PREDICTING THE CAUSES OF INPUT ERRORS IN THE DATABASE

Kirthigaah Devi, Mohd Zakree Nazri

Research Center for Software Technology and Management,
Fakulti Teknologi and Sains Maklumat, Universiti Kebangsaan Malaysia
43600 UKM Bangi, Selangor Malaysia.

Email: gp04913@siswa.ukm.edu.my, zakree@ukm.edu.my

ABSTRACT

Today, the world is filled with data like Oxygen. The amount of data being harvested and eaten up is flourishing vigorously in the digital world. The growing exploitation of companies that leads to the generation of huge quantities of data which can bring remarkable information if it is analysed properly. Organizations tries to reduce manufacturing cost by reducing design errors. Among the designs is database design. Organizations may undergo for analysis of predicting error in data modelling to have better decisions. Thus, data analytics is being paid attention in many companies in recent times. For finding the concealed values from preventing errors, each company requires new schemes or strategies. Predictive analytics comprises of several statistical and analytical techniques for developing strategies for predictions. Therefore, predicting performance becomes vital when an essential quantity of highly sensitive data must be handled. Based on the perceived events, future probabilities and measures are predicted. With the aid of available data mining techniques, predictive analytics predicts the events in future and can make as recommendations. Decision Trees(J48) has been used to develop a predictive model of modelling error in reducing turnover through database analysis. The analysis of the results in this study has been based on experimental results. The results show that Decision Trees performed well in terms of accuracy of predicting error. It is envisaged that knowledge that is gained from this study will aid decision makers in manufacturing strategic planning.

Keyword: data mining, decision tree (J48), Manufacturing strategic planning, modelling expertise framework (MEF)

I. INTRODUCTION

Researchers seek to study various methods in reducing design errors, such as database design because of the high cost of software and the high cost of employment to maintain errors is estimated to be between tens of millions of dollars per year (C. Kop, 2009). Forty-five to sixty-five percent of all errors or errors are made during design (Dedrick, 2017). The cost of correcting errors in development is proportional to the amount of time and the amount of errors that exist in the process. Savings are expected to increase between one and two magnitudes when an error is successfully detected and corrected near the moment of its creation point, rather than at the stage of execution (Dedrick, 2017).

Therefore, it is important to remove errors at the beginning of the design phase, for example during concept design. Before we can eliminate such errors, it is constructive to identify and analyze them, so that engineers and analysts understand the errors that are accidentally generated by the designer and understand why they occur. This research examines the expertise of database modelers as well as their impact on errors made in conceptual database design. Due to the significant cost of errors, previous research has looked at the impact or impact of various factors that cause design errors

such as issues in needs analysis, representation selection and Base methods, issues in application domains, and designer backgrounds (Dutta, 2009).

Arguably, some types of errors will be harder to avoid, such as errors caused by changes in user requirements. The attention of this study is on the error that can be attributed to the lack of expertise of Base. The meaning of expertise in this report is the level of knowledge and achievement of skills in the formalism of the conceptual Base (i.e. the Entity Relationship Model) required to construct a scheme effectively. Studying the impact and impact of Base skills or expertise is important for a few reasons. Among them is to show the relationship or often also called as a predictor (sign) between the level of expertise Base and error. If it can be shown in an analytical, then steps can be taken to improve the quality.

This study focuses on a scenario involving a real-world enterprise with the scope of this study covering employees aged between 20 to 39 years who contributed to the high error in an international enterprise in Negeri Sembilan, Malaysia. This study uses a measurement scale based on the actual data of the enterprise that is difficult. Reliability and validity tests are used to ensure the reliability of the data.

It is necessary to understand the causative factors of the type of error to support the action to prevent it from recurring with the main goal of reducing the number of errors. Unlike some other factors (e.g., Base techniques), Base expertise factors are unpopular and rarely researched by previous researchers as a major cause of design errors (P. Suraweera, 2002). Since expertise is a factor that can be influenced or changed through training (P. Suraweera, 2002), when combined with a detailed breakdown of error classes, this research will be able to prove empirically, that different training properties are required for different levels of expertise. This motivates researchers to research further to clarify errors and find ways to predict them and reduce them from recurring.

MEF (Modeling Expertise Framework) allows us to predict the type of error in a conceptual Base when the level of expertise is known. Experiment that designed to evaluate the framework to validate its validity. The study also developed algorithms to validate conceptual test schemes, with a solution scheme based on a subtle error classification to identify errors and analyze violations appropriate to each level of Bloom's Taxonomic Error (RBT) expertise.

To keep this report simple but complete, this study presents a report on the relationship of interactions. A complete group of ER constructs has been reported many times and this report will discuss ER constructs in general.

II. LITERATURE REVIEW

A. Overview of Modeling Expertise Framework (MEF)

Conceptual databases are known to be an important part of database development. Few studies have sought to find the underlying cause of cognitive challenges or errors made during this stage. Using the Modeling Expertise Framework (MEF) which uses base expertise to predict errors based on a revised Bloom (RBT) taxonomy. The use of RBT is in providing a classification of cognitive processes that can be applied to knowledge activities such as conceptual bases. MEFs can be used to map conceptual base tasks to different levels of cognitive complexity and classify current levels of modeling expertise. Experimental exercises confirm error predictions. This provides an understanding of why beginners can handle entity classes and identify binary relationships easily but find other components.
The MEF consists of three stages:
i) Early exploration,
ii) model identification and exploration,
iii) dissemination (application of models to new data to generate predictions).

Stage 1: Early exploration

Usually this stage begins with the preparation of data consisting of selecting a subset of data, filtering to bring the correct data to a controlled range depending on the method under consideration.

Stage 2: Model Introduction and Exploration

This stage involves selecting the best model based on their predictive performance. This also involves model validation that requires model training on data and data analysis.

Stage 3: Dissemination

This is the last stage which involves selecting the model at the previous level as best as possible and applying it to the new data to generate a prediction of the result. The successful implementation of this technique requires a methodology built on best practices. This will be discussed further in the section on the tasks performed.

B. Application of Forecasting Analysis Through Data Mining

Customer Relationship Management (CRM) - Analytical CRM is one of the most used forecasting analysis applications at present. Forecasting analysis under this field is applied to customer data to pursue and achieve the CRM objectives specified for the organization. CRM uses this analysis in applications to increase sales, marketing, and campaign targets (Bidisha Lahkar Das, 2013) (Jagun, 2015). This not only affects business growth, but also makes business customers eccentric by expanding the foundation for customer satisfaction (Gupta, 2010).

Child Protection - Child abuse is a serious error and child protection is highly sought after in any country (Gupta, 2010). Several child welfare organizations have used predictive analysis to identify high-risk cases of child abuse (Bass, 2018). Predictive models help in identifying from medical records, cases that may be under child abuse criteria. This approach is referred to as “innovative” by the Commission for the Elimination of Child Abuse and Neglect of Death (CECANF) (R.M. Felder, 2015). Using predictive analysis, crimes related to child abuse have been identified at an early stage to prevent further harm (Adeyemo, A., & Orialo, 2010).

C. Prediction in Database

Forecast base is a process that uses data mining to predict outcomes. Each model consists of several predictors, which are variables that tend to affect future outcomes. (B. Boehm & V.R. Basili 2001) After data were collected for relevant predictors, statistical models were formulated. Goals The database has the most direct application with powerful forecasting techniques. Automatic prospective analysis offered by mobile data mining goes beyond special analysis of decision support systems. They search the database to find predictive information that may have been missed by beginners because it was beyond their expectations (B. Boehm & V.R. Basili 2001)

Predictions in research fulfill one of the basic desires of humanity, which is to see the future. This prediction continues after the hypothesis to produce what will happen in the future (S. Dreyfus 1980). Massad et al. (2005) put forward two components to make predictions; the first is the forecast and the second projection. Prediction is quantitative to predict what might happen and projection is an attempt to describe what will happen to a hypothesis.

III. **RESEARCH MODEL**

To achieve the purpose of this research, several key phases have been followed, these phases are further divided into several steps. In the first phase, theoretical study and literature review in which analytical and predictive techniques were investigated. This has been presented in chapter 2. The second phase is the application process which involves two main parts, the application of association rules and the decision tree using the WEKA analytical tool kit on the data set. The next sub-section describes this phase and highlights the key processes for producing a model. In the final phase, this study evaluates

the performance of this study model in terms of accuracy in forecasting. This is followed by the results of the experiments presented in chapter 4 and finally the summary and conclusions in chapter 5.

A. Appropriateness Identification Techniques Used to Produce a Proposed Model

This phase begins with identifying the most relevant work. This focuses on understanding data mining and its applications for predicting errors in Databases in participating Companies. This is achieved by studying the latest forecast analysis to identify the predictive factors and accuracy of the current approach that drives us to conduct data mining methods and techniques and to predict errors in the Database. This is mining association rules and decisions (Agrawal et al 1999, WEKA, 2011).

There are so many options, tasks, techniques, tools, formats, and approaches to data mining that industrial engineers find it very difficult to plan and execute a project. Although methodologies are already in place, they are designed for specific software packages. Most of these methodologies use traditional statistical approaches. It is not yet clear that this approach to data mining is sufficient to obtain the large amount of data required for 24 industrial engineering applications. Therefore, data mining methodologies to meet the specific needs of industrial engineering are essential. Such a methodology should assist industrial engineers in selecting appropriate data mining tools and implementing data mining projects from a system perspective. A description of the process in achieving the purpose of predicting errors is presented.

IV. METHODS: PARTICIPANTS AND DATA COLLECTION

Demographic data used in this research work has been used in previous studies (Long and Bakar, 2011, Mousavi et al, 2013) collected from the Selangor state public works department combined with data from participating companies and has been presented to this study by the author. The data contains 10000 fields with 12 attributes. However, the data set is not complete, so this data is cleaned, normalized, attributes are selected and reduced and finally compiled into a suitable form for the data mining process.

All other variables are accurately categorized to accommodate all available information as shown in the Table below. 8 attributes selected from the original data; extraction is done carefully to avoid incomplete records. Selected attributes include Year, month, date, Start Time Machine, area, Machine Name, Material ID, Machine and Flag. For this study, the data combined with the data obtained from the participating companies located in Negeri Sembilan makes the total attributes to 12 to compile a model that quality.

V. RESULTS & DISCUSSION

Such a classifier will achieve an accuracy of 99.9%. Thus, this study will never predict errors, but the accuracy of the model is still very high. Therefore, for some specific problems, models with low accuracy, for example at 30% may do a better job than the exact 99.9%. Moreover, accuracy

differences may occur either measured on representative samples or with samples that are not completely balanced.

Experiment 2 is to test the accuracy value of the data from experiment 1. The purpose is to determine the machine that is causing the error in the database. Once a conclusion is made, a table will be built where it contains a summary of the results of Experiments 1 and 2.

Take the machine damage prediction as an example. This assessment technique is about positive positives, false positives, true negatives, and false negatives. These parameters are test results with historical data and failure detection and prediction maintenance approach approaches. For example, accuracy is defined as given values, we can calculate the probability of an error in which we break out as follows:

Where tp is positive positive, tn is negative, fp is false positive and fn is false negative. In recent literature, the most widely used metrics in the detection and prediction of maintenance failure types are accuracy, precision, mean square error and absolute mean percentage of errors. This assertion also applies to the detection of failure types and maintenance application predictions

J48 algorithm was chosen as the best algorithm for this research because it takes less than 0.19 seconds to model. According to (Mohamed et al, 2012) J48 is a good decision-making tree produce good results. This algorithm produces a simple tree structure with high accuracy in terms of classification rate. Pruning method or Pruning method is used to reduce the complexity of the tree structure without lowering the classification accuracy. While drawing conclusions about tree decision tree algorithms, this study builds a predictive model that varies confidence factors 0.1 and 0.5 to produce the best possible algorithm. Test of the experiment will show as per below.

Experiment 1

Algorithm	TP Rate	FP Rate	Ketepatan	Recall	F-measure	ROC Area	Waktu	C/C	MSE (mean square error)
J48	**1**	**0**	**1**	**1**	**1**	**1**	**0.19**	**99.9596**	**0.0201**
Random Tree	0.996	0.01	0.996	0.996	0.996	0.992	0.5	99.4942	0.0711
Rep Tree	1	0	1	1	1	1	0.54	99.9595	0.0201
Random Forest	1	0	1	1	1	1	4.86	99.9798	0.0296
Lad Tree	0.997	0.001	0.997	0.997	0.997	0.998	4	99.97269	0.0747
Stump Tree	0.882	0.426	0.897	0.882	0.864	0.722	0.18	88.1954	0.3204
JRip Rules	1	0	1	1	1	1	3.23	99.9595	0.02
Decision Table	0.997	0.001	0.997	0.997	0.997	1	0.63	99.7471	0.0447

Experiment 2

Modal Mesin	% SPLIT	F-M	ROC	WAKTU	C/C	MSE	PRE
ICP395158AM2	50-50	0.999	0.999	0.17	99.9393	0.0246	1
ICP395168AM3	60-40	0.999	0.999	0.16	99.9494	0.0225	1
ICP395768AM1	70-30	0.999	0.999	0.18	99.9326	0.026	0.999
ICP395769AM1	80-20	0.999	0.999	0.8	99.9494	0.0223	1
ICP395990AM2	90-10	0.999	0.999	0.18	99.8989	0.0315	1

B. Techniques Used to Produce a Proposed Model

Researchers developed a model based on data mining technology, this technique is used to capture the art and science of predictive analysis. The concept of decision tree (Decision Tree) was used to produce the proposed model and implemented in the Waikato Environment for Knowledge Analysis (WEKA). Several tests are performed to produce methods and reasons that produce the best prediction accuracy.

VI. CONCLUSION

To achieve the goals and objectives, the researcher laid the theoretical basis for the research work through a large enough literature review to find out what has been done in the field, problems and solutions by most researchers; other parameters that have been validated in previous works are also considered. Native demographic data related to performance through database analysis were obtained from participating companies. Researchers have developed a model based on data mining technology, this technique will be used to conquer the art and science of predictive analysis. Decision Tree (J48) and the concept of split percentage are used to produce the model proposed and implemented in the Waikato Environment for Knowledge Analysis (WEKA). There are several tests performed to produce the correct algorithm where it produces the best prediction accuracy as well.
However, research paves the way for future research to use additional important inputs, data sets and larger attributes. As can be seen in the work of this project, the model is unable to spend all the existing rules to adapt to all current world phenomena and as a result, systems for updated rules are used when needed and can be put together for more accurate predictions in specific areas where errors more likely.

ACKNOWLEDGEMENT

The authors would like to thanks, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia by giving the authors an opportunity to conduct this research.

REFERENCE

D e d r i c k, Information Technology and Economic Performance: A Critical Review of the Empirical Evidence. – ACM Computing Surveys, Vol. 35, March 2003, No 1, 1-28.
D u t t a. The Global Information Technology Report 2008-2009. World Economic Forum, Accessed 15 May, 2009.

H a n a f i z a d e h, M. R., A. S a g h a e i, P. H a n a f i z a d e h. An Index for Cross-Country Analysis of ICT Infrastructure and Access. – Telecommunications Policy, Vol. 33, 2009,385-405.

H e s h m a t i, A., W. Y a n g. Contribution of ICT to the Chinese Economic Growth. The RATIO Institute and Techno-Economics and Policy Program College of Engineering, Seoul National University, 2006.

H u v e n e e r s, C. ICT Diffusion and Firm-Level Performance. Case Studies for Belgium,Planning & Working Papers, 2003.

K r a k a r, Z., S. T o m i ć R o t i m. Assessment of Croatia's Readiness for Using the ICT Potentials. – In: Proceedings of 20th Central European Conference on Information and Intelligent Systems, Varaždin, 2009.
Suraweera, Vanichsetakul. Tree-Structured Classification via Generalized Discriminant Analysis (With Discussion). – Journal of the American Statistical Association, Vol. 83, 2012, 715-728.

Jagun, S. Wiedenbeck, J. Scholtz, Mental representations of programs by novices and experts, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Amsterdam, The Netherlands, 2015, pp. 74–79.

Bidisha Lahkar Das, How novices design business processes, Information Systems 37 (6) (2013) 557–573.

P. Suraweera KERMIT: a constraint-based tutor for database modeling, in: Proceedings of the Sixth International Conference on Intelligent Tutoring Systems (ITS 2002), Biarritz, France and San Sebastian, Spain, 2002, pp. 377–387.

D.L. Moody, G.G. Shanks, Improving the quality of data models: empirical validation of a quality management framework, Information Systems 28 (2003) 619–650.

Y. Wand, An ontological analysis of the relationship construct in conceptual modeling, ACM Transactions on Database Systems (TODS) 24 (1999) 494–528.

] R.M. Fuller, B.A. Schafer, The effects of data model representation method on task performance, Information & Management 47 (2010) 208–218.

C. Kop, Towards a combination of three representation techniques for conceptual data modeling, in: Advances in Databases, Knowledge, and Data Applications, 2009. DBKDA ‘09. First International Conference on, 2009, pp. 95–100.

RTI, The economic impacts of inadequate infrastructure for software testing, National Institute of Standards and Technology (NIST) Planning Report 02-3, May 2002.

G. Rush., A fast way to define system requirements, Computerworld (1985) 11–16. [6] M.I. Aguirre-Urreta, G.M. Marakas, Comparing conceptual modeling techniques: a critical review of the EER vs. OO empirical literature, SIGMIS Database 39 (2008) 9–32.

B. Boehm, V.R. Basili, Software defect reduction top 10 list, IEEE Computer 34 (2001) 135–137.

H. Topi, V. Ramesh, Toward an extended framework for human factors research on data modeling, in: K. Siau (Ed.), Advanced Topics in Database Research, vol. 3, ed: Idea Group, 2004, pp. 188–217.

S. Dreyfus, A Five-Stage Model of the Mental Activities Involved in Directed Skill Acquisition, 1980.

D. Batra, J. Davis, A study of conceptual data modeling in database design: similarities and differences between expert and novice designers, in: Proceedings of the Tenth International Conference on Information Systems (ICIS-89), Boston, MA, 1989, pp. 91–100.