furthermore, the models are able to provide some understanding and indication on the processes, which can be very useful during the set-up process. This can lead to the improvement of productivity as compared with the *trial and error* approach. Note that it is the most common practice to adjust the process parameters by the trial and error approach to obtain adequate response(s). This approach is often very time consuming and laborious. As such, the ANN prediction models can be useful as a support tool in dealing with operational problem with less reliance on skilled, experienced, knowledgeable workers or engineers.

In addition, the developed ANN model can be employed as an assisting tool to examine the effects of potential parameters without performing experimental trial on

expensive sample parts or in expensive investigation areas. This helps save cost of materials, machines, manpower, time that are needed in the experimental trials.

2.3 Comparison of ANN and RSM

As explained in section 2.1, process modelling and prediction has become an important task in many industrial applications such as pharmaceutical, biochemistry, food research, mechanical engineering, and manufacturing technology. The deployment of new techniques as an alternative to the traditional approaches, i.e., the trial and error approach, seem to be emerging as an answer to the increasing attention in solving the task.

The ANN and RSM are among the popular methods used in process modelling and prediction. Both the methods build their own predictive models through data. The difference between these two methods is on the way they build the models. Though an ANN has the ability to learn from data autonomously, it normally acts like a black-box which provides only the input-output relationship with a minimum understanding of the process. In other words, an ANN is a system which has the ability to learn from examples and to formulate a predictive model to tackle the modelling problem in an opaque manner. As opposed to an ANN, the RSM correlates the relationship between inputs-outputs through mathematical and statistical formulations. However, the RSM model is only valid within the range in which it is developed. A number of researchers have made comparison between these two methods, owing to their different ways in building the predictive models as discussed in the following paragraphs. The applications of ANN and RSM and their comparison have covered various industrial areas, from pharmaceutical to manufacturing technology.

In the pharmaceutical industry, an ANN-based intelligent learning system, which comprised the face-centered central composite design (CCD) and the Gaussian mixture model (GMM), had been developed for the prediction of drug release profiles (Lim *et al.,* 2002). The results were compared with those from multiple regression models. Another application of ANN in pharmaceutical application is to access experimental data from a tablet compression (Bourquin *et al.,* 1998). The aim of the study was to quantitatively characterize the influence of each excipient's concentration on the ejection and residual forces during tablet ejection. A comparison between the model and RSM was also made, and the result showed the MLP model achieved better data fitting and predicting abilities.

In the manufacturing industry, an attempt has been made to model a wire mechanical discharge machining (WEDM) process through ANN and RSM (Spedding & Wang*,* 1995). The pulse-width, the time between two pulses, the wire mechanical tension, and the injection set-point were selected as the factors, while the cutting speed, the surface roughness, and surface waviness were the responses. The use of ANN and RSM has covered the biochemical industry too. Based on the work done by Dutta *et al.* (2004), the ANN and RSM were used to build a predictive model of the combined effects of independent variables for extracellular protease production from a newly isolated *Pseudomonas sp*. The independent variables consisted of pH, temperature and inoculum volume.

As reported in most publications, ANNs outperformed the RSM with a higher prediction capability. However, the RSM is better than ANNs as it is able to provide two-dimensional and three-dimensional plots that correlate the input-output relationship. It is difficult to elucidate the nature of the relationship from an ANN model. As a result, the RSM has been used in unison with ANNs to overcome the information visualization problem. However, the RSM is limited by its ability in handling multi-

dimensional problems. The RSM is unable to produce a graphical plot if the dimension of the factors is more than two. Assume that one would like to investigate a process which involves multi-factors and multi-responses. With the help of the constrained-optimization approach, one can visualize the region where multi-responses overlay through a graphical method. Obviously, this approach is difficult to be applied when the numbers of factors exceed two. One can hardly visualize the factors-responses (inputs-outputs) relationship in the same graph if the number of factors exceeds two. Thus, this makes the use of RSM as the visualization tool of ANNs unattractive.

2.4 Data Visualization Methods

Data visualization is a method of mapping or projecting numerical data into graphical forms so that insight and knowledge can be gained. Since humans are good at perceiving information pictorially, it is important that data visualization methods are employed to elucidate the relationship between factors-responses. A major advantage of visualization techniques over other (semi)automatic data exploration and analysis techniques (from statistics, machine learning, neural network, etc) is that visualization allows a direct interaction with the user and provides an immediate feedback, as well as enables user steering which is difficult to achieve in most non-visual approaches (Keim, 2000).

It is always easier for humans to understand or correlate the relationship if the proposed model is more visual. Data visualization methods show the overall structure of a process and how the factors-responses relate to each other. They help gather all the relevant information in one place and make the complex relationship easier to be understood. According to Keim (2002), there are two main advantages of data visualization techniques as compared with other machine learning techniques. First,

they can easily deal with highly non-homogeneous and noisy data. Second, they are intuitive and require no understanding of complex mathematical or statistical algorithms or parameters. Therefore, it is easier for users to access or interact with the data through the use of data visualization techniques.

Data visualization techniques enable the data, especially multidimensional data to have an overview. An overview of large information spaces reduces the necessary search, allows the detection of overall patterns, and assists users in making the next move. Demian & Fruchter (2006) explained that the need for an overview is implied by the information foraging theory. According to this theory, users of information will modify their information seeking strategies or the structure of the environment to maximize their rate of gaining valuable information.

Having the right information is crucial for making the right decision (Keim *et al.,* 1995). Without an appropriate tool, the process of exploring and analyzing data may become tedious and difficult, which leads to inaccurate or at least suboptimal decisions. However, the process of analyzing data cannot fully depend on the computer since it still needs human intelligence in making judgment and decisions. Therefore, humans will continue to play an important role in exploring and analyzing the data. Humans need to be adequately supported by the computer in dealing with large amounts of data instead of depends on the computer blindly. Therefore, the ability to visualize data is an important way of supporting humans in exploring and analyzing the data. An appropriate data visualization tool is very useful in providing a quick overview on the large amounts of multidimensional data, providing the possibility of focusing on significant effects or patterns.

One of the applications of machine learning is to derive general knowledge from specific data sets by searching through possible hypotheses exemplified in the data.

The machine learning techniques vary from simple testing of sample features for statistical significance to sophisticated probabilistic modelling techniques. For most of the machine learning techniques, all of the inputs are provided to the system without carrying out an appropriate analysis. This happens a lot in the applications of ANN. In ANN applications, the use of more inputs than necessary would make the network over-specific to the training data, which would reduce the generalization capabilities of an ANN (Kavzoglu & Mather, 2000). As a result, the performance of the ANN in testing data (data which has not been seen by the network before) deteriorates.

In order to solve this problem, features selection needs to be carried out, and data visualization techniques are suitable to accomplish this task. According to Lerner *et al*. (1994), features selection for classification is defined as a search among all possible transformations, for the best subspace that preserves class separation as much as possible in the lowest possible dimension space. Features selection can improve the accuracy of the machine learning techniques by eliminating the insignificant inputs. The integration of machine learning and data visualization techniques is needed to improve the data analysis process through the combination of each other's strength. Several studies have shown the usefulness of the integration of machine learning and data visualization techniques (Johansson *et al*., 2004; McCarthy *et al*., 2004; Ruthkowska, 2005). In their studies, the data visualization technique had been used as a feature selection tool. With the help of data visualization techniques, one can understand and identify the significant features, and thus maximize valuable elucidation to be derived from the analysis. This helps reduce redundancy in data, otherwise, the accuracy of the predictive model under development can be affected.

2.4.1 Data Visualization Category

There are many types of data visualization techniques, which basically can be categorized into six categories; geometric projection techniques, icon-based techniques, hierarchical-based techniques, graph-based techniques, pixel-oriented techniques and combination thereof, as shown in Figure 2.1 (Keim, 2000). Nevertheless, it is clear that no one general set of visualization techniques is suitable to address all problems (Fayyad *et al*., 2002). Different techniques will be selected according to the task and data of the problems.

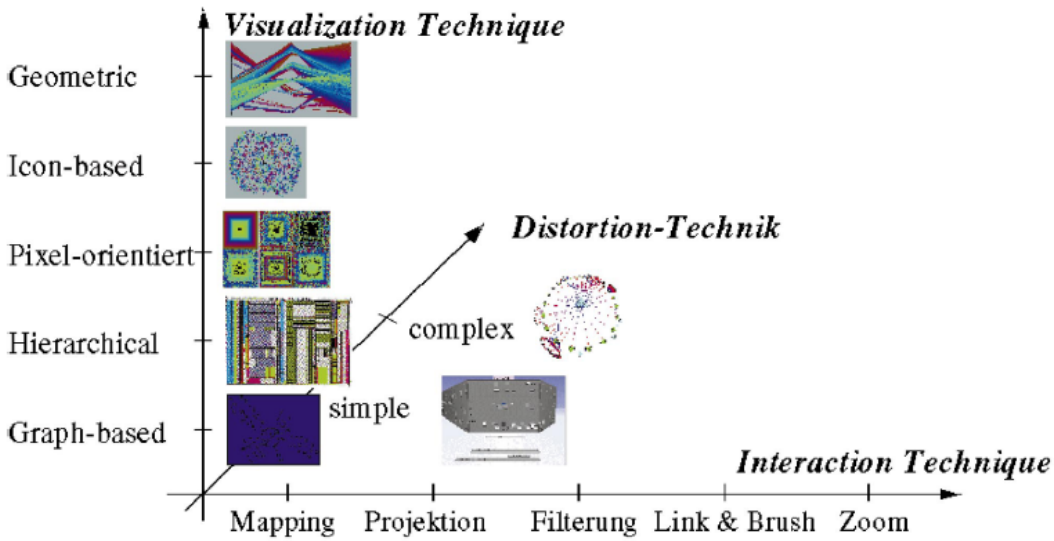

Figure 2.1 - Categories of data visualization.

2.4.1.1 Geometry Category

The parallel coordinates is an example of geometric projection techniques. The parallel coordinates is a two-dimensional technique to visualize multidimensional data set (Inselberg & Dimsdale, 1990). Assume a data set consists of *k* dimensions

(attributes). In the parallel-coordinates method, each of the dimensions is represented by a vertical line. The maximum and minimum of these dimensions are scaled to the upper and lower levels on these vertical lines. An example of the parallel coordinates is shown in Figure 2.2. A *k*-dimensional data sample is drawn by plotting the value of each dimension along the appropriate vertical lines, and a polygonal line then connects all the values. For a data set consists of many items, this technique produces a compact two-dimensional visualization of the whole multidimensional data set.

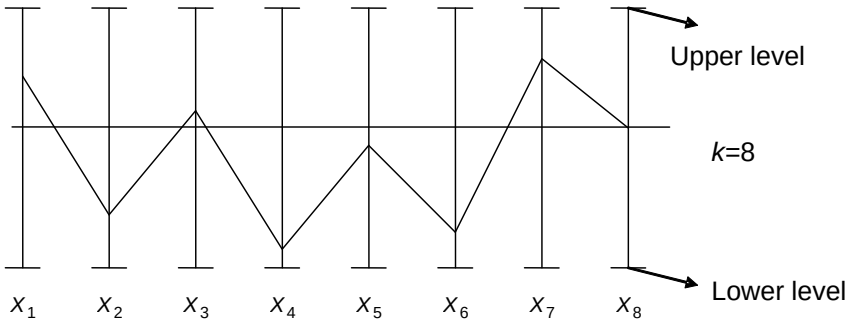

Figure 2.2 – Parallel coordinates shows an eight-dimensional data element by a polygonal line.

The parallel coordinates has been used directly or in modified manner to improve its performance in discovering new information in the data set. The applications of this technique have covered several areas. It was applied to the 2-region Sangren-Sheppard model of capillary exchange (King & Haris, 1999). The parallel coordinates-based model analysis was examined for all combinations of several values of flow, extravascular volume and capillary membrane permeability-surface area product, representing a total 900 simulations. The insight gained through this model has the potential to aid in the development of more rigorous parameter identification procedure, as well as aid in the design of experiments. In addition, new techniques are introduced to manipulate parallel coordinates in order to improve the performance of the parallel coordinates, i.e., dimension zooming and hierarchical clustering (Siirtola, 2000). The proposed techniques were applied to the cars data set, which consisted of 9 dimensions and 406 polylines, resulting in 3654 data items. The

car data set consisted data on the cars road tested by the Consumer Reports Magazine between 1971 and 1983. The two new techniques were used to enhance the performance of parallel coordinates in discovering new information and correlating the dimensions in the data set.

2.4.1.2 Icon-Based Category

Glyph or icon is defined as a shape or image by mapping data components to graphical attributes (Fayyad *et al*., 2002). The Chernoff faces (Figure 2.3) and star glyphs (Figure 2.4) are examples of icon-based visualization. In the Chernoff faces, the data dimensions can be mapped to facial features such as angle of the eyes and the width of the nose. In the star plot, the dimensions are represented as equal angular spokes radiating from the center of a circle. The outer end of each spoke (axis) represents the maximum value of the dimension, while the center of the circle represents the minimum value of the dimension. In a typical display, there is a star glyph for every *n*-dimensional data point.

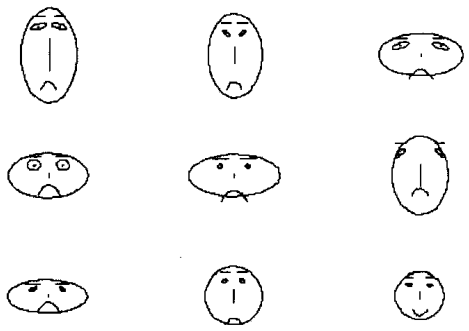

Figure 2.3 – Chernoff faces.