# Predicting the Quality of Object-Oriented Multidimensional (OOMD) Model of Data Warehouse using Decision Tree Technique

Kunwar Babar Ali, Anjana Gosain
University School of Information Technolgy, Guru Gobind Singh Inderprastha University, Dwarka.16-C, New Delhi

**Abstract**— Data warehouse is a powerful tool which makes decision faster and reliable in organizations where 'information' is the main asset of primary concern. It is necessary to assure the quality of the data warehouse information. Information quality depends on multidimensional model's quality of data warehouse. In the last few years' different authors have suggested several metrics to access the quality of multidimensional models of data warehouse. Validation of these metrics with traditional statistic techniques like correlation analysis, univariate and multivariate regression technique etc are not capable to derive reasonable results. The pure amount of data and the problem context of quality prediction, demand sophisticated analysis provided by machine learning techniques. In this work we focus on quality (in terms of understandability) prediction using decision tree learner on the metrics provided by Pittani et al. The main reason for choosing decision tree learners, instead of for example neural nets, was the goal of finding underlying rules which can be easily interpreted by humans. Our result shows that the proposed decision tree based technique is capable to predict the output with considerable accuracy.

**Index Terms**— Data warehouse, Information Quality, Multidimensional Model, Metrics, Quality Prediction, Machine Learning Techniques, Decision Tree

## 1 INTRODUCTION

Data warehouse is the backbone of most decision support system; it provides historical information to the decision makers. A be short of quality in the data warehouse can have disastrous consequences from both technical and organizational points of view: loss of clients, important financial losses or discontent amongst employees [2]

One way to assure quality of data warehouse is to guarantee the quality of the models (conceptual, logical and physical) used in to design of data warehouse. Quality of data warehouse multidimensional model has a great influence on the overall data warehouse quality and hence, in turn on information quality [5][1]. Few researchers have suggested quality factors for multidimensional model of data warehouse like maintainability, simplicity, completeness, consistency, minimality, etc [1][3][6]. Maintainability includes Understandability (a model's ability to be easily understood). Few researchers have proposed metrics for object-oriented multidimensional model of data warehouse to assess the understandability. In order to prove their practical utility, these metrics have been theoretically and empirically validated. Most of the empirical validation is done using statistical techniques like correlation analysis, univariate and multivariate regression analysis [2][5][6][11]. These techniques are unable to model the non linear relationship between the metrics and object-oriented multidimensional (OOMD) model quality.

In this paper we predict the understandability (an attribute of quality) of OOMD model of data warehouse using decision tree. We present an approach that applies decision tree learners on different metrics values of OOMD model. In literature a number of famous machine learning algorithms have been proposed and today many computing applications are using these algorithms. These algorithms have been engineered over the last several decades, and many are open-source and available for public usage. The decision tree is a popular utility for implementing decision based classification and adaptive learning over a training set. A decision tree is a decision-modeling tool that graphically displays the classification process of a given input for given output class labels. [12][13].

A decision tree is created by a process known as splitting on the value of attributes means testing the value of an attribute such as Outlook and then creating the branch for each of its possible values [12].Compared with statistical and neural/connectionist approaches to classification of data, decision trees (DT) have several advantages. First of all, there is no presumption of data distribution in DT. Second, since DT adopts a divide-and-conquer strategy, it is fast in training and execution. Most importantly, the resulting classification rules are presented in a tree form. [12][13][14][17].

We used WEKA tool to perform our experiments. Weka is an open-source Java application created by the University of Waikato in New Zealand. This software pack features an interface through which many of the aforementioned algorithms (including decision trees) can be utilized on preformatted data sets. Using this interface, several test-domains were experimented with to gain insight on the effectiveness of different methods of pruning an algorithmically induced decision tree.

In this paper we address the issue of predicting understandability of object-oriented multidimensional model of data warehouse. We present an approach that applies decision tree learners to the metrics values of 11 objects –oriented multidimensional model of data warehouse.

## 2 RELEATED WORK

In this work we concentrate on predicting the quality of OOMD model. The related work is primarily divided into two parts. The first part deals with the multidimensional modelling. Second part focus on the work associated with decision tree technique to build a quality prediction models.

In past various multidimensional data model have been proposed .Some of them fall into Logical level, some fall into frame model and some fall into conceptual model[02].Object-oriented multidimensional model comes under the conceptual model. Conceptual model provides a set of graphical notations that improve their use and reading. Some conceptual models are The Dimensional-Fact (DF) Model by Golfarelli et al. The Multidimensional ER (M/ER) Model by Sapia et al, The starER Model by Tryfona et al, the Model proposed by Hu¨seman et al., and The Yet Another Multidimensional Model (YAM2) by Abello´ et al.[02] But regrettably none of them has been accepted as a standard for design and maintain high quality data warehouse[02] .Lately a new model is projected that is capable to design efficient data warehouse that is Object-Oriented Multidimensional (OOMD) model [02].

N.Prat [02] has proposed metrics for multidimensional schemas analyzability and simplicity.Unfortunately none of these metrics proposed have been theoretically as well as empirically validated and therefore, have not proven their practical utility[07] .A proposed metrics has no value if it is not empirically validated then its practical value is zero..

Abraham at el [17] used decision tree method to find out the defect densities in source code files of various releases of Mozilla open source web browser project. The evolution data includes different source code, modification, and defect measures computed from seven recent Mozilla releases.

Sam Drazin and Matt Montag [12] used decision tree method to compare different pruning methods. Y.W [13] used decision tree algorithm in Materialized Projection and Selection View. In this he used a set of implementation steps for the data warehouse decision makers to improve the response time of queries. The study concludes that both attributes and tuples are important factors to be considered to improve the response time of a query. The adoption of data mining techniques in the physical design of data warehouses has been shown to be useful in practice.

Tibor Gyimothy at el [14] used decision tree learners to conduct Empirical Validation of object-oriented metrics on open source software for fault prediction. In this they also compared the metrics of several versions of Mozilla to see how the predicted fault proneness of the software system changed during its development cycle.

## 3 EXPERIMENTAL SETUP

Our experimental setup includes hypothesis and metrics of OOMD model and their values.

Using decision tree technique we conduct a series of experiments addressing the following hypotheses:

Hyp 1: We can predict the understandability of OOMD model of data warehouse.

Hyp 2: Larger metrics values produce poor understandability.

Hyp 3: We can identify the factors leading to high/poor understandability.

### 3.1 Metrics for Object -Oriented Multidimensional (OOMD) model:

Various authors stated that the complexity of a model may be calculated by the no. and variety of elements and no. and variety of relationship between them. Taking into account this statement we must take three different levels: class, star and diagram. [02] .In this paper we used star level metrics since the star schema is the main issue of data warehouse multidimensional model. Following table shows the proposed metrics (by piatiani et al [02]) for object-oriented multidimensional (OOMD) model

Table 1: OOMD model metrics

| Metrics | Description |
|---|---|
| NDC(S) | Number of dimension classes of the star S |
| NBC(S) | Number of base classes of the star S |
| NC(S) | Total number of classes of the star S<br>NC(S) = NDC(S) + NBC(S) + 1 |
| NA(S) | Total number of FA, D and DA attributes of the star S |

To carry out our experimental task we select the metric proposed by Pittani at el [02]. They selected representative examples of real world cases. Every schema had different metrics values. The selected metrics and their values are given in the following table.

Table 2: Metrics Values

| Schema Number | NDC | NBC | NC | NA |
|---|---|---|---|---|
| | | | | |
| 1 | 6 | 16 | 23 | 17 |
| 2 | 5 | 19 | 25 | 32 |
| 3 | 2 | 5 | 8 | 14 |
| 4 | 4 | 17 | 22 | 27 |
| 5 | 3 | 21 | 25 | 36 |
| 6 | 5 | 13 | 19 | 34 |
| 7 | 3 | 6 | 10 | 12 |
| 8 | 4 | 5 | 10 | 21 |
| 9 | 3 | 5 | 9 | 19 |
| 10 | 2 | 4 | 7 | 10 |
| 11 | 7 | 24 | 26 | 35 |

### 3.2 Experiments

We export the selected data to an arff file (i.e., a text based data file readable by the WEKA explorer), which is then loaded into the WEKA explorer. The classifier is the J48 tree learning algorithm provided by the WEKA tool. The accuracy is calculated with ten-fold cross validation.In this algorithm we used all available data set from Pittani at el to predict the understandability of object –oriented multidimensional model of data warehouse.

Fig 1 shows the generated decision tree. We can see that the attribute NC appear at the root. Attribute on the second level are NBC and last level includes NA and NDC. We got

```
Correctly Classified Instances          44               54.321  %
Incorrectly Classified Instances        37               45.679  %
```

This is good, looking at the confusion matrix

```
=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
  0  0  1  0  0 |  a = VERY_EASY
  0  0  7  0  0 |  b = EASY
  0  0 43  5  0 |  c = MODERATE
  0  0 19  1  0 |  d = POOR
  0  0  4  1  0 |  e = VERY_POOR
```

The top row of the confusion matrix shows the labels of the predicted classes. On the right are the actual classes. Each cell of the matrix denotes the number of instances (source files) classified as a, b, c, d, or e. For instance, the numbers in the bottom row state that 0 instances which are of the actual class, where classified correctly, 1 instances where wrongly classified as c. The confusion matrix illustrates the good performance of the classifier. By looking at the diagonal we see moderate dispersion of the values. If we count near misses, the prediction, especially for class e, is excellent. Following are all the detail of all the correctly and incorrectly instances produced by the tree learner.

1. No instance classified as wrong i.e. all 4 values of confusion matrix are non zero.
2. No instance classified as wrong i.e. all 4 values of confusion matrix are non zero.
3. In this 43 instances correctly classified by the algorithm and 5 instance which is d are incorrectly classified, means out of 48 instances 43 classified correctly and five were wrong defined by the algorithm.
4. In this 19 instances correctly classified and one instance classified wrongly means out of 20 only one instance is wrong.
5. In this 4 instances are correctly classified and only one instance incorrectly classified as d

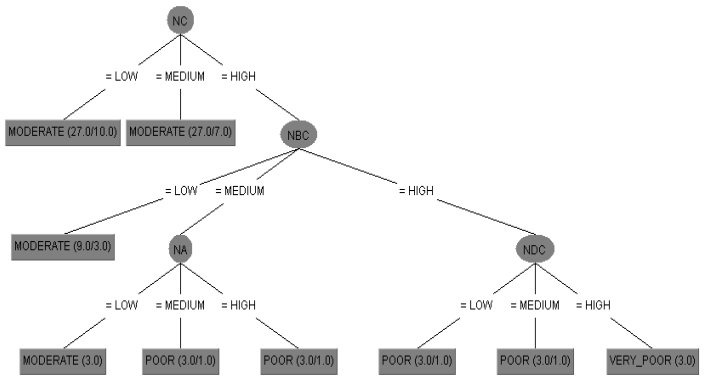


Figure 1: Generated Decision Tree of undersatandability by WEKA

In the last level we can see that number of dimension classes (NDC) in all cases i.e. LOW, MEDIUM and High classified as Poor, this confirms that if our model have high metrics values then it will be less understandable is wrong means understandability of the model not depend on the metrics values so it leads to reject hypothesis 2.

Due to complex relationships between the various metrics we could only partly identify factors that lead to high understandability. Using error classifier graph we can conclude that maximum classification problems occurred in the moderate mode. But these are not so much clears that why they occurred. This resulted in the partly rejection of Hypothesis 3.

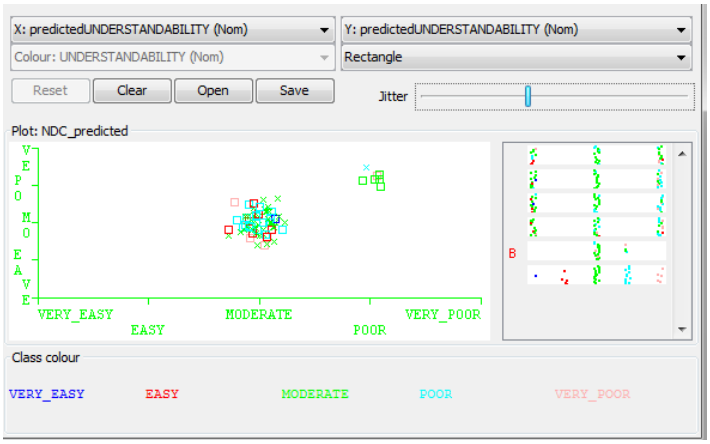


Figure 2:Error Classification Graph

## 4. CONCLUSION AND FUTURE WORK

In this paper we applied decision tree on OOMD model schema values to predict understandability. For this we stated a set of hypotheses that we addressed in a series of experiments with metrics proposed by Pittani et al. The data mining experiments showed that a decision tree learner (J48) can produce reasonable results with respect of stated hypothesis, with various values of input data. Future work is concerned with including detailed measures of OOMD model metrics (e.g., NBC, NDC, NA and NBC). With this additional information we can gain deeper insights into the internals of the implementation. Another area of future work is to use another data mining techniques and conduct our experiments with more metrics of OOMD model of data warehouse from the open source community as well as industrial software systems.

## REFERENCES

## References:

**[1]** Manuel Serrano, Coral Calero, Mario Piattini, "Experimental Validation of Multidimensional Data Models Metrics", Proceedings of the 36th Hawaii International Conference on System Sciences – 2003.

[2] Manuel Serrano , Juan Trujillo, Coral Calero, Mario Piattini , " Metrics for data warehouse conceptual models understanadability", Proceedings of the 36th Hawaii International Conference on System Sciences

[3] Manuel Angel Serrano, Coral Calero, Houari A. Sahraoui, Mario Piattini."Empirical studies to assess the understanadability of data warehouse schemas using structural metrics", Software Qual J (2008), 6:79–106 DOI 10.1007/s11219-007-9030-7

[4] Daniel L. Moody, Guttorm Sindre, Teqje Brasethvik, "Evaluating the Quality of Information Models: Empirical Testing of a Conceptual Model Quality Framework", 0-7695-1877-X/03 © 2003 IEEE.

[5] Calero C., Piattini M., Carolina Pascual, Serrano, M. A. "Towards Data warehouse quality metrics", 3rd International workshop on design and Management of Data warehouses (DMDW 2001), Interlaken, Switzerland

[6] Serrano, M.Calero, C.Piattini, M., "Validating metrics for data warehouses" IEEE Proceedings SOFTWARE, 2002.

[7] Fenton, N., & Fleeter, S. Software metrics: A rigorous approach (2nd Ed.). London: Chapman & Hall. 1997

[8] Serrano, M.Calero, C.,Sahraoui, H.,Piattini, M., Empirical Studies to Assess the Understandability of Data Warehouse Schemas using Structural Metrics, Software Quality Journal. Springer, 2008. Pages: 79- 106

[09] M. Serrano, C. Calero, J. Trujillo, S. Lujan, M. Piattini, Empirical validation of metrics for
conceptual models of data warehouse, 16th International Conference on Advanced Information Systems Engineering (CAISE'04), Riga, Latvia, 2004, pp. 506–520.

[10] Jarke, M., LenzerinI, I. M., Vassilou, Y., & Vassiliadis, P. Fundamentals of data warehouses, Springer, 2000.

[11] Serrano, M.Calero, C.Piattini, M., "Validating metrics for data warehouses", IEEE Proceedings SOFTWARE, 2002

[12] Sam Drazin and Matt Montag "Decision Tree Analysis using Weka" Machine Learning – Project II, University of Miami

[13] Y.W. The "Data Mining Techniques Using Decision Tree Model in Materialized Projection and Selection View" Mathware & Soft Computing 11 (2004) 51-66

[14] Tibor Gyimothy, Rudolf Ferenc, and Istva´n Siket, "Empirical Validation of Object-Oriented Metrics on Open Source Software for Fault Prediction" IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, VOL. 31, NO. 10, OCTOBER 2005.

[15] Yuming Zhou and Hareton Leung, "Empirical Analysis of Object-Oriented
Design Metrics for Predicting High and Low Severity Faults" IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, VOL. 32, NO. 10, OCTOBER 2006.

[16] R. Ferenc, _AA. Besze´des, M. Tarkiainen, and T. Gyimo´thy, "Columbus—Reverse Engineering Tool and Schema for

C++," Proc. 18th Int'l Conf. Software Maintenance (ICSM 2002), pp. 172-181, Oct. 2002

[17] Patrick Knab, Martin Pinzger, Abraham Bernstein "Predicting Defect Densities in Source Code Files with Decision Tree Learners", MSR'06, May 22–23, 2006, Shanghai, China. ACM 1-59593-085-X/06/0005