

Data Mining for Lifetime Prediction of Metallic Components

Esther Ge¹ Richi Nayak¹ Yue Xu² Yuefeng Li²

¹ School of Information Systems, ² School of Software Engineering and Data Communications

Queensland University of Technology

GPO Box 2434, Brisbane Qld 4001, Australia

{e.ge,r.nayak,yue.xu,y2.li}@qut.edu.au

Abstract

The ability to accurately predict the lifetime of building components is crucial to optimizing building design, material selection and scheduling of required maintenance. This paper discusses a number of possible data mining methods that can be applied to do the lifetime prediction of metallic components and how different sources of service life information could be integrated to form the basis of the lifetime prediction model.

Keywords: Data Mining, prediction, corrosion, civil engineering

1 Introduction

Our globe is increasingly challenged by growing populations and aging infrastructure. An escalating demand to maintain the infrastructure is always at place. Service life of building components is a key issue in predictive and optimizing design and management of buildings and civil infrastructures. It is influenced by many factors like materials, environment and maintenance etc. The corrosion of metallic components is the main factor that influences the service life of building.

Recent Australian research found there are over 300 metallic components with 2-3 materials and 2-3 coatings in a standard Australian house (I. Cole et al., 2006). Those components in general have been exposed to environments. Corrosion decay is very serious in metallic components due to sunlight, rain and salt deposition. Therefore, it is necessary to develop efficient means of estimating corrosion rate and then the service life. The need to develop accurate methods to predict the lifetime of metallic components has become an international recognition. For example, the European Performance Based Building network and the CIB working group W80 on design life of buildings is working on the further development of the Factorial Approach to predict the service life of building components (I. Cole et al., 2006).

The material should be selected to match the severity of the environment. For example, in severe marine locations, very durable materials need to be selected while in benign environments lower quality products can be used. As with

materials selection, the timing of maintenance and building design would be tailored to the severity of the environment. Through these ways, substantial cost savings can be made. For example, it has been estimated that nearly \$5 million was spent by Queensland Department of Public Works (QDPW) in 03/04 in replacing corroded metallic components in Queensland schools (I. Cole et al., March 2005).

Data mining is a powerful technology to solve prediction problem (Fayyad, Piatetsky-Shapiro, & Smyth, 1995b). It has been effectively applied to civil engineering for corrosion prediction. For example, Kessler et al. (1994) improved prediction of the corrosion behavior of car body steel using a Kohonen self organizing map. Furuta et al. (1995) developed a practical decision support system for the structural damage assessment due to corrosion using Neural Network. More recently, Morcoux et al. (2002) proposed a case-based reasoning system for modelling infrastructure deterioration. Melhem and Cheng (2003) first used KNN and Decision Tree for estimating the remaining service life of bridge decks. And 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. However, limited research was conducted on comparing the prediction accuracy of various methods and how we can get the best prediction accuracy to the corrosion rates.

This paper discusses a number of possible data mining methods that can be applied to do the lifetime prediction of metallic components and how different sources of service life information could be integrated to form the basis of the lifetime prediction model. Experiments are conducted with Weka, a public data mining tool.

2 Data Acquisition

The data sets include three different sources of service life information: Delphi Survey, Maintenance database, and Holistic Corrosion Model. The Delphi Survey includes the estimation of service life for a range of metallic components by experts in the field such as builders, architects, academics and scientists. Maintenance database is derived from the maintenance records that provide a repository of past experiences on component lifetime predictions under specific conditions. Holistic Model is based on a theoretical understanding of the basic corrosion processes (I. S. Cole, Furman, & Ganther, 2001). It provides the required knowledge for computing the lifetime of metallic components. An independent model for Colorbond is included in the Holistic model since Colorbond has different features from other materials. Details of these data sets are presented in Table 1.

Data Set	Number of Cases	Number of attributes	Target attribute
Delphi Survey	683	10	Mean
Maintenance Database	1297	18	Zincalume Life
			Galvanized Life
Holistic Model	9640	11	MLannual
Colorbond	4780	20	Life of gutter at 600um

Table 1: Details of Data Sets

The Delphi survey data set contains the predicted life information for over 30 components, 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 data set is an estimated components life. The estimated life was stored in two forms: the mode and the mean as well as a standard deviation for the mean. The mean is the average years of service life, aesthetic life or time to first maintenance. As the Delphi dataset is the result of surveys, the final dataset was examined in three ways to determine its accuracy and reliability. They 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 (I. Cole et al., March 2005).

The maintenance data set contains life information of roof component for schools in Queensland. They are the results of analysing over 10000 records with regard to significant maintenance. The outputs are service life of Zincalume and Galvanized Steel materials for roofs.

The holistic data set contains theoretical information of corrosion for gutters in Queensland schools. The overall model is a reflection of influence of climatic conditions and material/environment interactions on corrosion. The output of this data set is the annual mass loss of Zincalume or Galvanized steel. Once the mass loss of material is determined, its service life is measured with appropriate formulas (I. Cole et al., March 2005).

Because Holistic model has no facility for handing the particular material Colorbond, the rules for the degradation of Colorbond is devised separately. The Colorbond data set includes this information. The output of Colorbond is service life of Colorbond for gutters.

In general, Delphi Survey is expert opinions; Maintenance database is operational while Holistic Model is theoretical. They form three important source of information for predicting lifetime of metallic components. They are independent but complement each other. Delphi Survey

can be used for analyzing correlation with other two data sets on component performance and consistency with expected trends based on knowledge of materials performance while Maintenance database and Holistic Model provide de facto and theoretical proof respectively for prediction. Maintenance, Holistic and Colorbond relate to different component types with different material while Delphi contains all component types with all material, which can be used to check for consistency. More specifically, Maintenance is for roofs with Galvanized Steel and Zincalume, Holistic is for gutters with Galvanized Steel and Zincalume, Colorbond dataset is for gutters with Colorbond 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 outcome from Maintenance, Holistic and Colorbond while the predicted outcome from them can be compared with the outcome from Delphi.

3 Data Mining for Lifetime Prediction

In this section, we explore various predictive data mining techniques to apply for lifetime prediction problem and to find a best one. Before a learning algorithm is applied, the data must be preprocessed (Olafsson, 2006).

3.1 Data Pre-processing

Data quality is a key aspect in performing data mining on a real-world data. Raw data generally include many noisy, inconsistent and missing values and redundant information. In this section, we explain how data is pre-processed to make data ready for mining.

3.1.1 Feature Selection

Feature selection is for removing those attributes irrelevant to mining results. In our data sets, some attributes like Centre Code, Centre Name and LocID only provide identification information. They have no mining value. Similarly, some attributes such as Building Type and Material in Colorbond contain only one value. They were also ignored during mining.

For Delphi Survey, the estimated life was stored in two forms: the mode and the mean as well as a standard deviation (SD) for the mean. As we want a real value to be the final predicted result, the attribute 'mean' is chosen as the target attribute. All other attributes are kept as inputs to know their influence to the target value. They are as follows:

Building type | Component | Measure | Environment | Material | Maintenance | Criteria | Mean

For maintenance database, there are two target attributes: Zincalume Life and Galvanized Life. After examining all attributes carefully, we found that some attributes are only related to 'Zincalume Life' while others are only related to 'Galvanized Life'. Therefore, we divided maintenance database into two parts: one is for 'Zincalume Life' and the other is for 'Galvanized Life'. The attribute 'Centre Code' and 'Centre Name' are removed since they are

identification information. The final attributes for ‘Zincalume Life’ and ‘Galvanized Life’ are as follows:

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

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

For Holistic Model, as we describe in Data Acquisition section, the service life is calculated based upon ‘MLannual’. We create a target variable named ‘Service Life’ which is calculated from formulas (I. Cole et al., March 2005). Similarly, ‘LocID’ and ‘Location’ are removed because they are identification information. ‘State’ and ‘Building Type’ are also ignored since they only have one value. Therefore, the final attributes are as follows:

XLong YLat SALannual Material Gutter Position Gutter Maintenance MLannual Service Life
--

Similar process has been done for Colorbond. ‘LocID’, ‘Building Type’, ‘Position’, ‘Material’, ‘Building Face’ and ‘BuildingFacePos’ are ignored because they are either identification information or only have one value. The final attributes are as follows:

SALannual Exposure PositionVsExposure 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.

3.1.2 Data Cleaning

In our data sets, the percent of missing values are very low. For example, for Delphi Survey, only the attribute ‘mode’ has 8% missing values while all other attributes have no missing values. For Colorbond, all attributes have no missing values. However, inconsistent values do exist in every data set. An example is the use of lowercases and capitals such as ‘Steel’ and ‘steel’. More examples are different spellings but same meaning like ‘Galvanised’ and ‘Galvanized’ or different words but same meaning like ‘Steel in Hardwood’ and ‘Steel-Hardwood’. More spaces are included in values could be another reason to cause inconsistency like ‘Residential’ and ‘Residential ’. Data mining tool will treat those kinds of values as different values and hence influence the predicted results. All such errors are recovered during data cleaning. For example, the ‘Material’ attribute in Delphi Survey originally has 36 values. After cleaning, there are only 29 values.

3.1.3 Data Discretization

Data discretization is considered because some learning algorithms are better able to handle discrete data. We discretized all numeric attributes including target attributes to nominal type by dividing them into ranges before applying Naïve Bayes and Decision Tree mining algorithm. For example, ‘Mean’ contains values from 3 to

58. It is divided into 10 ranges: [3-13], (13-17], (17-21], (21-25], (25-29], (29-33], (33-37], (37-41], (41-45], (45-58]. While for other classification data mining methods like Neural Network and SVM, we keep all continuous values.

3.2 Data Modelling and Mining

Our main objective in this research is to make an accurate prediction for the lifetime of metallic components. Therefore, our problem is a prediction data mining problem. The overflow of prediction model is given in Figure 1.

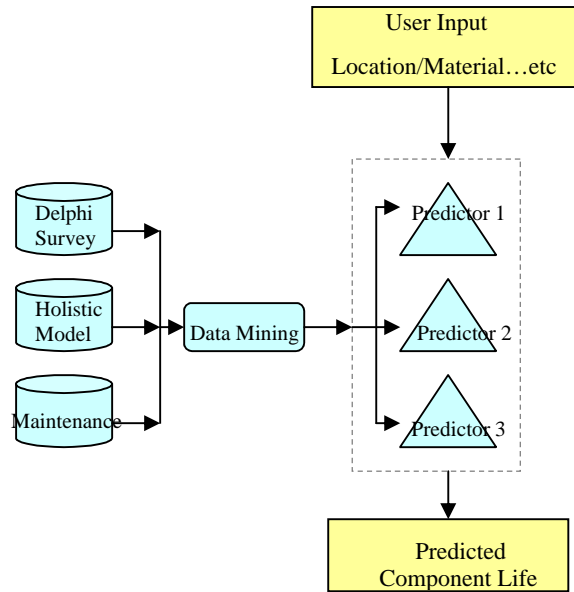


Figure 1: Overflow of Prediction Model

Data mining methods are applied to all three data sets to build three predictors first. After that, these three predictors can make predictions for user’s inputs. The final predicted life is either a multiple choice provided by three predictors or a value combined from the outputs of three predictors.

In order to get accurate predicted results, we have applied various data mining methods including Naïve Bayes, K Nearest Neighbors (KNN), Decision Tree (DT), Neural Network (NN), Support Vector Machine (SVM) and M5 Model Trees on these data sets. 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, Piatetsky-Shapiro, & Smyth, 1995a). 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 target.

Support vector machine (Vapnik, 1995) is relatively new method. It can solve the problem of efficient learning from a limited training set. M5 Model trees (Quinlan, 1992) is an effective learning method for predicting real values. Model trees, like regression trees, are efficient for large datasets. However, model trees are generally much smaller

than regression trees and prove more accurate (Quinlan, 1992).

All those are traditional data mining methods. We also used bagging (Breiman, 1996) to improve the performance of these methods. Bagging generates multiple predictors and uses these to get an aggregated predictor, which has better performance.

4 Experimental Results and Discussion

Although there can be many performance measures for a predictor such as the training time and comprehensibility, the most important measure of performance is the prediction accuracy in real-world predictive modelling problems (Zhang, Eddy Patuwo, & Y. Hu, 1998). For classification problem, prediction accuracy is defined as the number of correctly classified instances divided by total number of instances. For regression problem, correlation coefficient is often used to evaluate the performance. Correlation coefficient measures the statistical correlation between the predicted and actual values.

Data set	Prediction Accuracy	
	Naive Bayes	Decision Tree
Delphi Survey	30.0587%	36.217%
Holistic Model	89.744%	90.125%
Colorbond	94.728%	96.548%
Maintenance for Galvanized	93.138%	94.603%
Maintenance for Zincalume	91.904%	93.215%

Table 2: Prediction accuracy of Naïve Bayes & DT

In our data sets, all targets are continuous values. However, Naïve Bayes and Decision Tree implemented in Weka can only work for classification problem. Therefore, before using these two methods, all numeric attributes are discretized to nominal type. The average accuracy over 10-CV of these algorithms on these data sets is reported in Table 2.

KNN, NN, SVM and M5 implemented in Weka can work for regression problem. The average correlation coefficients over 10-CV of these algorithms on these data sets are reported in Table 3.

The results in Table 2 show that for Naive Bayes and Decision Tree, prediction accuracy is around 90% except Delphi Survey. Both Naive Bayes and Decision Tree are not good for Delphi Survey (only 30.0587% and 36.217% prediction accuracy that means more than half cases are not classified correctly). The highest accuracy is for Colorbond (94.728% from Naïve Bayes and 96.548% from DT). Decision Tree is a very 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

classification problem by discretizing continuous values to categorical values proved not suitable on our datasets, especially for Delphi Survey.

Data set	Correlation coefficient (cc)			
	KNN	NN	SVM	M5
Delphi Survey	0.797	0.9299	0.928	0.9333
Holistic Model	0.9960	0.979	0.8412	0.9892
Colorbond	0.9962	1	0.9999	1
Maintenance for Galvanized	0.9915	0.9994	0.9737	0.9883
Maintenance for Zincalume	0.9886	0.999	0.9889	0.9971

Table 3: Correlation coefficient of KNN, NN, SVM & M5

Table 4 shows the number of classes after discretizing the target attribute. We can find that the numbers of classes for all data sets are almost same while the number of cases varies from 683 to 9640. There are 10 classes while only 683 cases in Delphi Survey. Therefore, it may be the truth that decision tree is prone to errors in classification problems with many classes and relatively small training set.

Data Set	No. of cases	No. of target classes	No. of input attributes	Numerical attributes (%)	Categorical attributes (%)
Delphi Survey	683	10	7	0%	100%
Holistic Model	9640	10	6	50%	50%
Colorbond	4780	10	13	76.92%	23.08%
Maintenance for Galvanized	1297	10	12	91.67%	8.33%
Maintenance for Zincalume	1297	9	6	83.33%	16.67%

Table 4: Details of Data Sets

The results in Table 3 show that for KNN, NN, SVM and M5, very good results are achieved. Most of Correlation

coefficients (cc) are above 0.95. The lowest cc is 0.797 (KNN for Delphi Survey) and the highest is 1 (NN and M5 for Colorbond). NN works very well for all data sets, getting very high cc for all data sets. This result proves that NN is very efficient for handling numerical values and well-suited for predicting numerical target because most of attributes in our data sets are numerical values (The last two columns of Table 4 show the percentage of numerical and categorical attributes. We can find that almost all data sets have more than 50% numerical attributes). M5 is learned efficiently as NN. Especially, it is better for Delphi Survey and Holistic Model than NN.

The results from SVM are similar to NN, but reduced more for Holistic Model. The results from KNN are also similar to NN, even better for Holistic Model. But KNN got the worst result for Delphi Survey. This may prove that KNN is quite effective if the training set is large. Because there are 9640 cases in Holistic Model, 4780 cases in Colorbond, 1297 cases in maintenance while only 683 cases in Delphi Survey.

From the view of each data set, Colorbond gets the best result. The cc from all methods for Colorbond is very high (The highest reaches 1 while the lowest is also 0.9962).

The results for 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 like KNN, NN, SVM and M5 are better than those that have to discretize continuous values like Naïve Bayes and DT.

However, the interesting fact is that no one method is always best for all three data sets. M5 is the best method for Delphi Survey (cc is 0.9333), KNN is the best method for Holistic Model (cc is 0.9960), NN and M5 are the best methods for Colorbond (cc is 1) and NN is the best method for Maintenance database (cc is 0.999).

In next step, we experiment bagging (Breiman, 1996) to improve the prediction performance. Further experiments were performed on the best method for each data set. Results are shown in Table 5.

Data set	Correlation coefficient	
	M5 / KNN / NN	Bagged M5 / KNN / NN
Delphi Survey	0.9333	0.9454
Holistic Model	0.9960	0.9967
Colorbond	1	1
Maintenance for Galvanized	0.9994	0.9997
Maintenance for Zincalume	0.999	0.9995

Table 5: Results from bagging

From the results in Table 5, we find that bigger correlation coefficient can be obtained using bagging for M5, KNN and NN. It indicates that bagging is more accurate than the individual predictors.

So far, we have got five best models for our data sets. In order to see if the predicted service lives from different data sets are consistent, we choose some test cases as input data to produce predicted service lives from those models. Some examples of test cases are as follows:

-
- 1 | Windsor State school | Roof | Zincalume | Maintenance: Yes | Not Marine
 - 2 | Bald Hills State School | Roof | Zincalume | Maintenance: Yes | Marine
 - 3 | Beenleigh State School | Roof | Galvanized Steel | Maintenance: No | Not Marine
 - 4 | Allora State School | Gutters | Galvanized Steel | Maintenance: Yes | Not Marine
 - 5 | Calliope State School | Gutters | Colorbond | Maintenance: No | Not Marine
-

These test cases are in different environments (Marine or Not Marine), using different materials (Zincalume, Galvanized Steel or Colorbond) for different components (Roof or Gutters) with or without maintenance. They are selected in order to verify the predicted service lives under different conditions. The predicted service lives from different data sets for these test cases are shown in Table 6.

Because Holistic Model is only for Gutters of Galvanized Steel and Zincalume, Colorbond data set is only for Gutters of Colorbond, Maintenance data set is only for Roof of Galvanized Steel and Zincalume and Delphi is for a range of components and different materials, we compare the results of case 1, 2, 3 from Delphi and Maintenance and the results of case 4, 5 from Delphi, Holistic and Colorbond. From the above results, we found that for the first test case we got 51.877 from Delphi while only 29.928 from Maintenance. Similar contradiction happened to the fifth test case (36.64 from Delphi and 68.786 from Colorbond). For case 2,3,4, almost consistent results are achieved.

ID	Delphi	Mainten-ance	Holistic	Color-bond
1	51.877	29.928	N/A	N/A
2	27.185	30.449	N/A	N/A
3	35.929	26.338	N/A	N/A
4	33.151	N/A	30.951	N/A
5	36.64	N/A	N/A	68.786

Table 6: Predicted Service Life (years) for test cases

4.1 Existing Problems

Although no one method is always best for all three data sets, we can build independent model using the most suitable method for each data set. Sometimes a conflict exists among predicted values from three predictors for a given situation. One example is the first and the fifth test cases as shown in Table 6. There are twofold reasons for these contradictions (1) an inconsistency exists among three data sets, and (2) there exists an error during the mining process. If the first one is the case, the

inconsistencies need to be fixed with an expert opinion. The problem also arises how to choose the most appropriate answer for a given situation in case of inconsistencies.

The expert (or knowledgeable) user will have some prior knowledge to indicate the right choice. However, a naive user will not be able to make a decision depending on the result of the system.

The ideal way is to do some post-processing for the predicted result before presenting it to users. The post-processing should eliminate the conflict and select a best answer for users from multiple choices provided by multiple predictors.

Moreover, as mentioned in the data cleaning section, the data sets contains few missing values. The predictors are built based on data sets with few missing values. The user, however, may not be able to provide all inputs to get a precise answer from the use of data mining system. A method to deal with incomplete and vague queries is also required.

4.2 Possible Solutions

A knowledge base will be built in this solution. This knowledge base is a set of rules which are extracted from three predictors built already. They should identify the service life of a component using a material in a location. The framework of this solution is shown in Figure 2.

When the user queries the framework, the knowledge base is first consulted to search for matching between existing rules and user inputs. If user inputs are matched to a rule, we produce the result directly from the rule. If we can not find a matching rule, new data should be input into predictors to produce a result. Before doing this, user inputs should be pre-processed first for missing values.

Although some data mining algorithms can handle missing values automatically, for example, they replace missing values using most frequent value or average value; the ways they are using usually are not suitable for our case.

Case-based reasoning (Maher, Balachandran, & Zhang, 1995) is chosen to deal with the missing values of user inputs in this solution because the values are very close for similar cases. For example, if user only provides location and material, we can get mass loss of this material from other case using the same material and get salt deposition from other case in close location. After that, user inputs are fed into the predictors to produce the results.

To deal with the conflictions in the results of the three independent predictors and with the rules in knowledge base, post-processing of results is conducted. First we check for the consistency of the results. If they are consistent, we compare them with rules in knowledge base to see if the results are reasonable. For example, a roof in a severe marine location will not last longer than one in benign environment, and stainless steel should last longer than galvanized steel etc. We check the results to see if they match such rules. If they are not logical, adjust the results according to knowledge base. Otherwise, output the results.

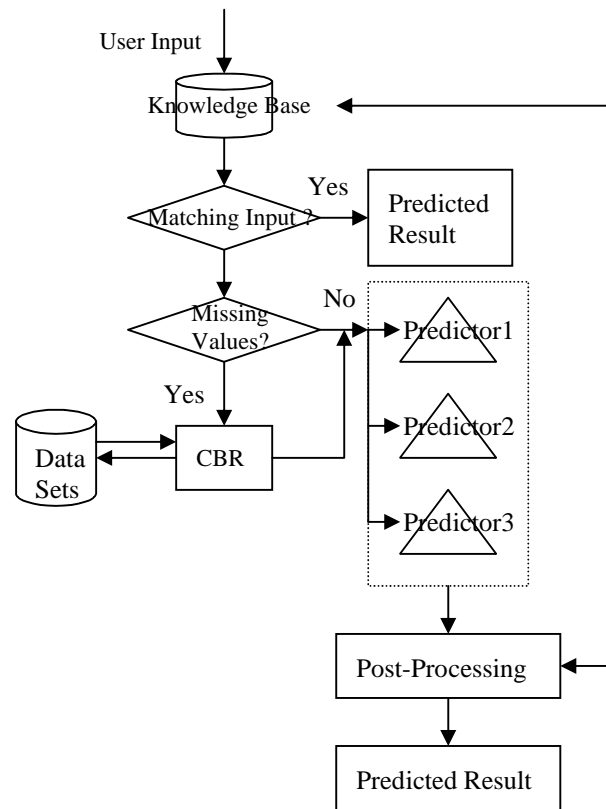


Figure 2: Framework of Solution

If the three predictors' results are inconsistent, the most reasonable (closest) result according to knowledge base is selected. In other words, the predictors' results compared with rules in knowledge base and the illogical results are deleted.

Finally, the result of each new case will be saved as a new rule in the knowledge base for later use.

The key of this solution is the construction of knowledge base. Experience and knowledge of domain experts can guide to construct the knowledge base. The cases covered by rules in knowledge base should be as many as possible. As a result, this solution can be human cooperated mining (Cao & Zhang, 2006).

5 Conclusions and Future work

Lifetime prediction of metallic components is significant in civil engineering. This paper has demonstrated that it is possible to apply data mining methods to solve this problem. We compare a number of data mining methods on the data sets provided by our industry partners and analyse what kind of methods are suitable for what kind of data.

Firstly, traditional data mining methods like Naïve Bayes, Decision Tree, Neural Network, SVM and M5 etc were applied to build a number of independent predictors for each data set. The results indicate that the best method for predicting the service life depends on the data set used to train the model. Moreover, those methods which can deal with continuous values directly like KNN, NN and M5 are better than those that have to discretize continuous values like Naïve Bayes and Decision Tree.

Further experiments for improving the performance were done on those best methods for each data set. We found the improvement to KNN, NN and M5 using bagging was obvious. We also analysed the predicted service life from each data set using certain test cases. The testing shows that in some situations, inconsistent predicted results may be presented by the data mining systems due to using three different data sets (information) for a same test case.

To solve this problem, we propose a possible solution. The key of this solution is the construction of knowledge base which contains a set of rules. When we evaluated the data mining methods, we focused on the prediction accuracy. However, if we need to extract rules from those models for building a knowledge base, we should consider the comprehensibility of those models. For example, we have found Neural Network is almost the best method for our data sets. However, it is very difficult to extract rules from Neural Network model. Moreover, what kind of rules shall we really need? It should be a case-based rule (eg. Contain many attributes like location, material etc and a service life) or if-else rules? Those questions should be answered by future research.

In summary, this study is a preliminary work. The aim of this paper was to explore various data mining methods to predict lifetime of metallic components and find a best one. Future research needs to prove the feasibility and effectiveness of the possible solution described above and develop a real lifetime prediction tool.

6 Acknowledgement

This work is partly supported by the CRC-CI project 2005-003-B funding. We would like to thank all the team members from CSIRO, Queensland Department of Public Works (QDPW) and Queensland Department of Main Roads (QDMR).

7 References

- Breiman, L. (1996). Bagging Predictors. *Machine Learning*, 24(2), 123-140.
- Cao, L., & Zhang, C. (2006). Domain-Driven Actionable Knowledge Discovery in the Real World. In *Lecture Notes in Computer Science* (3918 ed., pp. 821-830).
- Cole, I., Ball, M., Carse, A., Chan, W. Y., Corrigan, P., Ganther, W., et al. (2006). *Methods for the Service Life Estimation of Metal Building Products as Applied to Selected Facilities and Bridges*.
- Cole, I., Ball, M., Carse, A., Chan, W. Y., Corrigan, P., Ganther, W., et al. (March 2005). *Case-Based Reasoning in Construction and Infrastructure Projects - Final Report* (No. 2002-059-B).
- Cole, I. S., Furman, S. A., & Ganther, W. D. (2001). A holistic model of atmospheric corrosion. *Elec Soc S*, 2001(22), 722-732.
- Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1995a). Bayesian Networks for Knowledge Discovery. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth & R. Uthurusamy (Eds.), *Advances in Knowledge Discovery and Data Mining* (pp. 273 - 305). Menlo Park: AAAI Press.
- Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1995b). From Data Mining to Knowledge Discovery: An Overview. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth & R. Uthurusamy (Eds.), *Advances in Knowledge Discovery and Data Mining* (pp. 1 - 34). Menlo Park: AAAI Press.
- Furuta, H., Deguchi, T., & Kushida, M. (1995). *Neural network analysis of structural damage due to corrosion*. Paper presented at the 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.
- Kessler, W., Kessler, R. W., Kraus, M., Kubler, R., & Weinberger, K. (1994). *Improved prediction of the corrosion behaviour of car body steel using a Kohonen self organising map*. Paper presented at the Advances in Neural Networks for Control and Systems, IEE Colloquium on.
- Maher, M. L., Balachandran, M. B., & Zhang, D. M. (1995). *Case-Based Reasoning in Design*. Mahwah, NJ, USA: Lawrence Erlbaum Associates, Inc.
- Melhem, H. G., & Cheng, Y. (2003). Prediction of remaining service life of bridge decks using machine learning. *Journal of Computing in Civil Engineering*, 17(1), 1-9.
- Melhem, H. G., Cheng, Y., Kossler, D., & Scherschligt, D. (2003). Wrapper Methods for Inductive Learning: Example Application to Bridge Decks. *Journal of Computing in Civil Engineering*, 17(1), 46-57.
- Morcous, G., Rivard, H., & Hanna, A. M. (2002). Case-Based Reasoning System for Modeling Infrastructure Deterioration. *Journal of Computing in Civil Engineering*, 16(2), 104-114.
- Olafsson, S. (2006). Introduction to operations research and data mining. *Computers & Operations Research* Part Special Issue: *Operations Research and Data Mining*, 33(11), 3067-3069.
- Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1), 81-106.
- Quinlan, J. R. (1992). *Learning with Continuous Classes*. Paper presented at the 5th Australian Joint Conference on Artificial Intelligence.
- Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. New York: Springer-Verlag.
- Zhang, G., Eddy Patuwo, B., & Y. Hu, M. (1998). Forecasting with artificial neural networks:: The state of the art. *International Journal of Forecasting*, 14(1), 35-62.