CRC Construction Innovation

BUILDING OUR FUTURE

Report
Data Mining of Life Prediction Data Bases

Research Project No: 2005-003-B-11

The research described in this report was carried out by:

Project Leader	Penny Corrigan
Researchers	Angela Bradbury Ivan Cole Robin Drogemuller Stephen Egan Wayne Ganther Tim Muster David Paterson Gerry Trinidad Natalie Sherman Andrew Martin Wan Yee Chan Richi Nayak Esther Ge
Project Affiliates	Peter Hope Michael Ball Frank Turvey Lee Wade Wayne Muller Lex Vanderstaay

Research Program: B
Sustainable Built Assets

Project: 2005-2003-B
Learning System for Life Prediction of Infrastructure

Date: November 2007

Leaders in Construction and Property Research

Distribution List

Cooperative Research Centre for Construction Innovation
Authors

Disclaimer

The Client makes use of this Report or any information provided by the Cooperative Research Centre for Construction Innovation in relation to the Consultancy Services at its own risk. Construction Innovation will not be responsible for the results of any actions taken by the Client or third parties on the basis of the information in this Report or other information provided by Construction Innovation nor for any errors or omissions that may be contained in this Report. Construction Innovation expressly disclaims any liability or responsibility to any person in respect of any thing done or omitted to be done by any person in reliance on this Report or any information provided.

Please direct all enquiries to:

Chief Executive Officer
Cooperative Research Centre for Construction Innovation
9th Floor, L Block, QUT, 2 George St
Brisbane Qld 4000
AUSTRALIA
T: 61 7 3138 9291
F: 61 7 3138 9151
E: enquiries@construction-innovation.info
W: www.construction-innovation.info

Table of Contents

1 INTRODUCTION

The ability to accurately predict the lifetime of building components is crucial to optimising building design, material selection and scheduling of required maintenance (Cole et al. 2005). The material should be selected to match the severity of the environment. For example, in a severe marine location, very durable materials need to be selected, while in a benign environment lower quality products can be used. Along with materials selection, the timing of maintenance and building design would be tailored to the severity of the environment. Through the building service life prediction methods, substantial cost savings can be made. For example, it has been estimated that nearly $5 million was spent by the Queensland Department of Public Works (QDPW) in 03/04 in replacing corroded metallic components in Queensland schools (Cole et al. 2005). Organisations such as this urgently require a lifetime prediction tool for atmospheric corrosion which can offer potential savings for this cost.

A wide range of techniques has been undertaken at CSIRO to predict service life. Approaches include the development of fundamental degradation and microclimate models [4], and the development of a database on expert performance (Delphi Study, Cole *et al*, 2004). These methods should be viewed as complementary rather than as discrete alternatives. They form different data sources of service life information. The problem is how they could be combined to determine the most appropriate answer for any given situation. Data mining is considered to be an ideal method that links together the different data sources and provides intelligent decisions.

Data mining (DM) has been driven by the need to solve practical problems since its inception (Melli et al., 2006). In order to achieve a greater usability of the data mining models, there are three main phases in the lifecycle of a data mining project: (1) training of the model, (2) evaluation (or testing) of the model and (3) using the final trained model in practice. The third phase is usually carried out by the business managers or a typical user of the system. A number of Knowledge Discovery and Data Mining (KDDM) process models have been established to organise the lifecycle of a data mining project within a common framework. However, the existing KDDM process models end up with the deployment phase and do not consider the use of the trained model in practice. This has led to a gap of knowledge that may limit widespread use of the trained model.

DM has been successfully applied in many areas such as business, marketing, medical and financial fields (Kantardzic and Zurada, 2005). Civil engineering is one of the areas where a variety of successful real-world data mining applications are reported in building construction (Melham and Cheng, 2003; Leu et al., 2001; Furuta et al., 1995; Morcous et al., 2002a; Morcous et al., 2002b, Mita and Hagiwara, 2003; KamrunNahar and Urquidi-MacDonald, 2005; Brence and Brown, 2002; Skomorokhov, 2000; Kessler et al., 1994; Melhem et al., 2003). One such application is metallic corrosion prediction in buildings. The corrosion prediction applications can be classified into two main categories: 1) building the predictive models using various traditional data mining techniques; and 2) improving the prediction accuracy using new hybrid methods.

All of these predictive models in the above two categories assume that the inputs that users will provide in using the model are the same as the input features used for training the models. However, if users have information of limited inputs only, the predicted results will not be as good as they were during the training and evaluation phases of the data mining system. In other words, the performance of the predictive model degrades due to the absence of many input values. For example, a predictive data mining model is built to predict the "Service Life" of the building components based on the input features such as "Location", "Component", "Material", "Salt Deposition", and "Mass Loss" (shown as Figure 1). Suppose builders (typical users of the predictive model or tool) want to know the service life of a "Gutter" with "Galvanized Steel" at a location (shown as Figure 2). However, the user does not know the "Salt Deposition" and "Mass Loss" in that location. The user query will include two missing values. In such a case, the predicted service life by the predictive data mining tool may not be as accurate as the service life tested in the evaluation phase of the predictive model, especially when the missing features play key roles in building the model. On the other hand, if the "Salt Deposition" and "Mass Loss" features are excluded from the model building, the performance of the model may not be acceptable. Hence, a major problem that still needs to be solved is how to select appropriate features to build the model for a real situation when users have information on limited inputs only.

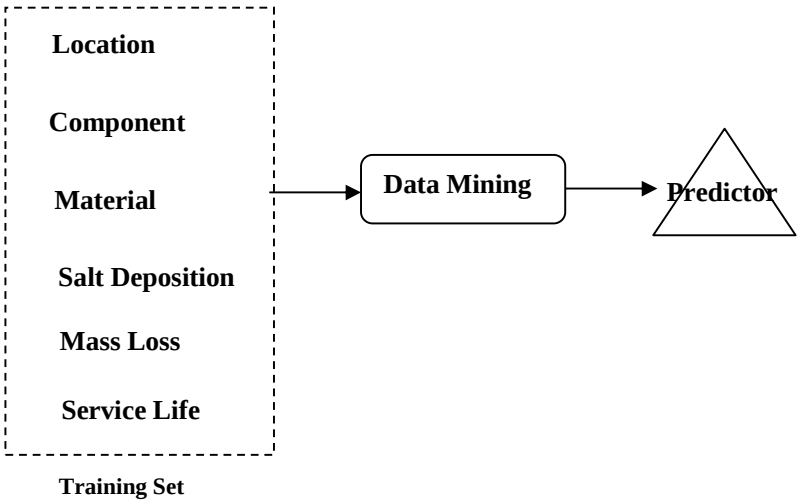

Figure 1. Training of the Model

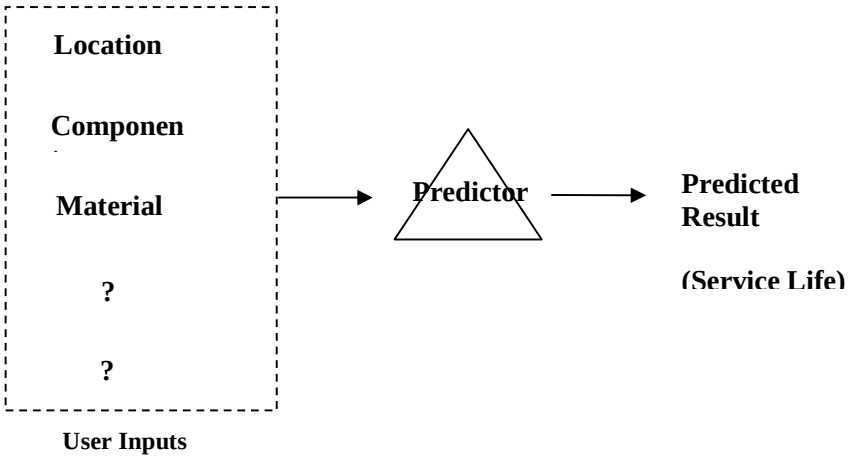

Figure 2. Using the Trained Model

This research proposes a learning system framework, namely the Query Based Learning System (QBLS), for improving the performance of predictive models in practice where not all inputs are available for querying to the system.

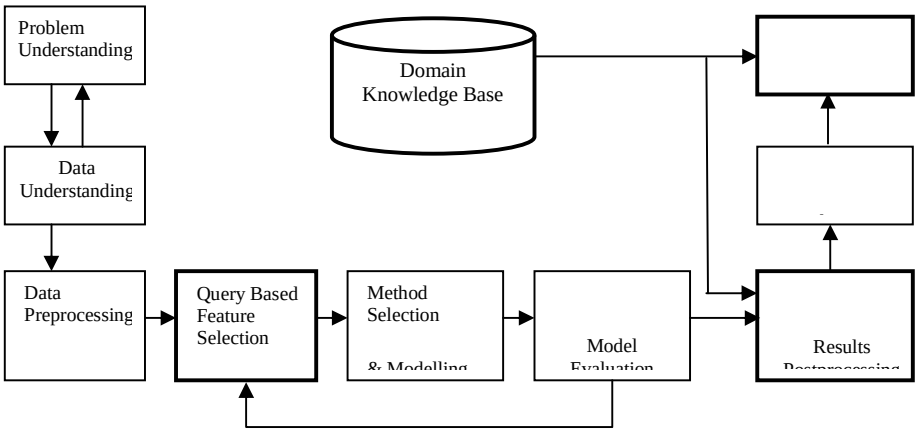

Figure 3. Query Based Learning System

The QBLS model consists of nine phases (as shown in Figure 3), which are structured as sequences of predefined steps. The arrows indicate the most important and frequent dependencies between phases. A domain knowledge base is involved in the results post-processing and the use of model phase. More specifically, the domain knowledge is used for pre-processing the incomplete queries and post-processing the inconsistent results. Based on this model, a practical system is developed for predicting the lifetime of metallic components. The system is evaluated on the data provided by CSIRO.

1.1 Background Information on Data Mining

Data mining, also referred to as knowledge discovery, is a powerful new technology with great potential to help companies to focus on the most important information in their data warehouses or database. It extracts hidden valued information from large databases (Fayyad et al, 1995a; Chapple, 2006). Through the use of automatic or semiautomatic algorithms, data mining extracts patterns from the data and transfers the data to knowledge. Data mining techniques can be applied to many applications, answering various types of business questions such as cross-selling, fraud detection and banking (Kantardzic and Zurada, 2005). A poll about successful data mining applications in 2005 was presented on KDnuggets website (2005), which shows that the most common applications are still the traditional areas of Banking, Direct Marketing, and Fraud detection.

1.1.1 Basic Data Mining Tasks

Based on the nature of data mining problems, the data mining tasks can be grouped into the following main categories: classification, regression, clustering and association rules.

Classification

Classification is one of the most popular data mining tasks. Classification assigns tuples in the dataset into predefined classes based on a target attribute. Each tuple contains a set of attributes, one of which is the target attribute and others can be chosen as input attributes. The purpose is to find a model that describes the target attribute as a function of input attributes. Classification can be considered as supervised learning since it requires a target to learn.

Prediction can be viewed as a type of classification when the target is a categorical attribute; namely, prediction can be thought of as classifying an attribute value into one of a set of possible classes.

Typical classification algorithms include K Nearest Neighbors (Aha et al, 1991), decision trees (Quinlan, 1986), neural network (Resampling Stats??), Naïve Bayes (Fayyad et al., 1995b) and support vector machine (Vapnik, 1995).

Regression

The regression task is similar to classification. The main difference is that the target attribute is a continuous value. Just as prediction for class values can be viewed as a classification problem, numeric prediction can be regarded as a regression problem. Therefore, the proposed research problem belongs to this category.

Although all classification algorithms can automatically deal with continuous values (they usually divide them into ranges, e.g. decision trees), most of them can not be used to solve a regression problem directly (e.g. decision trees and Naïve Bayes) unless numeric target is discretised to nominal type. However, the discretisation level chosen dramatically affects the learning of the problem and, not incidentally, the utility of the results. Therefore, the best solution to a regression problem is regression techniques. Linear regression and logistic regression are the most popular regression methods. Other regression techniques include regression trees(Breiman et al., 1984), model trees (Quinlan, 1993), neural networks and support vector machine (Vapnik, 1984), in which a neural network and support vector machine can also be applied to the classification problem.

Clustering & Association Rules

Clustering and association rules are another two popular data mining tasks. Clustering partitions or segments the data into groups (clusters). The most similar data are grouped into the same group. It is similar to classification except the groups are not predefined, but rather based on a set of attributes. From this point of view, clustering is an unsupervised learning.

Association rules, also called market basket analysis, refer to the data mining task of finding the relationships between data items. The form of an association rule is $X \Rightarrow Y$, where X and Y are sets of items called itemsets. Support and confidence are used to measure an association rule, in which support is the percentage of transactions in the database that contain $X \cup Y$ and confidence is the ratio of the number of transactions that contain $X \cup Y$ to the number of transactions that contain X (Dunham, 2003). The common usage of association rules is to identify common sets of items and rules for the purpose of cross-selling (Chapple, 2006).

1.1.2 Knowledge Discovery and Data Mining Process Model

A Knowledge Discovery and Data Mining (KDDM) process model consists of a set of processing steps to be followed by practitioners when executing KDDM projects. The concept of a KDDM process model was originally discussed during the first workshop on KDD in 1989 (Piatetsky-Shapiro, 1991). The main reason for defining and implementing KDDM process models is to ensure that the end product will be useful to the user (Fayyad et al, 1996a). The basic structure of the model was proposed by Fayyad et al. (1996b). Since then, several different KDDM models have been developed in both academia and industry. The human-centric and data-centric models are two major types of process models. The human-centric model emphasised the interactive involvement of a data analyst during the process, and the data-centric model emphasised the iterative and interactive nature of the data analysis tasks (Fayyad et al., 1996b). . Kurgan et al. (2006) conducted a survey of knowledge discovery and data mining process models, presenting a historical overview and a comprehensive comparison of several leading process models.

The CRISP-DM (CRoss-Industry Standard Process for Data Mining) (2003) process model is currently the most popular and broadly adopted data-centric model. It was first proposed in early 1996 by a consortium of three companies: SPSS (then ISL), NCR and DaimlerChrysler (then Daimler-Benz). It was later sponsored by the European Commission research fund. This model is very industry-oriented and enjoys strong industrial support. In fact it has already been assessed as meeting industrial needs (Kurgan et al., 2006).

The CRISP-DM model consists of six phases, as shown in Figure 4. The centre of the CRISP-DM model is the data. The possible relationships between all data mining phases most importantly depend on the data. The arrows indicate the most important and frequent dependencies between phases. The outer circle in the figure symbolises the cyclic nature of data mining itself. A data mining process continues after a solution has been deployed. The lessons learned during the process can trigger new, often more focused business questions.

Below follows a brief outline of the phases:

Business Understanding

This initial phase focuses on understanding business objectives and requirements, which are converted into a data mining problem definition.

Data Understanding

The data understanding phase includes data collection, identification of data quality problems, data exploration and detection of interesting subsets.

Data Preparation

The data preparation phase covers all activities about preparation of the final dataset which will be fed into the modeling tool(s). The tasks include table, record, and attribute selection as well as data transformation and cleaning.

Modeling

The modeling phase selects and applies various data mining techniques to the prepared data and generates the knowledge (patterns) from data or constructs the model from data.

Evaluation

The evaluation phase evaluates the generated knowledge/model from the business perspective, to be certain it properly achieves the business objectives.

Deployment

The deployment phase includes presentation of the discovered knowledge, generation of a report or implementation of deployment in order to actually make use of the created models.

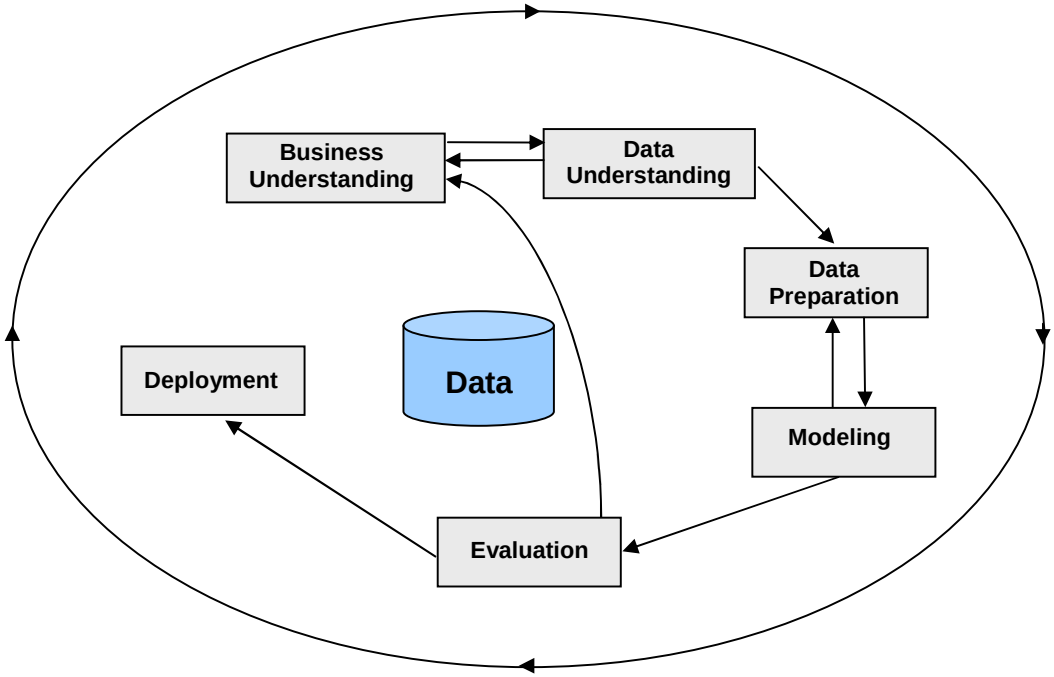

Figure 4. Phases of the CRISP-DM Process Model

In general, data-centric models are structured as sequences of steps that focus on performing manipulation and analysis of data and information surrounding the data. In such models, the user's role is to ensure that specific objectives for each step are met (Kurgan et al., 2006). Therefore, one major limitation of such models is their lack of user interaction. As the main purpose of KDDM process models is to ensure that the end product will be useful to the user, the success of a process model depends upon providing results to suit user needs. This success could be achieved when the user interacts with the process model by constraining the process to suit his/her needs. Another limitation is such models do not consider the use of model phase, which is usually carried out by the customer after the model is deployed. The problems or user needs sometimes arise during this phase as described earlier. Such problems/needs will trigger new, often more constrained data mining processes.

2. RELATED DATA MINING APPLICATIONS

A number of successful corrosion prediction applications in civil engineering have been reported.

Furuta et al. (1995) developed a practical decision support system for structural damage assessment due to corrosion using the Neural Network. This system aimed to aid inexperienced inspectors to judge whether a certain bridge should be repaired or not. It proved the learning ability of the Neural Network in damage assessment.

Morcous et al. (2002a) proposed a case-based reasoning system for modeling infrastructure deterioration (CBRMID). It was a CBR system developed to provide government agencies with practical, accurate, and versatile deterioration models. The architecture of CBRMID was described in terms of case representation, case retrieval, case adaptation, and case accumulation. Later Morcous et al. (2002b) presented an application example generated using CBRMID for modeling the deterioration of concrete bridge decks.

Melhem and Cheng (2003) first used KNN and the decision tree for estimating the remaining service life of bridge decks. Their work showed the prediction accuracy generated by KNN (50%) was higher than that produced by C4.5 (41.8%). However, both of these values were considered low from a machine learning standpoint. They attributed this to the fact that the deterioration model used to compute the remaining service life turned out to be inadequate. Later Melhem et al. (2003) investigated the use of wrapper methods to improve the prediction accuracy of the decision tree algorithm for the application of bridge decks. Bagging, boosting and automatic feature selection were chosen to compare the results. Their experiments showed all three methods could provide improvement to the decision tree. However, the improvement obtained by the feature selection method can be misleading because the attributes selected were not the ones most important to the problem domain. Therefore, what may be an improvement from the machine learning or data mining viewpoint, can turn out to be a mistake from an engineering perspective. They concluded that the general purpose feature selection was not recommended in this case.

Skomorokhov (2000) presented a rule extraction algorithm for a real life problem, which is to find automatic rules to describe the corrosion rate of steel in sodium as a function of alloy additions. The input data were experimental data of corrosion rate measured for different steel samples. The output is a set of IF-THEN rules, which describe the dependence of corrosion rate on alloy additions.

Brence and Brown (2002) described the use of data mining (multiple linear regression, regression trees, polynomial networks and ordinal logistic regression) to predict corrosion damage from non-destructive test (NDT) data with aircraft. Their results showed that while a variety of modeling techniques can predict corrosion with reasonable accuracy, regression trees are particularly effective in uncovering the complexity of the corrosion-NDT relationship.

Others like Kessler et al. (1994) improved prediction of the corrosion behaviour of car body steel using a Kohonen self organising map. Leu et al. (2001) presented a data mining approach to the prediction of tunnel support stability using artificial neural networks. Mita and Hagiwara (2003) proposed a method using the support vector machine to detect local damage in a building structure with a limited number of sensors. KamrunNahar and Urquidi-Macdonald (2005) used Neural Network to predict the corrosion behaviour and in turn, the life of metals and alloys over extended periods of time in specific environments.

Although the above applications utilise various data mining techniques to predict the corrosion or service life of building components, they can be classified into two main groups: 1) Building the models using various traditional data mining techniques (Melham and Cheng, 2003; Leu et al., 2001; Furuta et al., 1995; Morcous et al., 2002a; Morcous et al., 2002b, Mita and Hagiwara, 2003; KamrunNahar and Urquidi-MacDonald, 2005;

Brence and Brown, 2002; Skomorokhov, 2000) and 2) Improving the prediction accuracy using new hybrid methods (Kessler et al., 1994; Melhem et al., 2003). None of them involves solving the problem of reduced performance in a real situation when users only have knowledge of limited inputs.

3. DATA ANALYSIS AND REPRESENTATION

This section describes the datasets to be used in this project, data pre-processing and existing problems to be solved.

3.1 Data Acquisition

The objective is to predict the service life of metallic components in Queensland school buildings. The datasets include two different sources of service life information: the Delphi Survey and Holistic Corrosion Model, in which the Holistic Model includes three datasets named Holistic-I, Holistic-II and Holistic-III for different components and materials respectively. The Delphi Survey, conducted by the CSIRO, includes the estimation of service life for a range of metallic components by experts in the field such as builders, architects, academics and scientists. The Holistic Model is based on a theoretical understanding of the basic corrosion processes. It provides the required knowledge for computing the lifetime of metallic components through grounded theories and principles. Details of these datasets are presented in Table 1.

Table 1. Details of Datasets

Data Set	***Number of cases***	***Number of attributes***	***Building Component***	***Building Material***	***Target attribute***
Delphi Survey	*683*	*10*	*Roofs, Gutters, Others*	*Galvanized Steel, Zincalume, Colorbond, Others*	*Mean*
Holistic-I	*9640*	*11*	*Gutters*	*Galvanized Steel and Zincalume*	*MLannual*
Holistic-II	*4780*	*22*	*Gutters*	*Colorbond*	*Life of gutter at 600um*
Holistic-III	*1297*	*18*	*Roofs*	*Galvanized Steel and*	*Zincalume Life* *Galvanized Life*

				Zincalume	

3.1.1 Delphi Survey

The Delphi Survey dataset contains the predicted life information for over 30 components and 29 materials, for marine, industrial and benign environments of both service (with and without maintenance) and aesthetic life. They are knowledge of domain experts. The output of this dataset is an estimated service life of metallic components. As the Delphi dataset is the result of surveys, the final dataset was examined in three ways to determine its accuracy and reliability. These were analysis for internal consistency of the data, analysis for consistency with expected trends based on knowledge of materials performance and correlation with existing databases on component performance. In all of these comparisons, the Delphi dataset showed good agreement (Cole et al., 2005). Table 2 contains the details of the Delphi Survey dataset.

Table 2. Details of Delphi Survey

Attribute ID	***Attribute Name***	***Type***	***Description***
1	*Building type*	*Nominal*	*e.g. Commercial, Residential*
2	*Component*	*Nominal*	*e.g. Gutters, Roof, Door Handles*
3	*Measure*	*Nominal*	*e.g. Service Life, Aesthetic Life*
4	*Environment*	*Nominal*	*e.g. Benign, Industrial, Marine*
5	*Material*	*Nominal*	*e.g. Aluminium, Galvanised Steel, Zincalume*
6	*Maintenance*	*Boolean*	*Yes / No*
7	*Mode (years)*	*Nominal*	*The range of Service Life, Aesthetic Life or Time to First Maintenance (e.g. <5 means less than 5 years,* *5-10 means from 5 to 10 years)*
8	*SD (years)*	*Numeric*	*standard deviation for the mean*
9	*Mean (years)*	*Numeric*	*The average years of Service Life, Aesthetic Life or Time to First Maintenance*
10	*Criteria*	*Nominal*	*How good the agreement was in the responses from the survey* *Rated 1,2,3,4*

3.1.2 Holistic-I

The Holistic-I dataset contains theoretical information of corrosion for gutters with Galvanized Steel and Zincalume in Queensland schools. The overall model is a reflection of the influence of climatic conditions and material/environment interactions on corrosion. Table 3 contains the details of the Holistic-I dataset. The output of this dataset is the annual mass loss of Zincalume or Galvanized Steel. Once the mass loss of material is determined, its service life is measured with formula 1 (Cole et al., 2005).

$$\textit{Service life} = \min\left(\exp\left(\frac{\log \frac{effective\_coating\_mass}{MLannual \times 1.5}}{n}\right), 100\right) \quad \ldots \text{Eqn(1)}$$

Where MLannual is the annual mass loss (last attribute of Holistic-I dataset), effective_coating_mass = 56.25 for Zincalume and 103.13 for Galvanized, n = 0.60 for Zincalume and 0.62 for Galvanized.

Table 3. Details of Holistic-I

Attribute ID	***Attribute Name***	***Type***	***Description***
1	*LocID*	*Numeric*	*Location ID for each school*
2	*XLong*	*Numeric*	*Longitude and Latitude of school*
3	*YLat*	*Numeric*	
4	*Location*	*Nominal*	*School name*
5	*State*	*Nominal*	*QLD*
6	*SALannual*	*Numeric*	*Annual salt accumulation*
7	*Building Type*	*Nominal*	*Gutters*
8	*Material*	*Nominal*	*Zincalume or Galvanized*
9	*Gutter Position*	*Nominal*	*Bottom-interior, outside or sides-interior*
10	*Gutter Maintenance*	*Boolean*	*Cleaned or not cleaned*
11	*MLannual*	*Numeric*	*Annual Mass Loss of Zincalume/Galvanized*

3.1.3 Holistic-II

The Holistic-II dataset is for gutters with Colorbond material in Queensland schools. This model is also generated with theoretical information. But the rules for the degradation of Colorbond are devised separately. The output of this dataset is the service life of gutters with Colorbond material. Table 4 presents the details of the Holistic-II dataset.

Table 4. Details of Holistic-II

Attribute ID	***Attribute Name***	***Type***	***Description***
1	*LocID*	*Numeric*	*Location ID for each school*
2	*XLong*	*Numeric*	*Longitude and Latitude of school*
3	*YLat*	*Numeric*	
4	*SALannual*	*Numeric*	*Annual salt accumulation*
5	*Building Type*	*Nominal*	*Gutters*
6	*Position*	*Nominal*	*Facade of buildings*
7	*Exposure*	*Nominal*	*Open or sheltered*
8	*Material*	*Nominal*	*Colorbond*
9	*PositionVsExposure*	*Nominal*	*Openly exposed to rain and sky or sheltered from rain and sky*
10	*Building Face*	*Nominal*	*Front face*
11	*BuildingFacePos*	*Nominal*	*Edges*
12	*Gutter Type*	*Nominal*	*One-sided topcoat or two-sided topcoat*
13	*rain_annual_mm*	*Numeric*	*Annual rainfall*
14	*cum_MZa_2ndYear*	*Numeric*	*Cumulative Zincalume mass loss of 2nd year*
15	*cum_dSTEEL_2ndYear*	*Numeric*	*Cumulative Steel corrosion of 2nd year*
16	*remCr*	*Numeric*	*The amount of chromate remaining in the 25um area surrounding the defect*
17	*normCr*	*Numeric*	
18	*accelerated_corrosion_rate*	*Numeric*	*An increased corrosion rate of Zincalume*
19	*Time to White Rust of Zincalume*	*Numeric*	*Time to occur Zincalume Mass Loss*
20	*Time to penetration of Zincalume*	*Numeric*	*Time to penetrate Zincalume coating*
21	*Time to onset of Red Rust*	*Numeric*	*Time to occur Steel Mass Loss*
22	*Life of gutter at 600um*	*Numeric*	*Service life of gutter*

3.1.4 Holistic-III

The Holistic-III dataset contains life information of roof components for schools in Queensland. They are the results of analysing over 10000 records with regard to significant maintenance. The output of this dataset is service life of roofs with Zincalume and Galvanized Steel materials. Table 5 presents the details of the Holistic-III dataset.

Table 5. Details of Holistic-III

Attribute ID	***Attribute Name***	***Type***	***Description***
1	*Centre Code*	*Numeric*	*Identification for each school*
2	*Centre Name*	*Nominal*	*School name*
3	*Longitude*	*Numeric*	*Geographic location of school*
4	*Latitude*	*Numeric*	
5	*Salt Deposition*	*Numeric*	*A parameter pertinent to corrosion*
6	*Zinc Mass Loss*	*Numeric*	*Damage to Zinc, Steel and Zincalume*
7	*Steel Mass Loss*	*Numeric*	
8	*Zincalume Mass Loss*	*Numeric*	
9	*Marine*	*Boolean*	*True / False*
10	*Nzinc*	*Numeric*	*A constant that depends on Zinc Mass Loss*
11	*Nsteel*	*Numeric*	*A constant that depends on Steel Mass Loss*
12	*L*	*Numeric*	*Calculated based upon Zinc Mass Loss*
13	*M*	*Numeric*	*Calculated based upon Steel Mass Loss*
14	*N*	*Numeric*	*Calculated based upon Zincalume Mass Loss*
15	*Zinc Life*	*Numeric*	*Calculated based upon Nzinc and L*
16	*Steel Life*	*Numeric*	*Calculated based upon Nsteel and M*
17	*Zincalume Life*	*Numeric*	*Calculated based upon N*
18	*Galvanized Life*	*Numeric*	*Calculated based upon Zinc Life and Steel Life*

In general, the Delphi Survey is expert opinions while Holistic-I, -II and -III are theoretical. They form four important sources of information for predicting the lifetime of metallic components. They are independent but complementary to each other. The Delphi Survey can be used for analysing correlation with the other three datasets on component performance and consistency with expected trends based on knowledge of materials performance, while Holistic-I, -II and -III provide theoretical proof for prediction. Holistic-I, -II and -III relate to different component types with different materials while Delphi contains all component types with all materials. More specifically, Holistic-I is for gutters with Galvanized Steel and Zincalume, Holistic-II is for gutters with Colorbond, Holistic-III is for roofs with Galvanized Steel and Zincalume and Delphi is for a range of components including roofs and gutters with different materials including Galvanized Steel, Zincalume and Colorbond. There is no overlap of predicted outcomes from Holistic-I, -II and -III while the predicted outcome from them can be compared with the outcomes from Delphi.

3.2 Data Preprocessing

Data quality is a key aspect in performing data mining on real-world data. Raw data generally include many noisy, inconsistent and missing values and redundant information. This section describes how data is pre-processed in terms of data cleaning and data reduction.

3.2.1 Data Cleaning

Data cleaning consists of dealing with missing data and inconsistent data. In our datasets, the percentage of missing values is very low. For the Delphi Survey, only the attribute ‘mode’ has 8% missing values while all other attributes have no missing values. For Holistic-I, only the attribute ‘Gutter Maintenance’ has 51% missing values. For Holistic-II and -III, all attributes have no missing values. Due to the low percentage of missing values, we do not apply cleaning on the missing values. Inconsistent data were also dealt with during the data cleaning phase. An example for inconsistent data is the use of lowercases and capitals such as ‘Steel’ and ‘steel’. More examples are different spellings but the same meaning like ‘Galvanised’ and ‘Galvanized’ or different words but the same meaning like ‘Steel in Hardwood’ and ‘Steel-Hardwood’. More spaces included in values like ‘Residential ’ and ‘Residential ’ is another reason to cause inconsistency. The data mining tool will treat those kinds of values as different values and hence will influence the predicted results. All such kind of inconsistency is recovered during the data cleaning phase. For example, the ‘Material’ attribute in the Delphi Survey originally consists of 36 values. After cleaning, there are total 29 different values (instances of Material) in the data set.

3.2.2 Data Reduction

Data reduction includes dimension reduction and instance selection. This section describes these two tasks for each of the datasets.

Delphi

The original Delphi dataset has ten attributes. They are 'Building type', 'Component', 'Measure', 'Environment', 'Material', 'Maintenance', 'Mode', 'Mean', 'SD' and 'Criteria'. The estimated service life was stored in two forms: the mode and the mean as well as a standard deviation (SD) for the mean. The mode is the range (e.g. 5-10) of 'service life', 'aesthetic life' or 'time to first maintenance'. The mean is the average year of 'service life', 'aesthetic life' or 'time to first maintenance'. As we want a real value to be the final predicted result, the attribute 'mean' is chosen as the target attribute and hence the 'Mode' is removed since 'Mean' and 'Mode' are different forms for the same information. 'SD' can not be considered as input because it is a part of output. 'Criteria' relates to how good the agreement was in the response from the Delphi Survey. It is not useful in mining and should be removed. This dataset contains life information of service life, aesthetic life and time to first maintenance. As we are only interested in service life, those instances whose value of 'Measure' is not equal to 'Service Life' are removed. After removing those instances, the attribute 'Measure' becomes unary and hence should be removed. The remaining attributes that are included in analysis are as follows:

Building type	Component	Environment	Material	Maintenance	Mean

Holistic-I

The original Holistic-I dataset has 11 attributes, in which 'LocID' and 'Location' are identification information and 'State' and 'Building Type' only have one value. After removing those irrelevant attributes, the attributes are as follows:

XLong	YLat	SALannual	Material	Gutter Position	Gutter Maintenance	MLannual

As described previously, the service life is calculated based upon 'MLannual'. We create a target variable named 'Service Life' and remove the false predictor 'MLannual'. Therefore, the attributes are as follows:

XLong	YLat	SALannual	Material	Gutter Position	Gutter Maintenance	Service Life

Holistic-II

The original Holistic-II dataset has 22 attributes, in which 'LocID' is identification information and 'Building Type', 'Position', 'Material', 'Building Face' and 'BuildingFacePos' only have one value. 'Exposure' and 'PositionVsExposure' are two

attributes which are correlated to each other. For example, when ‘Exposure’ is equal to ‘open’, ‘PositionVsExposure’ must be equal to ‘openly exposed to rain and sky’. Therefore, they are redundant to each other and one of them should be removed. After removing these irrelevant attributes, the attributes are as follows:

XLong | YLat | SALannual | Exposure | Gutter Type | rain_annual_mm | cum_MZa_2ndYear | cum_dSTEEL_2ndYear | remCr | normCr | accelerated_corrosion_rate | Time to White Rust of Zincalume | Time to penetration of Zincalume | Time to onset of Red Rust | Life of gutter at 600um

‘Life of gutter at 600um’ is the target attribute.

Holistic-III

The Holistic-III dataset is divided into two parts in terms of different target attributes: one is for ‘Zincalume Life’ named Holistic-III_Zi and the other is for ‘Galvanized Life’ named Holistic-III_Ga. The attribute ‘Centre Code’ and ‘Centre Name’ are ignored since they are identification information. After that, their attributes are as follows:

Holistic-III_Zi:

Longitude | Latitude | Salt Deposition | Zincalume Mass Loss | Marine | N | Zincalume Life

Holistic-III_Ga:

Longitude | Latitude | Salt Deposition | Zinc Mass Loss | Steel Mass Loss | Marine | Nzinc | Nsteel | L | M | Zinc Life | Steel Life | Galvanized Life

3.3 Data Analysis

After data pre-processing, the datasets were analysed in terms of the type of features and their availability of values as user inputs in order to determine the learning method and the input attributes. For all datasets, both discrete and continuous features exist. Therefore, a learning method for handling both discrete and continuous data is required.

The data mining system lifecycle includes three main phases: (1) training of the model, (2) evaluation (or testing) of the model and (3) using the trained model in practice. If the user can not provide the same inputs as used for training the model in the use of the model phase, the performance of the predictive model degrades due to the absence of many input values. Therefore, the availability of features in the use of model phase is an important aspect to influence the model performance. Based on the availability of

features in the use of the model phase, we simply divide all features into two groups: *available features* which are features that can be provided by users and *unavailable features* which are features that can not be provided by users. Our datasets contain some unavailable features. More specifically, all features in Delphi are available features while in Holistic-I, 'SALannual' is an unavailable feature; in Holistic-II, all other features except 'XLong', 'YLat', 'Exposure' and 'Gutter Type' are unavailable and in Holistic-III, only 'Longitude', 'Latitude' and 'Marine' are available to users. Hence, how to deal with these unavailable features is a research issue to be addressed. The literature on related data mining applications shows that most research work [9-19] aims to build the predictive models and improve the prediction accuracy. None of the existing work involves solving the problem of the reduced performance of the predictive model when the model is trained with some unavailable features.

Moreover, our datasets include multiple data sources of service life information. These sources can not be combined and the models are required to be constructed independently from each of them. However, the predicted results from different models can be compared to each other. For example, both Delphi and Holistic-II can be used to predict the lifetime of gutters with Colorbond material. The results from Delphi and Holistic-II may be inconsistent. Hence, we used the knowledge base based on the expert knowledge to choose the most appropriate answer for a given situation in case of inconsistencies in the results of different models.

4. QUERY BASED LEARNING SYSTEM

As discussed previously, the current KDDM process models are data-oriented rather than user-oriented. The data-oriented process models emphasise the data analysis tasks surrounding the data and lack the interactive involvement of users during the process. Hence, they do not suffice to address the problems that are due to user interaction during the use of the model phase. This section will propose a user-oriented learning system, namely the Query Based Learning System (QBLS), which is based on a data-centric model with extensions to provide support for user interaction.

4.1 Motivations for QBLS

Due to user interaction, problems arise during the use of the model phase. One such problem is the availability of features in the use of the models. Neither keeping both available and unavailable features nor simply removing unavailable features is a good solution. A suitable feature selection algorithm is required to minimise the number of unavailable features and maximise the classification accuracy. Meanwhile, when the user can not input the values of those unavailable features for querying to the system, some pre-processing should be done for missing input values. Moreover, the data mining process is usually carried out by a data analyst and the knowledge or model generated from the data mining process is too complex to be understood by the user. In

order to ensure the end product (knowledge or model) will be useful to the user, some post-processing is needed, such as interpreting the discovered knowledge in such a way that the user can use it. In our case, post-processing can eliminate the conflicting results from multiple data sources. Hence, we propose a new learning system framework, called the Query Based Learning System (QBLS), which is based on the data-centric process model. A domain knowledge base is introduced for pre-processing missing input values and post-processing inconsistent results.

4.2 Overview of QBLS

The QBLS is developed based on an industry standard data mining process model, CRISP-DM (Cross Industry Standard Process for Data Mining) (2003). Four procedures that are different from the CRISP-DM are highlighted in Figure 1. The three procedures - Query Based Feature Selection, Results Post-processing and the Use of Model - are critical for the success of the proposed QBLS model. The Query Based Feature Selection is separated from the data pre-processing step as it has the involvement of users or domain experts and hence is different from the usual feature selection. The Results Post-processing and the Use of Model phase are added into the model in order to ensure the results are useful to users. An external domain knowledge base is involved in results post-processing and missing inputs pre-processing in the Use of Model phase. The next section will discuss each phase of the QBLS model.

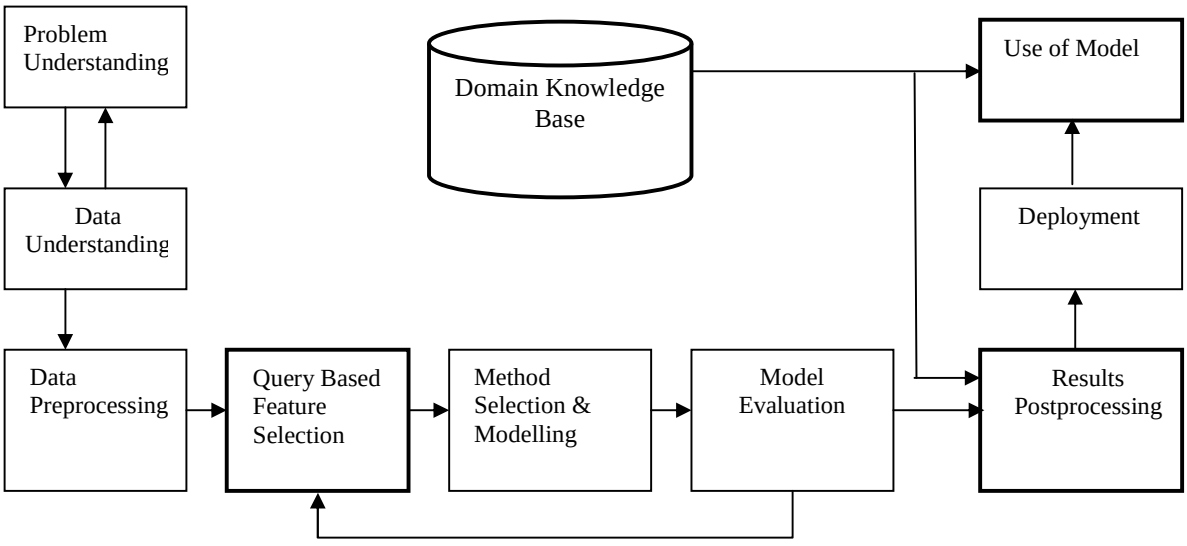

Figure 1. Query Based Learning System

4.3 Phases of QBLS

The Problem Understanding phase, like the Business Understanding phase in the CRISP-DM model, focuses on understanding the project objectives and requirements and then converting them into a data mining problem definition.

The Data Understanding phase is for identifying data quality problems and exploring the interesting subsets of data.

The Data Pre-processing phase involves preparing the datasets for applying the Query Base Feature Selection algorithm, which includes data cleaning and data reduction.

The Query Based Feature Selection phase involves selecting the final features of the dataset, which will be used to build the model. The basic idea of this phase is to select a minimum subset of relevant features with which the predictive model provides an acceptable performance, as well as, to make the selected features available to users when the model is used in practice.

The Method Selection and Modeling phase is for selecting and applying various data mining techniques to the prepared data. The models are constructed in this phase.

The Model Evaluation phase includes performance measures from both a technical perspective and business perspective.

The Results Post-processing phase includes interpretation of the mined patterns/ discovered knowledge and elimination of unreasonable results to ensure the end product will be useful.

The Deployment phase covers presentation of the generated knowledge in a customer-oriented way or deploying the created model as a customer-oriented system.

The Use of Model phase involves using the deployed system in practice. In many cases, it will be the customer, not the data analyst, who will carry out this phase. The user needs in this phase will trigger new, often more constrained data mining processes.

4.4 Query Based Feature Selection

The first step of QBFS involves removing the features such as features for identification. Let $A = \{a_1, a_2, \ldots, a_k, a_{k+1}, \ldots, a_m, a_{m+1}, \ldots, a_n\}$ be a set of remaining features in a dataset. The remaining features are clustered into three groups according to their easy availability to users as follows:

- Group 1 (a_1 - a_k): Features that the user can easily provide while using the model
- Group 2 (a_{k+1} - a_m): Features that can not be provided by the user but can be obtained from the external domain knowledge

- Group 3 (a_{m+1} - a_n): Features that can not be provided by the user or obtained from domain knowledge

Group 1 will be included in the final model because features in Group 1 are not only useful in mining but also can be provided by users while they are using the model. Group 3 will be rejected because they can not be provided in model use although they have mining value. If we include the features of Group 3 in the final model, their values in new data will be missing. As a result, the generalization accuracy will decrease. A decision has to be made for features in Group 2, as they can not be provided by users but they can be obtained from external domain knowledge. If we include all the features of Group 2, the measurements to obtain some of these values may be too complex and computationally expensive. If we exclude those features, the performance of the model may not be accepted by users.

The datasets include four different sources of service life information from the Delphi Survey, Holistic-I, -II and -III, where Holistic-III was divided into two parts in terms of different target features. The multiple sources are independent but complementary to each other. Holistic-I, -II and -III relate to different component types with different materials while Delphi contains all component types with all materials. Each data source contains completely different features in which some can not be provided by users or domain knowledge.

4.4.1 Categorisation of features

Features of each data source are divided into three groups.

Holistic-I

Group 1: { XLong, YLat, Material, Gutter Position, Gutter Maintenance }

Group 2: { SALannual }

There is no feature in Group 3.

Holistic-II

Group 1: { XLong, YLat, Exposure, Gutter Type }

Group 2: { SALannual, rain_annual_mm, cum_MZa_2ndYear,
cum_dSTEEL_2ndYear, remCr, normCr, accelerated_corrosion_rate }

Group 3: { Time to White Rust of Zincalume, Time to penetration of
Zincalume, Time to onset of Red Rust }

Holistic-III_Zi

Group 1: { Longitude, Latitude, Marine }

Group 2: { Salt Deposition }

Group 3: { Zincalume Mass Loss, N }

Holistic-III_Ga

Group 1: { Longitude, Latitude, Marine }

Group 2: { Salt Deposition }

Group 3: { Zinc Mass Loss, Steel Mass Loss, Nzinc, Nsteel, L, M, Zinc Life, Steel Life }

Delphi

Group 1: { Building type, Component, Environment, Material, Maintenance }

There is no feature in Groups 2 and 3.

4.5 Domain Knowledge Base

Domain knowledge can be included in the process of data mining from the beginning of the problem understanding to the end when the result inferred by the predictive model is presented to the users while used in practice. It is necessary to understand the project objectives and requirements and then convert them into a data mining problem definition. In the proposed process model, QBLS, a domain knowledge base is used especially for results post-processing and missing input values pre-processing in the Use of Model phase. Some features included in the final model may not be directly provided by users but can be inferred by the domain knowledge base. For example, “annual rainfall” is an important factor in determining the service life of building components in civil engineering. However, while using the data mining model to predict the service life of a building component, the user will most likely provide the location and material as an input. The user may not be aware of the exact value of rainfall in the area. However, a domain knowledge base will have such information. This information can now be treated as one of the input values for the model.

Furthermore, the domain knowledge base can be used in reinforcing the outputs inferred by the predictive model. Since the real-life data mining models are for solving practical problems, the final results should be significant to users. However, mining errors are inevitable even for a perfect model. The domain knowledge base is used to confirm that the results predicted by the data mining system do abide by the rules of the domain and/or domain experts. For example, it is domain knowledge in civil engineering that (1) a roof in a severe marine location will not last longer than one in a benign environment, and (2) a stainless steel roof should last longer than one with galvanized steel. Such in-

built rules will be checked to ensure the correctness of the results processed by the models.

In general, the external domain knowledge base assists to deal with the vague queries in use of the model phase and with eliminating illogical outcomes in post-processing. The domain knowledge base is extensible with the use of the system in real-life practice.

5. PREDICTOR SELECTION

This section will explore various predictive data mining techniques to apply to the selected features for building the predictors to determine the service life of metallic components in buildings. The primary objective is to find the best method for the building service life prediction problem. For this purpose, two types of data mining methods, namely classification methods and regression methods are applied for comparison. The following sections will discuss each of the methods involved and present the experimental results conducted to achieve the research objective. An integrated method of combining M5 and KNN will also be provided for improving the performance of predictors.

5.1 Methods Selection

There are various data mining methods such as Naïve Bayes (Fayyad et al., 1995b), K Nearest Neighbors (KNN) (Aha et al., 1991), regression modelling, support vector modelling (SVM) (Vapnik, 1995), decision tree (DT) (Quinlan, 1986) and neural network (NN) (Resample, 2003) that can be considered to undertake prediction tasks. These methods can be categorised into two groups, namely classification methods and regression methods, based on the type of target feature. Classification methods require categorical class as the target feature while regression methods work for numeric prediction. Typical classification methods include Naïve Bayes, KNN, DT, NN, and SVM. Naïve Bayes is a statistical-based algorithm. It is useful in predicting the probability that a sample belongs to a particular class or grouping (Fayyad et al., 1995b). KNN is based on the use of distance measures. Both DT (Quinlan,1986) and NN are very popular methods in data mining. DT is easy to understand and better in classification problems while NN can not produce comprehensible models in general and is more efficient for predicting numerical targets. SVM is relatively new method. It can solve the problem of efficient learning from a limited training set. For Naïve Bayes and DT, before they are applied to do numeric prediction tasks, the target feature needs to be discretised to a nominal type. Others like KNN, NN and SVM can predict the continuous value directly.

Linear regression, logistic regression, regression trees, KNN, M5 model trees (Quinlan, 1992), NN and SVM are typical regression methods. Linear regression and logistic regression are statistical-based algorithms and they are the most popular regression techniques. Model trees and regression trees are tree-based algorithms and efficient for

large datasets. Model trees are generally much smaller than regression trees and prove to be more accurate (Quinlan, 1997).

For comparison purposes, experiments were conducted on both classification and regression methods. Naïve Bayes and DT (C4.5) were chosen as representative classification methods as they are statistical-based and tree-based algorithms respectively. Linear regression, KNN, M5 model trees, NN and SVM were also chosen as representative regression methods as they are based on different theory. All the experiments were conducted in a WEKA environment and tenfold cross validation (10-CV) was used throughout the experiments described in this chapter.

The *n*-fold cross validation (*n*-CV) is a popular method used to test the performance. The idea behind *n*-fold cross validation is that a dataset is randomly evenly divided into *n* parts, *n*-1 parts of which are used as a training set for building a predictive model and the remainder is used as a test set. This process is repeated *n* times. Each time a different one of *n* parts is chosen as the test set. The performance is reported as average of *n* runs.

5.2 Experiments using Classification Methods

The first experiments were conducted using classification methods, that is, Naïve Bayes and DT (C4.5). The MDL discretisation method (Fayyad and Irani, 1992) was applied first to discretise the target feature to a nominal type. Table 1 shows the number of target classes after discretisation and the percentage of numerical and categorical attributes in datasets. Table 2 presents the classification accuracy of Naive Bayes and C4.5.

Table 1. Details of Datasets

Dataset	***No. of Cases***	***No. of Target Classes***	***No. of Input Attributes***	***Numerical Attributes (%)***	***Categorical Attributes (%)***
Delphi Survey	*683*	*10*	*7*	*0%*	*100%*
Holistic-I	*9640*	*10*	*6*	*50%*	*50%*
Holistic-II	*4780*	*10*	*13*	*76.92%*	*23.08%*
Holistic-III_Ga	*1297*	*10*	*12*	*91.67%*	*8.33%*
Holistic-III_Zi	*1297*	*9*	*6*	*83.33%*	*16.67%*

Table 2. Classification Accuracy of Naïve Bayes & DT (C4.5)

Dataset	***Classification Accuracy***	
	Naive Bayes	***DT (C4.5)***

Delphi Survey	*30.0587%*	*36.217%*
Holistic-I	*89.744%*	*90.125%*
Holistic-II	*94.728%*	*96.548%*
Holistic-III_Ga	*93.138%*	*94.603%*
Holistic-III_Zi	*91.904%*	*93.215%*

The results from Table 2 show that for Naive Bayes and C4.5, classification accuracy is around 90% except for the Delphi Survey. Both Naive Bayes and C4.5 are not good for the Delphi Survey (only 30.0587% and 36.217% classification accuracy - that means more than half the cases are not correctly classified). The highest accuracy is for Holistic-II (94.728% from Naïve Bayes and 96.548% from C4.5). Decision tree is a good classification method but seems less appropriate for estimation tasks where the goal is to predict the value of a continuous attribute. Transforming our prediction problem to a classification problem by discretising continuous values to categorical values proved not suitable on our datasets, especially for the Delphi Survey.

Moreover, we can observe from Table 6 that the numbers of classes for all datasets are almost the same while the number of cases varies from 683 to 9640. There are ten classes while only 683 cases in the Delphi Survey. Therefore, it may be true that the decision tree is prone to errors in classification problems with many classes and a relatively small training set.

5.3 Experiments using Regression Methods

The second experiments were conducted using regression methods, that is, linear regression, KNN, M5, NN and SVM. The average correlation coefficients over 10-CV of these algorithms on our datasets are reported in Table 3.

The results in Table 3 show that good results are achieved for all methods. Most of the correlation coefficients (CCs) are above 0.95. The lowest CC is 0.797 (KNN for Delphi Survey) and the highest is 1 (NN and M5 for Holistic-II). NN works best for all datasets, getting very high CC for all datasets. This result proves that NN is very efficient for handling numerical values and well-suited for predicting a numerical target because most of the attributes in our datasets are numerical values (the last two columns of Table 1 show the percentage of numerical and categorical attributes - it is obvious that almost all datasets have more than 50% numerical attributes). The correlation coefficients of SVM are closer to NN, only the value for Holistic-I is much reduced. The results from KNN are also similar to NN, even better for Holistic-I. KNN obtained the worst result for the Delphi Survey. This may prove that KNN is quite effective if the training set is large. There are 9640 cases in Holistic-I, 4780 cases in Holistic-II, 1297 cases in Holistic-III while only 683 cases in the Delphi Survey. M5 is learned efficiently as NN. Especially, it is better for the Delphi Survey than NN.

Table 3. Correlation Coefficient of KNN, NN, SVM & M5

Dataset	*Correlation Coefficient (CC)*				
	Linear regression	***KNN***	***NN***	***SVM***	***M5***
Delphi Survey	*0.9320*	*0.7970*	*0.9299*	*0.9280*	*0.9333*
Holistic-I	*0.8679*	*0.9960*	*0.9790*	*0.8412*	*0.9892*
Holistic-II	*0.9999*	*0.9962*	*1*	*0.9999*	*1*
Holistic-III_Ga	*0.9678*	*0.9915*	*0.9994*	*0.9737*	*0.9883*
Holistic-III_Zi	*0.9038*	*0.9886*	*0.9990*	*0.9889*	*0.9971*

From the view of each dataset, Holistic-II gets the best result. That is because Holistic-II contains more valuable features than others for predicting service life. The CC from all methods for Holistic-II is very high (the highest reaches 1 while the lowest is also 0.9962). The results for the Delphi Survey are the worst (the highest is only 0.9333 while the lowest is 0.797).

All results indicate those methods which can deal with continuous values directly such as KNN, NN, SVM and M5 are better than those that have to discretise continuous values such as Naïve Bayes and DT. However, the interesting fact is that no one method is always best for all five datasets. In order to clearly show the best method for each dataset, the information in Table 3 is presented graphically in Figure .1

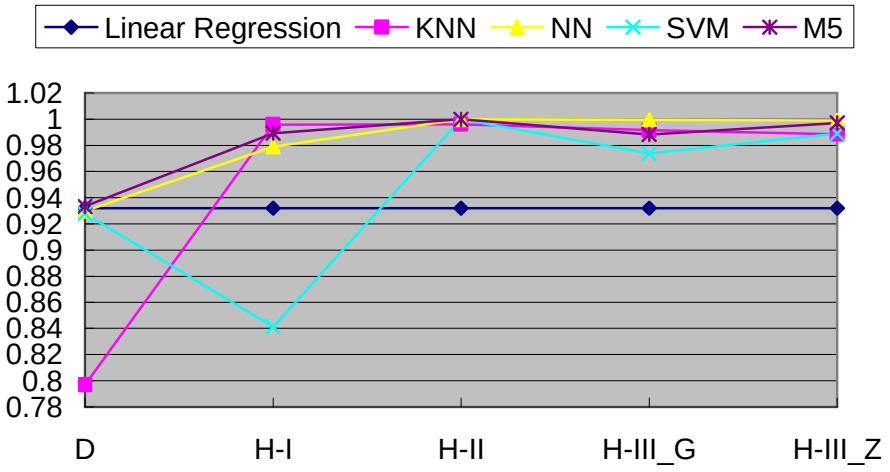

Figure .1: **Correlation Coefficient of KNN, NN, SVM & M5**

Figure .1 clearly indicates that M5 is the best method for the Delphi Survey (CC is 0.9333), KNN is the best method for Holistic-I (CC is 0.9960), NN and M5 are the best methods for Holistic-II (CC is 1) and NN is the best method for Holistic-III (CC is 0.999).

Considering the balance of accuracy and comprehensibility of predictors, M5 was chosen as the final learning method.

5.4 Predictors constructed using M5

Having chosen M5 as the learning method, it was then applied on the features selected by QBFS to build the predictors for each of the datasets. As the whole predictive model for each of the datasets is very large, a part of the M5 model tree output for Holistic-I is given as an example.

Predictor for Holistic-I

Part of M5 pruned model tree:

GutterMaintenance=cleaned <= 0.5 : LM1 (2410/2.632%)

GutterMaintenance=cleaned > 0.5 :

| GutterPosition=sides-interior,outside <= 0.5 :

| | XLong <= 151.184 :

| | | XLong <= 145.486 :

| | | | XLong <= 141.351 :

| | | | | YLat <= -21.646 :

| | | | | | XLong <= 139.491 : LM2 (4/0%)

| | | | | | XLong > 139.491 : LM3 (4/3.365%)

| | | | | YLat > -21.646 :

| | | | | | XLong <= 140.027 : LM4 (22/4.599%)

LM num: 1

ServiceLife =

*-0.0116 * XLong*

*- 0.0064 * YLat*

*- 0.0002 * SALannual*

*+ 0.0085 * Material=Zincalume*

*+ 0.0689 * GutterPosition=sides-interior,outside*

*+ 0.039 * GutterPosition=outside*

*+ 0.0345 * GutterMaintenance=cleaned*

+ 1.9424 …Eqn(2)

This is a part of the M5 model tree output using the attributes 'XLong', 'YLat', 'SALannual', 'Material', 'GutterPosition' and 'GutterMaintenance' for Holistic-I. The first part of the output shows the tree structure of the model. The output on a leaf node is a reference to a function. For example, there is a rule in the tree:

GutterMaintenance=cleaned <= 0.5 : LM1 (2410/2.632%)

This means that if this rule is true, then the output, 'ServiceLife' in this case, is decided by the linear regression equation with label LM1, namely the second part of the output above (Equation 6.2). The numerical values in parentheses (2410/2.632%) tell us 2410 instances satisfy the rule and 2.632% in the training set do not satisfy the rule.

To evaluate Equation 2, simply replace all numerical attributes (XLong, YLat and SALannual in this example) with their value for the particular instance and replace categorical expressions (such as Material=Zincalume) with the value 1 if the attribute is equal to any of the listed attributes (they are comma-delimited) or with 0 if they are false. This is the same in the model tree; any rules that involve categorical values, such as

GutterMaintenance=cleaned <= 0.5

Simply replace 'GutterMaintenance=cleaned' with the value 1 if 'GutterMaintenance' is equal to 'cleaned' or with 0 if it is false.

5.5 Improvement of Performance

The QBFS feature selection algorithm may result in some useful features being rejected; as a result, this may reduce the performance of the predictive models. The model-based learning (M5) is combined with the instance-based learning (Quinlan, 1993) to improve the performance. This method first uses the instance-based approach to find a set of instances similar to the target instance. Then, the class values of similar instances are adjusted using the value predicted by the model tree before they are combined. The

detailed algorithm is given in Figure 2. We use the KNN (K=3) for the instance-based method.

Input:

T: the Training Set

M: A predictive model constructed by the model-based method

U: an unseen instance

Output:

$V(U)$: predicted class value for U

1. $M(U) \leftarrow$ the value predicted for U by M
2. Let $P \leftarrow \{P_1, P_2, \ldots, P_k\}$ be a subset of instances similar to U by using the instance-based method
3. Let $VP \leftarrow \{V(P_1), V(P_2), \ldots, V(P_k)\}$ be a subset of class values for P
4. For i =1 to k

   $M(P_i) \leftarrow$ the value predicted for P_i by M

   $diff(i) = M(P_i) - M(U)$

   $V(P_i)' = V(P_i) - diff(i)$

5. $V(U) = \dfrac{\sum_{i=1}^{k} V(P_i)'}{k}$

Figure 2. M5 + KNN Algorithm

Therefore, the final predictors are built using M5+KNN on the features selected by QBFS. The performance of this M5+KNN combined model is compared with the M5 model and the ensemble model with bagging (Breiman, 1996). Correlation coefficient and Mean Absolute Error of M5, M5+KNN and bagging are presented in Table 4 and Table 5.

Table 4. Correlation Coefficient of M5, M5 + KNN and Bagging

Dataset	***Correlation Coefficient (CC)***		
	M5	***M5 + KNN***	***Bagging***
Delphi Survey	*0.9198*	*0.94555*	*0.9467*
Holistic-I	*0.9790*	*0.97990*	*0.9904*
Holistic-II	*0.9103*	*0.97628*	*0.9158*
Holistic-III_Ga	*0.9421*	*0.97520*	*0.9416*
Holistic-III_Zi	*0.8692*	*0.94859*	*0.8770*

Table 5: Mean Absolute Error of M5, M5 + KNN and Bagging

Dataset	*Mean Absolute Error*		
	M5	*M5 + KNN*	*Bagging*
Delphi Survey	*3.3272*	*2.7526*	*2.7686*
Holistic-I	*0.9113*	*0.5094*	*3.0823*
Holistic-II	*2.3758*	*1.1414*	*2.3177*
Holistic-III_Ga	*2.1044*	*0.9857*	*2.1486*
Holistic-III_Zi	*2.9378*	*1.2025*	*2.9157*

The same information is presented graphically in Figures 8 and 9. From Figures 8 and 9, we can observe that the better correlation coefficient and lower mean absolute error are obtained by combining the M5 and KNN learning methods. The method seems to provide significant improvement for relatively weaker models such as the Holistic-II and Holistic-III_Zi, whereas the improvement for the near-perfect models such as Holistic-I, is not so obvious. The combined M5+KNN model also outperforms the ensemble model with bagging. Bagging can not always improve the performance such as for Holistic-I as shown in Figure 3 and 9.

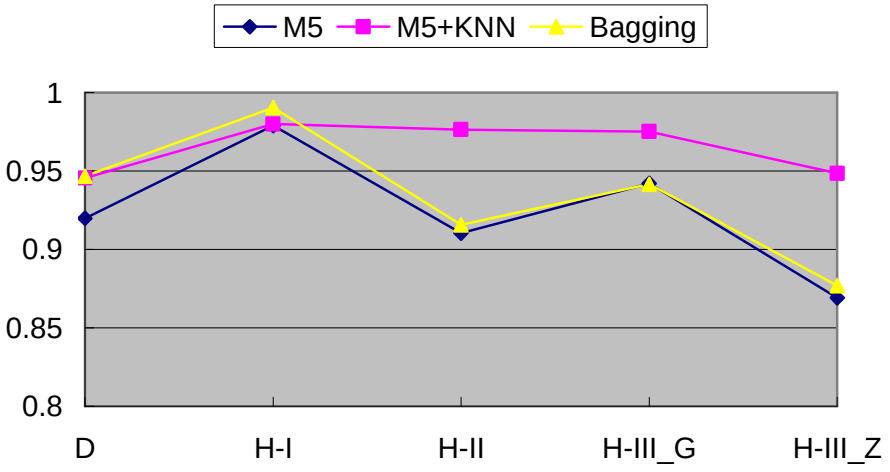

Figure 3. Correlation Coefficient of M5, M5+KNN and Bagging (D=Delphi, H-I=Holistic-I, H-II= Holistic-II, H-III_G=Holistic-III for Galvanized Steel, H-III_Z=Holistic-III for Zincalume)

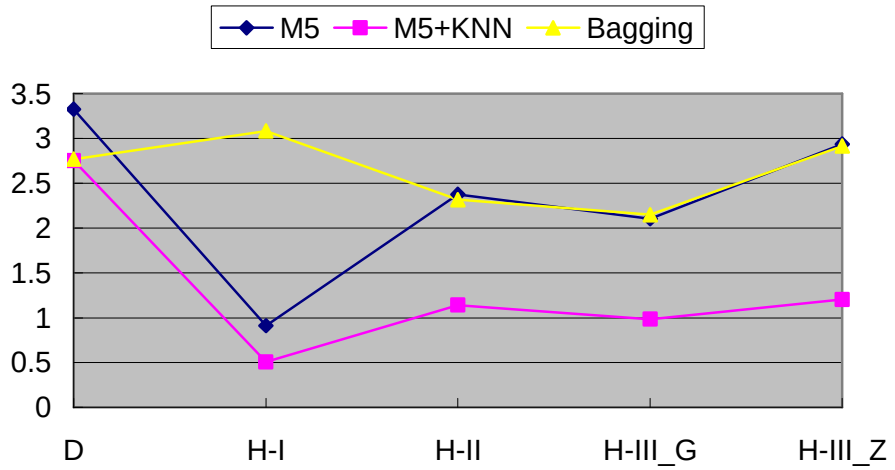

Figure 4. Mean Absolute Error of M5, M5+KNN and Bagging, (D=Delphi, H-I=Holistic-I, H-II= Holistic-II, H-III_G=Holistic-III for Galvanized Steel, H-III_Z=Holistic-III for Zincalume)

6. OVERALL SOLUTION

In the previous section, we proposed a learning system framework, the QBLS model. We also presented a summary of experimental results for choosing the best learning method. Based on the theoretical framework and practical experiments, we propose an overall solution to predict the service life of metallic components in Queensland schools. This section will describe the solution in detail and provide an example of prediction using the developed system.

6.1 Overview of the System

The overview of the system is given in Figure 1. This system basically consists of three main parts: feature selection, predictors and domain knowledge. The Query Based Feature Selection is first applied to the datasets to select a minimum subset of features which can be provided by users. Then, a hybrid method M5+KNN is applied on the selected features to build the predictors for all of the datasets. The predictors are used to carry out prediction for user input queries. The domain knowledge base consists of three parts: salt deposition knowledge, rainfall knowledge and generalised rules extracted from domain expert opinions. Because the features selected to build the predictors include features of ‘Salt Deposition’ and ‘Rainfall Annual’, the salt deposition and rainfall database is included in the knowledge base, which is for pre-processing user inputs. Generalised rules are used in post-processing the predicted results, for example, solving the inconsistency in predicted results.

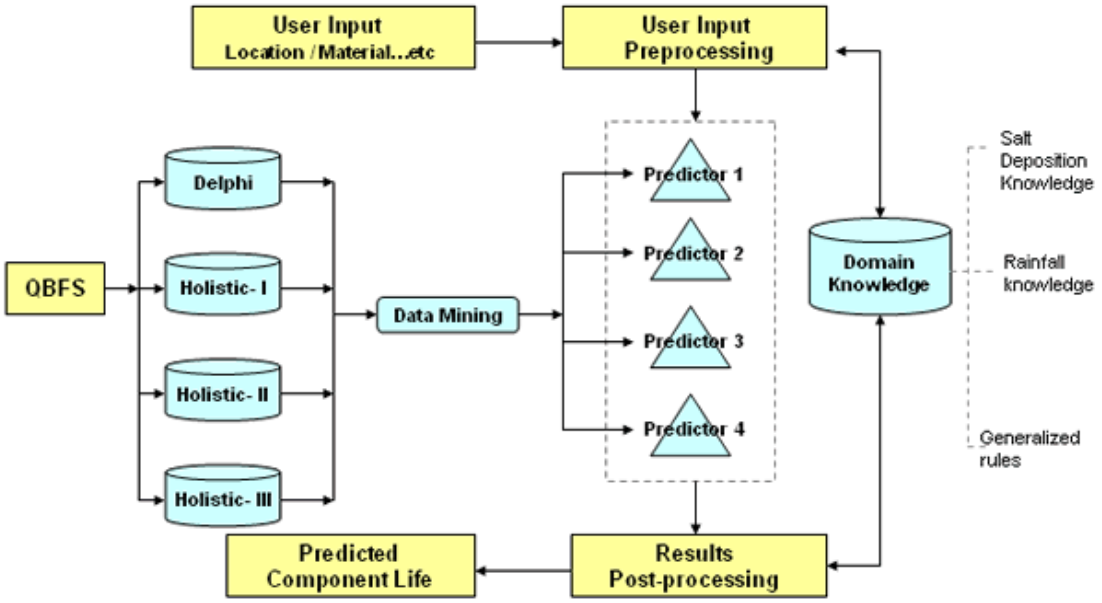

Figure 1. Overview of System

6.2 Representation of the Knowledge Base

Construction of the knowledge base, consisting of the salt deposition, annual rainfall and generalised rules, has been generated for the purpose of pre-processing vague queries and post-processing inconsistent results. The knowledge is represented as items in the database. Some of the salt deposition knowledge in the generated knowledge base is presented in Table 1.

Table 1. Salt Deposition Knowledge

XLong	*YLat*	*Salt Deposition*
151.986	*-28.0373*	*3.80842*
153.007	*-27.3206*	*4.42054*
147.633	*-22.8372*	*3.77518*

Some of the rainfall knowledge in the generated knowledge base is presented in Table 2.

Table 2. Rainfall Knowledge

XLong	*YLat*	*Rain Annual (mm)*
151.986	*-28.0373*	*1595*
153.007	*-27.3206*	*1595*

147.633	*-22.8372*	*783*

And some of the generalised rules in the generated knowledge base are presented in Table 3.

Table 3. Generalised Rules

Component	***Environment***	***Material***	***Min (years)***	***Max (years)***
Gutters	*Marine*	*Galvanised Steel*	*5*	*15*
Gutters	*Benign*	*Galvanised Steel*	*30*	*50*
Gutters	*Benign*	*Colorbond*	*20*	*50*
Roof	*Marine*	*Colorbond*	*15*	*30*

Once the knowledge base is created, it can be used for pre-processing the user inputs and post-processing the predicted results.

As the location (longitude and latitude) that users input may not exactly match the salt deposition and rainfall knowledge, a similarity principle is employed to obtain the value of salt deposition and rainfall. The similarity principle means that the nearest geographic location will have the most similar value for salt deposition and annual rainfall. The distance *D* between two points on the surface on the earth is computed by the formula 3 (Cole et al., 2005).

$$D = R \times \cos^{-1}\left\{\left[\sin\left(\frac{latitude_1}{57.2956}\right)\times\sin\left(\frac{latitude_2}{57.2956}\right)\right]+\left[\cos\left(\frac{latitude_1}{57.2956}\right)\times\cos\left(\frac{latitude_2}{57.2956}\right)\times\cos\left(\frac{longitude_2}{57.2956}-\frac{longitude_1}{57.2956}\right)\right]\right\}$$

…(Eqn.3)

Where:

The location of the first point is given by ($longitude_1$, $latitude_1$);

The location of the second point is given by ($longitude_2$, $latitude_2$);

And longitudes and latitudes are measure in decimal degrees;

R is the radius of the earth taken as 6378.7 km.

To covert latitude or longitude from decimal degrees to radians, the latitude and longitude values are divided by $180/\pi \approx 57.2956$ (taking π to be 3.1416).

Once the user inputs longitude and latitude, the system will find the nearest location from the knowledge base and then get the value of salt deposition and rainfall. These values can then be treated as user inputs for the predictors.

In terms of predicted results, the system checks them with the generalised rules. If the component, material and environment are matched and the predicted service life is in

the range, the results are reasonable. Otherwise we suggest that the result does not abide by the generalised rules.

6.3 An Example of Prediction using the System

A prediction system has been developed in this research project. Figure 2 shows the user interface of the system.

As shown in the user interface, the location, component and material are compulsory inputs for querying to the system. Based on these three inputs, different predictors will be used to do the prediction. Here we provide an example for using the system. Suppose the user wants to know the service life of gutters with galvanized steel in location (151, -28). He/she first inputs (151, -28) as location, gutters as component and galvanized steel as material. The location inputs can also be directly selected from the geo-spatial database using the GIS system. Then Holistic-I and Delphi options are activated for more inputs needed by these two predictors. After the user inputs the gutter position, maintenance and environment etcetera, the system automatically gets values from domain knowledge for other features needed by the predictors. For example, the Holistic-I predictor requires salt deposition in this location as an input as well. The system gets the salt deposition from the salt database and predicts the service life is 14.5004 years from the Holistic-I predictor. A similar process is done by the Delphi predictor and the predicted service life is 14.4165 years. The results for this example are quite consistent. However, sometimes the results from different predictors will conflict with each other. An example of such a case is the service life of roof with Zincalume in location (153.0310, -27.4315). The predicted result from the Delphi predictor is 51.877 years while from the Holistic-III predictor is only 29.928 years. In such a case, domain knowledge is also used to eliminate unreasonable results.

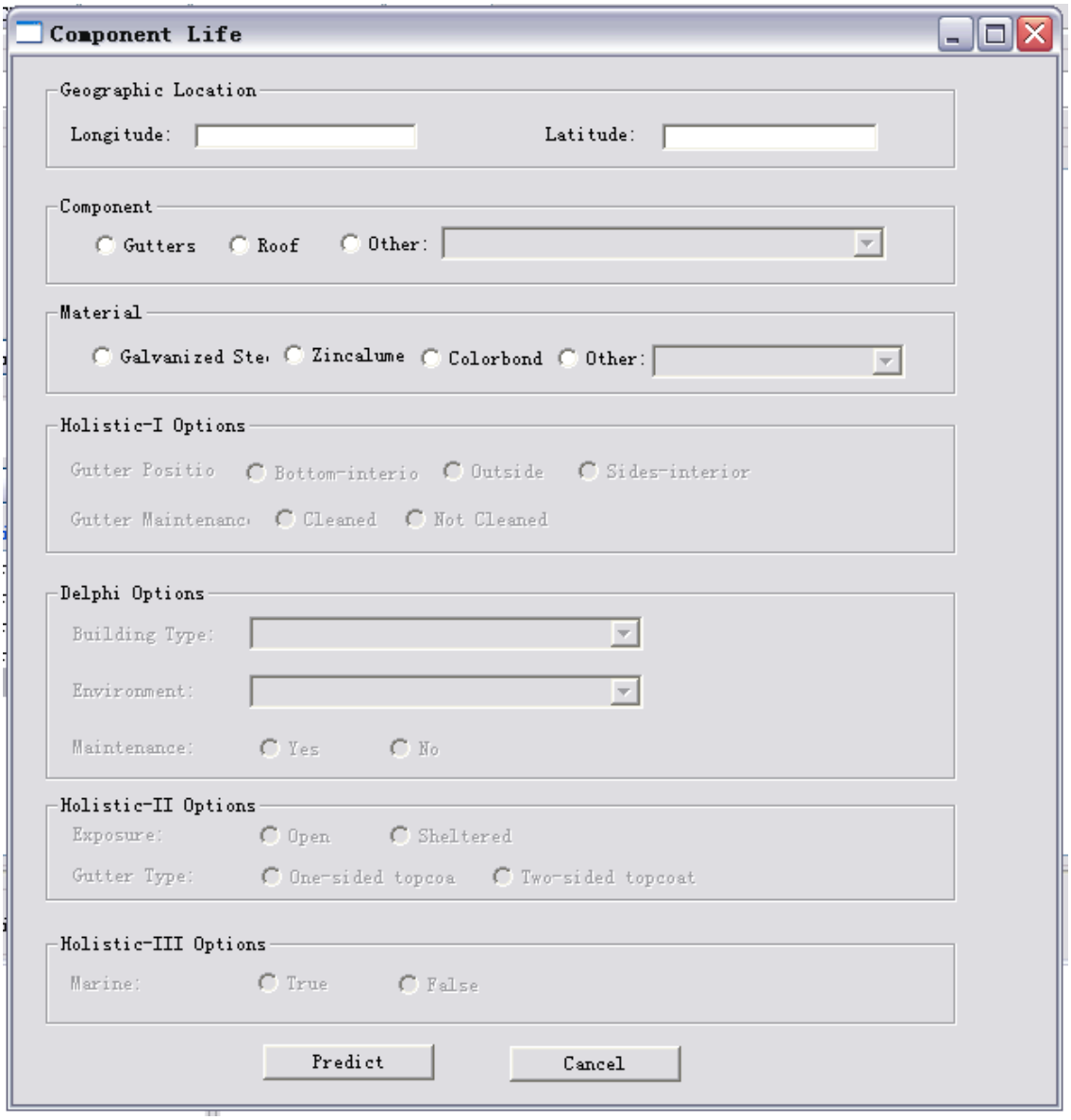

Figure 2. User Interface of System

7. CONCLUSIONS

The main objective of this research is to develop a prediction tool for accurately estimating the service life of metallic components and hence provide economic benefits to industry partners of this project. To achieve this objective, we have proposed a user-oriented learning system framework, namely QBLS, for solving the problem of using the data mining models in a real-world situation where the user can not provide all the inputs

with which the model is built. A practical prediction system is developed based on the QBLS framework, which provides high accuracy in practice where not all inputs are available for querying to the system.

8. REFERENCES

Aha, D. W., Kibler, D. and Albert, M. K., (1991) "Instance-Based Learning Algorithms," *Machine Learning*, vol. 6, pp. 37 - 66.

Breiman, L., (1996) "Bagging Predictors," *Machine Learning*, vol. 24, pp. 123-140.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J., (1984) "Classification and Regression Tree".

Brence, J. R., and Brown, D. E., (2002) "Data mining corrosion from eddy current non-destructive tests," *Computers & Industrial Engineering*, vol. 43, pp. 821-840.

Chapple, M., (2006) "Data Mining: An Introduction," vol. 2006.

Cole, I.S., Ball, M., Carse, A., Chan, W. Y., Corrigan, P., Ganther, W., Muster, T., Paterson, D., Trinidad, G., Maher, M. L. and Liew, P.S, (2005) "Case-Based Reasoning in Construction and Infrastructure Projects - Final Report," 2002-059-B, March 2005.

Cole, I.S., Trinidad, G., Bradbury, A., McFallen, S., Chen, S.-E., MacKee, J., Gilbert, D. and Shutt, G.(2004) "Final Report of Delphi study," CRC Report 2002-020-B.

CRISP-DM, "Cross Industry Standard Process for Data Mining", 2003.

Dunham, M. H., (2003) *Data mining introductory and advanced topics*: Upper Saddle River, NJ : Prentice Hall/Pearson Education.

Fayyad, U. M. and Irani, K. B., (1992) "On the Handling of Continuous-Valued Attributes in Decision Tree Generation," *Machine Learning*, vol. 8, pp. 87—102.

Fayyad, U. M., Piatetsky-Shapiro, G. and Smyth, P., (1995a) "From Data Mining to Knowledge Discovery: An Overview," in *Advances in Knowledge Discovery and Data Mining*, U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds. Menlo Park: AAAI Press, pp. 1 - 34.

Fayyad, U. M., Piatetsky-Shapiro, G. and Smyth, P., (1995b)"Bayesian Networks for Knowledge Discovery," in *Advances in Knowledge Discovery and Data Mining*, U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds. Menlo Park: AAAI Press, pp. 273 - 305.

Fayyad, U. M., Piatetsky-Shapiro, G. and Smyth, P, (1996a) "Knowledge discovery and data mining: towards a unifying framework," presented at Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R., (1996b) *Advances in Knowledge Discovery and Data Mining*: AAAI Press.

Furuta, H., Deguchi, T. and Kushida, M., (1995) "Neural network analysis of structural damage due to corrosion," presented at Proceedings of ISUMA - NAFIPS '95 The Third International Symposium on Uncertainty Modeling and Analysis and Annual Conference of the North American Fuzzy Information Processing Society.

KamrunNahar, M., and Urquidi-Macdonald, M., (2005) "Data mining of experimental corrosion data using Neural Network," presented at 208th Meeting of the Electrochemical Society, Oct 16-21 2005, Los Angeles, CA, United States.

Kantardzic, M. and Zurada, J. , (2005) *Next Generation of Data-Mining Applications*: Wiley-IEEE Press.

KDnuggets, (2005)"Successful Data Mining Applications,".

Kessler, W., Kessler, R. W., Kraus, M., Kubler, R. and Weinberger, K., (1994) "Improved prediction of the corrosion behaviour of car body steel using a Kohonen self organising map," presented at Advances in Neural Networks for Control and Systems, IEE Colloquium.

Kurgan, L. A., Alberta, K. and Musilek, P., (2006) "A survey of Knowledge Discovery and Data Mining process models," *The Knowledge Engineering Review*, vol. 21, pp. 1 - 24.

Leu, S.-S., Chen, C.-N. and Chang, S.-L., (2001) "Data mining for tunnel support stability: neural network approach," *Automation in Construction*, vol. 10, pp. 429-441.

Melhem, H. G. and Cheng, Y. (2003) "Prediction of remaining service life of bridge decks using machine learning," *Journal of Computing in Civil Engineering*, vol. 17, pp. 1-9.

Melhem, H. G., Cheng, Y., Kossler, D. and Scherschligt, D., (2003) "Wrapper Methods for Inductive Learning: Example Application to Bridge Decks," *Journal of Computing in Civil Engineering*, vol. 17, pp. 46-57.

Melli, G., Zaiane, O.R., and Kitts, B. (2006) "Introduction to the special issue on successful real-world data mining applications," *SIGKDD Explor. Newsl.*, vol. 8, pp. 1-2.

Mita, A. and Hagiwara, H., (2003) "Damage Diagnosis of a Building Structure Using Support Vector Machine and Modal Frequency Patterns," presented at Smart Structures and Materials 2003: Smart Systems and Nondestructive Evaluation for Civil Infrastructures, Mar 3-6 2003, San Diego, CA, United States.

Morcous, G., Rivard, H., and Hanna, A. M., (2002a) "Case-Based Reasoning System for Modeling Infrastructure Deterioration," *Journal of Computing in Civil Engineering*, vol. 16, pp. 104-114.

Morcous, G., Rivard, H., ASCE, A., Hanna, A.M., A. M. and ASCE, F., (2002b) "Modeling Bridge Deterioration Using Case-based Reasoning," *Journal of Infrastructure Systems*, vol. 8, pp. 86-95.

Piatetsky-Shapiro, G., (1991) "Knowledge discovery in real databases: A report on the IJCAI-89 Workshop," *AI Magazine*, vol. 11, pp. 68 - 70.

Quinlan, J. R., (1986) "Induction of decision trees," *Machine Learning*, vol. 1, pp. 81-106.

Quinlan, J. R., (1992) "Learning with Continuous Classes," presented at 5th Australian Joint Conference on Artificial Intelligence.

Quinlan, J. R., (1993) "Combining instance-based and model-based learning," presented at Proceedings of the Tenth International Conference on Machine Learning, Amherst, Massachusetts.

Resampling and Stats, "Neural Networks Classification," vol. 2006, 2003.

Skomorokhov, A. O., (2000) "A knowledge discovery method - APL implementation and application," presented at Proceedings of the APL Berlin 2000 Conference, Jul 24-27 2000, Berlin, Germany.

Vapnik, V. N., (1995) *The Nature of Statistical Learning Theory*. New York: Springer-Verlag.

Cooperative Research Centre for Construction Innovation

9th Floor, L Block
QUT Gardens Point
2 George Street
BRISBANE QLD 4001
AUSTRALIA

Tel: +61 7 3138 9291

Fax: +61 7 3138 9151

Email:
enquiries@construction-innovation.info

Web:
www.construction-innovation.info

Established and supported under the Australian Government's Cooperative Research Centres Program