An Interactive Predictive Data Mining System for Informed Decision

Esther Ge and Richi Nayak

Faculty of Information Technology
Queensland University of Technology,
Brisbane, Australia
t.ge@student.qut.edu.au, r.nayak@qut.edu.au

Abstract. There exists a need to utilize the predictive data mining models for querying to obtain the predicted outcome based on user provided inputs in its real use. This demo illustrates a real-world situation in which the trained predictive data mining system is being deployed and now users can interact with the model for informed decision.

Keywords: Data Mining, predictive model, interactive model

1 Introduction

The training of a predictive data mining model and then understanding rules and patterns inferred by the mining model should not be the end of the prediction task. The data mining model should allow the user to query the system for future cases. The created data mining model needs to be deployed in practice as a user-driven prediction system so that the data mining system can be used for querying to obtain the predicted outcome based on user provided inputs. In some real-world applications, the training set contains relatively complete attribute information while the unseen cases (user queries) do contain many missing attribute values. Consider a predictive data mining model that is built to predict the “Service Life” of the building components based on the input attributes such as “Location”, “Component”, “Material”, “Salt Deposition”, and “Mass Loss”. Suppose a builder (a typical user of the predictive model or tool or system) wants to know the service life of a “Gutter” with “Galvanized Steel” at a particular location. However, the user does not explicitly 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 will not be as accurate as tested in the evaluation phase of the predictive model, especially when the missing attributes play key roles in predicting the outcome. 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 needs to be solved is how to select the appropriate attributes to build the model for a real-world situation when the users can not provide all the inputs for querying to the system. In other words, how to deal with the missing attribute values in user queries (unseen cases). We developed an interactive data mining model for predicting the service life of metallic components in buildings which allows the user to input the queries based on their limited knowledge, while maintaining the accuracy of the predicted outcome.

2 An Interactive Predictive Data Mining System

The proposed interactive predictive data mining system consists of nine phases structured as sequences of predefined steps. The system includes the standard data pre-processing, data analysis and result post-processing phases. Additionally, it includes the phase of Query Based Feature Selection (QBFS) separated from the data pre-processing step. The QBFS phase 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 predictive data mining system is being used in practice by users. An external domain knowledge base is involved in results post-processing and missing inputs pre-processing in the Use of Model phase.

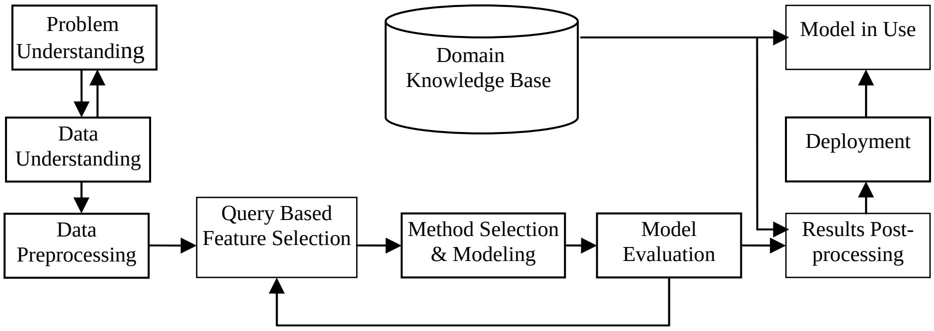

Fig. 1. An Interactive Predictive Data Mining System

In order to select the appropriate attributes to build the predictive model, the Query Based Feature Selection (QBFS) algorithm [1] is applied to the datasets. The QBFS algorithm allows selecting the attributes according to the interest of a user/domain expert. This algorithm first divides the attributes available for training into three categories according to their accessibility for querying. The first group contains attributes that are easily accessible to users. The second group contains attributes that are difficult for users to access. In other words, users cannot directly provide the values of these attributes while querying, but, it is still feasible to get those values via some indirect sources. The third group contains attributes that users can never provide values for querying. The attributes belonging to third groups are not included in model building. The QBFS algorithm selects a minimum subset of features which can be provided by users or obtained from domain knowledge based on the groups 1 and 2, while maintaining the acceptable accuracy of the model as well. The QBFS has been proven [1] to provide good generalization accuracy.

The datasets used in this system are from real life including four different sources of service life information [1]. The features selected by the QBFS include some which can be provided by users such as “longitude”, “latitude”, “component”, “material” and others which can not be provided by users such as “Salt Deposition” and “Rainfall”. Therefore, the domain knowledge is used to get these two attribute values in user queries. The knowledge is represented as items in the database. For example, an item for Salt Deposition knowledge is (longitude, latitude, Salt Deposition). Once the user inputs the location (longitude, latitude), the SQL query language is used to search the knowledge base to find the same location or the nearest location and accordingly the values of Salt Deposition and the Rainfall are obtained. As the predictors are built from different datasets, the predicted result may not be consistent. The domain knowledge base also includes some generalised rules to post-process the inconsistent results. One example of the generalised rules is (Component, Environment, Material, Min years, Max years). These generalised rules give a reasonable range of service life for matched user inputs.

3 An Example of Prediction of Service Life using the System

As shown in the user interface (Figure 2), the location, component and material are compulsory inputs for querying to the system. Based on these three inputs, different predictors according to each distinct dataset will be used to do the prediction. Here is an example for using the system. The user query is to predict the service life of gutters (as component) with galvanized steel (as material) in location (151, -28). The location inputs can also be directly selected from the geo-spatial database using GIS. The Holistic-I and Delphi models can predict the service life based on these inputs. These models need more inputs so user is prompted to get those values. After the user inputs the gutter position, maintenance and environment, 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 to be 14.5 years from the Holistic-I predictor. A similar process is done for the Delphi predictor and the predicted service life is predicted as 14.4 years.

Sometimes the results of predictors 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 of the Delphi model is 51.8 years while of the Holistic-III predictor is only 29.9 years. In such a case, the system consults the “domain knowledge base” that includes the generalized rule set. For example, in this case, the rule that matches is “The range of service life for roofs with Zincalume in benign environment is greater than 50 years”. Based on this rule, the system finally displays the service life to be 51.8 and discards the life value of 29.9 based on Holistics-III predictor.

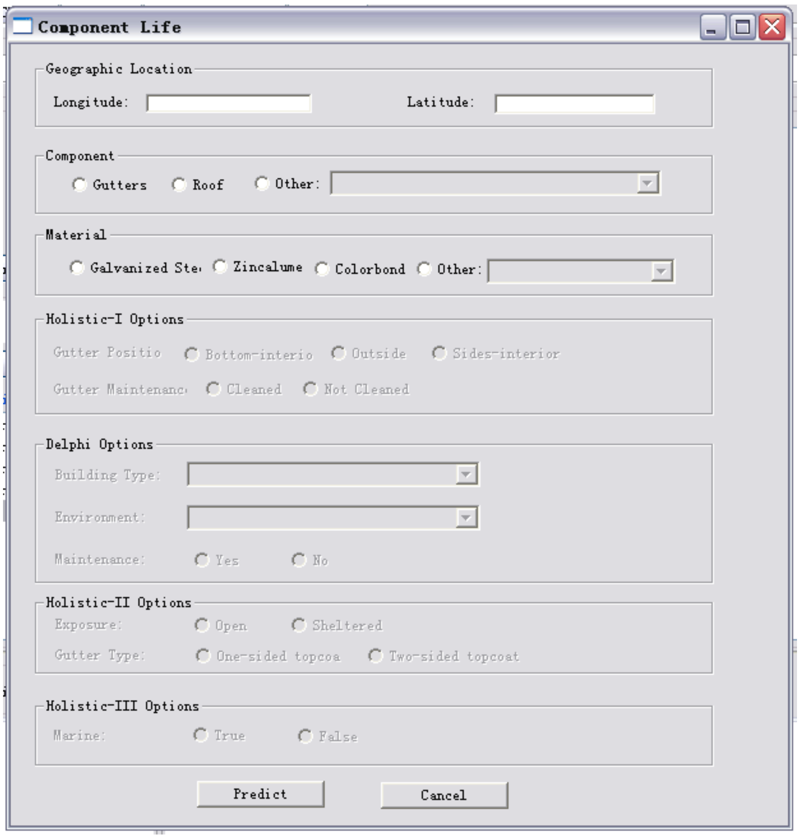
Component Life
Geographic Location
Longitude:
Latitude:
Component
○ Gutters ○ Roof ○ Other:
Material
○ Galvanized Ste ○ Zincalume ○ Colorbond ○ Other:
Holistic-I Options
Gutter Positio ○ Bottom-interio ○ Outside ○ Sides-interior
Gutter Maintenanc ○ Cleaned ○ Not Cleaned
Delphi Options
Building Type:
Environment:
Maintenance: ○ Yes ○ No
Holistic-II Options
Exposure: ○ Open ○ Sheltered
Gutter Type: ○ One-sided topcoa ○ Two-sided topcoat
Holistic-III Options
Marine: ○ True ○ False
Predict
Cancel

Fig. 2. User Interface of the Interactive Predictive Data Mining System

5 Conclusions

This paper presents a real-world situation in which the learned predictive data mining model is deployed to a user-oriented prediction system. The developed system is easy to use for people with little expertise in data mining and domain non-experts. It provides accurate prediction where not all inputs are available for querying to the system with incorporation of query based feature selection algorithm.

References

[1] E. Ge, R. Nayak, Y. Xu, and Y. Li, "A User Driven Data Mining Process Model and Learning System," presented at the 13th International Conference on Database Systems for Advance Applications, New Delhi, India, 2008.