Classification and Regression Trees as a Part of Data Mining in Six Sigma Methodology

Andrej Trnka, *Member, IAENG*

***Abstract*—The paper deals with implementation of the classification and regression trees into the DMAIC phases of Six Sigma methodology. Six Sigma methodology seeks to improve the quality of manufacturing process by identifying and minimizing variability of this process. Using the classification, regression and segmentation trees as a part of the Data Mining methods could improve results of DMAIC phases. This improvement has a direct impact on the Sigma performance level of processes. The author introduces research results of implementation Data Mining algorithms into retail sales promotion.**

***Index Terms*—Classification, Data Mining, DMAIC, regression, Six Sigma.**

I. Introduction to Six Sigma Methodology

Six Sigma is a rigorous, focused, and highly effective implementation of proven quality principles and techniques. Sigma (σ) is a letter in the Greek alphabet used by statisticians to measure the variability in any processes. A company's performance is measured by sigma level of their business processes. Traditionally companies accept three or four sigma performance levels as a norm. However, these processes created between 6,200 and 67,000 problems (or defects) per million opportunities (DPMO)! The Six Sigma standard of 3.4 problems-per-million opportunities is a response to the increasing expectations of customers and the increased complexity of modern products and processes.

TABLE I SIX SIGMA VALUES

Sigma	DPMO	Yield (%)
1	690 000	31
2	308 000	69.2
3	66 800	93.32
4	6 210	99.379
5	230	99.977
6	3,4	99.9997

Table I shows Six Sigma values with corresponding DPMO. The Yield column shows the yield of the process in percentage.

The Six Sigma methodology tools are frequently applied within a performance-improvement model known as Define-Measure-Analyze-Improve-Control (DMAIC) for an existing process or Define-Measure-Analyze-Design-Verify (DMADV) for a new process.

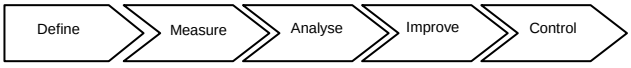

Fig. 1 DMAIC Model

Fig. 1 shows the model of DMAIC phases. In the Define phase the goals of the improvement activity are defined. At the top level the goals will be the strategic objectives of the organization. At the operations level, a goal might be to increase the throughput of a production department. At the project level goals might be to reduce the defect level and increase the throughput. In the Measure phase, the existing system is measured. Establish valid and reliable metrics to help monitor progress towards the goal(s) defined in the previous step. In the Analyze phase, the system is analyzed to identify ways to eliminate the gap between the current performance of the system or process and the desired goal. In the Improve phase the system is improved. In the Control phase the new system is controlled. [6], [8], [9]

The use of classification and regression trees could improve results of individual DMAIC phases. These sub-results have direct impact to the Sigma performance level of process. As shown in Table 1, Sigma performance level is tied to the DPMO value. And finally, the yield of the process is directly proportional on the DPMO value.

The goal of implementation of classification and regression trees (and other Data Mining methods) to DMAIC phases is to lower DPMO value and thereby reduce production costs.

The first step of the classification and regression trees implementation into Six Sigma methodology is to understand how the Knowledge Discovery in Databases (KDD) process works.

II. Knowledge Discovery in Databases

KDD is a process of finding interesting, useful and novel data. This fact is a reason why we describe this process. Implementation of classification and regression trees has to bring a new knowledge from production processes. If we have a lot of data, we need to find new relations, addictions and patterns. Classification and regression trees are one of possible ways how to find them.

Manuscript received May 12, 2010.

Andrej Trnka is with the University of. Ss. Cyril and Methodius in Trnava, Faculty of Natural Sciences, Department of Applied Informatics, Trnava, Nam. J. Herdu 2, 917 01 Slovak Republic (corresponding author to provide e-mail: andrej.trnka@ucm.sk).

Fig. 2 shows all steps of KDD process. Implementation of classification and regression trees must be made in Data Mining step. It is very important to realize that Data Mining is only one step of KDD process. So, the data from the process must go over all previous steps.

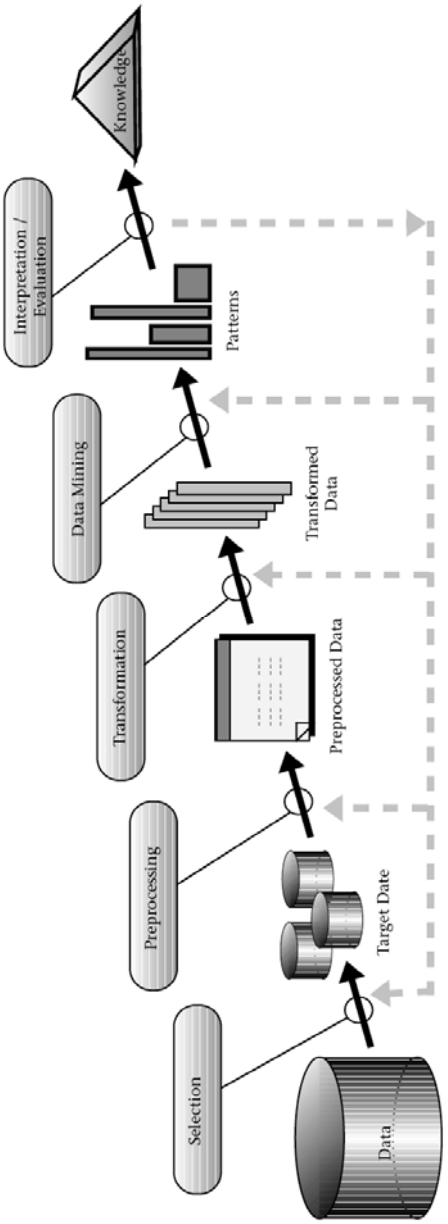

Fig. 2 Steps of the KDD process

III. Classification and Regression Trees

Tree models begin by producing a classification of observations into groups and then obtain a score for each group. The tree models are usually divided into regression trees (the response variable is continuous) and classification trees (the response variable is quantitative discrete or qualitative – categorical). However, as most concepts apply equally well to both, here we do not distinguish between them, unless otherwise specified. Tree models can be defined as a recursive procedure, through which a set of n statistical units are progressively divided into groups, according to a division rule that aims to maximize a homogeneity or purity measure of the response variable in each of the obtained groups. At each step of the procedure, a division rule is specified by the choice of an explanatory variable to split and the choice of a splitting rule for the variable, which establishes how to partition the observations. The main result of a tree model is a final partition of the observations. [1], [3]

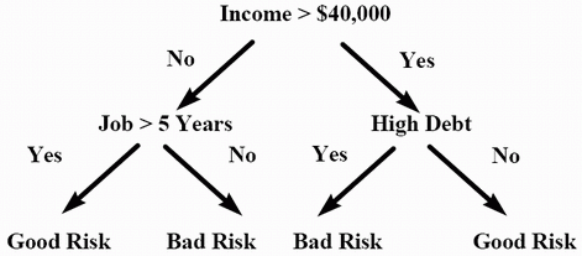

Fig. 3 Classification tree

We can use several tree models in Six Sigma methodology. The most popular are:

- CART
- CHAID
- QUEST
- C5.0

These trees can be used in many areas of production processes. [4]

CART stands for Classification and Regression Trees, originally described in the book with the same title [9]. CART divides up the data into two subsets so that the records within each subset are more homogeneous than in the previous one. It is a recursive process – each of those two subsets is then split again, and the process is repeated until the homogenenity criterion is reached or until another stopping criterion is satisfied (as do all of the tree-growing methods). The same predictor field may be used several times at different levels in the tree. It uses surrogate splitting to make the best use of data with missing values.

CHAID stands for Chi-squared Automatic Interaction Detector. It is a highly efficient statistical technique for segmentation, or tree growing, developed by Kass (1980). Using the significance of a statistical test as a criterion, CHAID evaluates all of the values of a potential predictor field. It merges values that are judged to be statistically homogeneous (similar) with respect to the target variable and maintains all other values that are heterogeneous (dissimilar). It then selects the best predictor to form the first branch in the decision tree, such that each child node is made of a group of homogeneous values of the selected field. This process continues recursively until the tree is fully grown. The statistical test used depends upon the measurement level of the target field. If the target field is continuous, an F test is used. If the target field is categorical, a chi-squared test is used. CHAID is not a binary tree method. It can produce more than two categories at any particular level in the tree. Therefore, it tends to create a wider tree than the binary growing methods do. It works for all types of variables and it accepts the both case weights and frequency variables. It handles missing values by treating them all as a single valid category.

QUEST stands for Quick, Unbiased and Efficient Statistical Tree. It is a relatively new binary tree-growing algorithm. It deals with split field selection and split-point selection separately. The univariate split in QUEST performs approximately the same unbiased field selection. That means,

ISBN: 978-988-17012-0-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

if all predictor fields are equally informative with respect to the target field, QUEST selects any of the predictor fields with equal probability. QUEST affords many of the advantages of CART, but, like CART, trees can become unwieldy. Automatic cost-complexity pruning can be applied to a QUEST tree to cut down its size. QUEST uses surrogate splitting to handle missing values.

C5.0 is sophisticated Data Mining tool for discovering patterns that delineate categories, assembling them into classifiers and using them to make predictions. C5.0 has been designed to analyze substantial databases containing thousands to hundreds of thousands of records and tens to hundreds of numeric, time, date, or nominal fields. C5.0 also takes advantage of processors with quad cores, up to four CPUs, or Intel Hyper-Threading to speed up the analysis. To maximize interpretability, C5.0 classifiers are expressed as decision trees or sets of if-then rules, forms that are generally easier to understand than neural networks. C5.0 is easy to use and does not presume any special knowledge of Statistics or Machine Learning. [2], [5]

IV. Field of the Application

The implementation of classification and regression trees into DMAIC phases can be applied in many fields of production process or business process generally.

CHAID – describing which customers are most likely to respond to the promotion

QUEST – identifying promotional target customers (with RFM analysis)

CART – retailing sales promotion, preparing data for analysis

C5.0 – machine condition monitoring, market basket analyzing, identifying promotional target customers (with RFM analysis)

Each implementation of Data Mining methods to Six Sigma methodology should be evaluated. [7]

In our research we decided to implement these tree-based tasks into the following phases of Six Sigma methodology.

Define – preparing data for analysis, identifying promotional target customer.

Measure – describing which customers are most likely to respond to the promotion.

Analyze – analyzing market basket.

Improve – retail sales promotion.

Control – machine condition monitoring.

The implementation of classification and regression trees into DMAIC phases requires other Data Mining methods, too (e.g. neural networks).

V. Results

A part of our research is the strategy of implementing Data Mining algorithms into DMAIC phases. We decided to implement the retail sales promotion into the Improve phase. The aim was to predict the effects of sales promotion in future. The process of Data Mining consisted of:

- exploration,
- data preparation,
- training,
- testing.

Each record in our dataset contained data about:

- Type – type of the product,
- Cost – unit price,
- Revenue1 – revenue before promotion,
- Revenue2 – revenue after promotion,
- Increase (derived data) – increase in revenue after the promotion.

Dataset contained 142 records (Fig. 4). The field Increase was derived, because the two revenue fields were expressed in absolute terms.

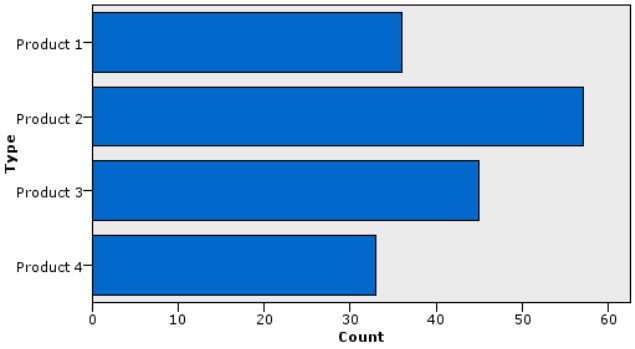

Fig. 4 Frequency of products

Figure 5 shows the histogram of increase in revenue overlay with color of product's type.

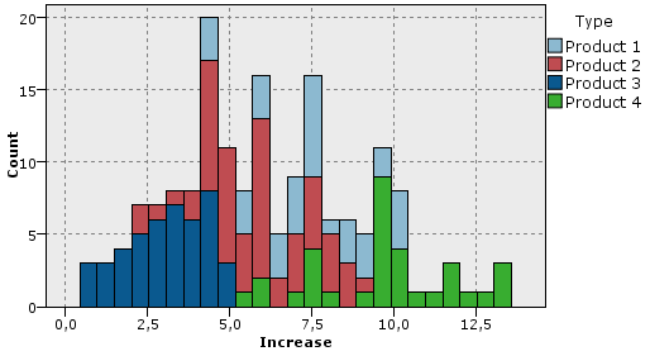

Fig. 5 Histogram of increase in revenue (after promotion)

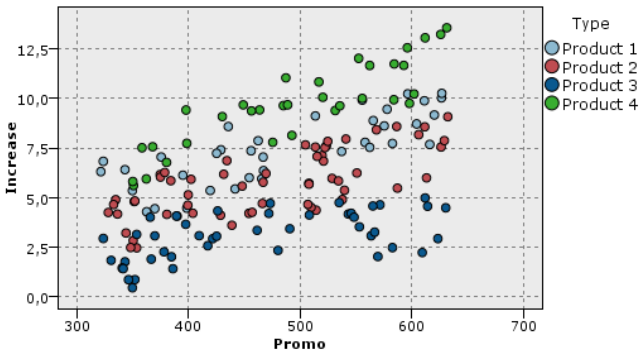

Fig. 6 Scatterplot Promotion vs. Increase

The scatterplot in Fig. 6 shows that for each type of a product, the relationship between the increase in revenue and the cost of the promotion exists. This is a reason, why we decided to apply the decision tree (esp. CART) to predict the effect of sales promotion.

Figure 7 shows the proposed model built in IBM SPSS Modeler.

ISBN: 978-988-17012-0-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

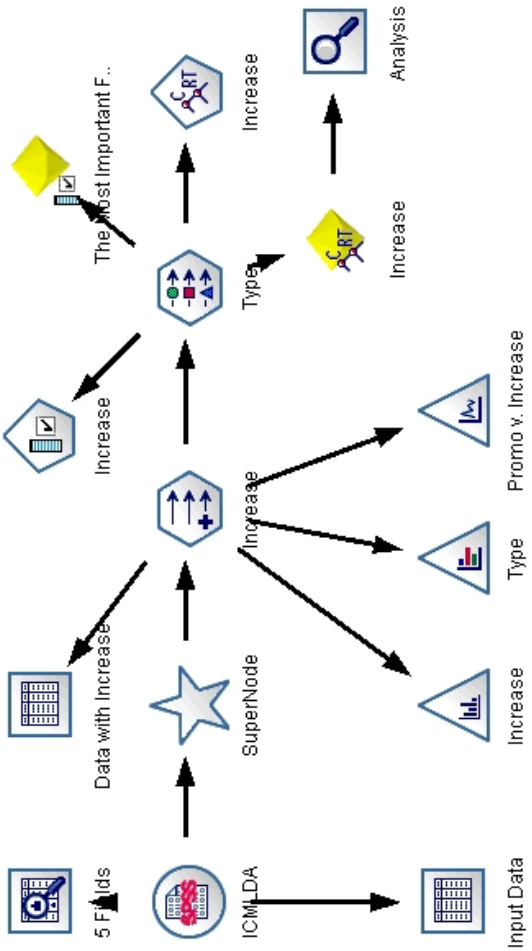

Fig. 7 Model built with CART algorithm

After training the model we are able to make the prediction of revenue increase. We can test the results of the learning process with changing the input data and executing the Analysis node. From the linear correlation between the predicted increase and the correct answers, we are able to find that the trained system predicts with a high degree of success (Table II).

TABLE II RESULTS FOR OUTPUT FIELD

Minimum Error	-3.278
Maximum Error	2.719
Mean Error	-0.118
Mean Absolute Error	1.144
Standard Deviation	1.399
Linear Correlation	0.876
Occurences	91

We suggest implementing this model to the Improve phase (Fig. 8) of Six Sigma methodology, because historical data from old processes can improve decision concerning the new promotion. The correct prediction will improve the future results and related revenue (direct impact to Six Sigma performance level).

The yield of the manufacturing process before implementation was 97.7%, which corresponds with 3.5 Sigma. After implementation, the yield was increased to 98.6%, which corresponds with 3.7 Sigma.

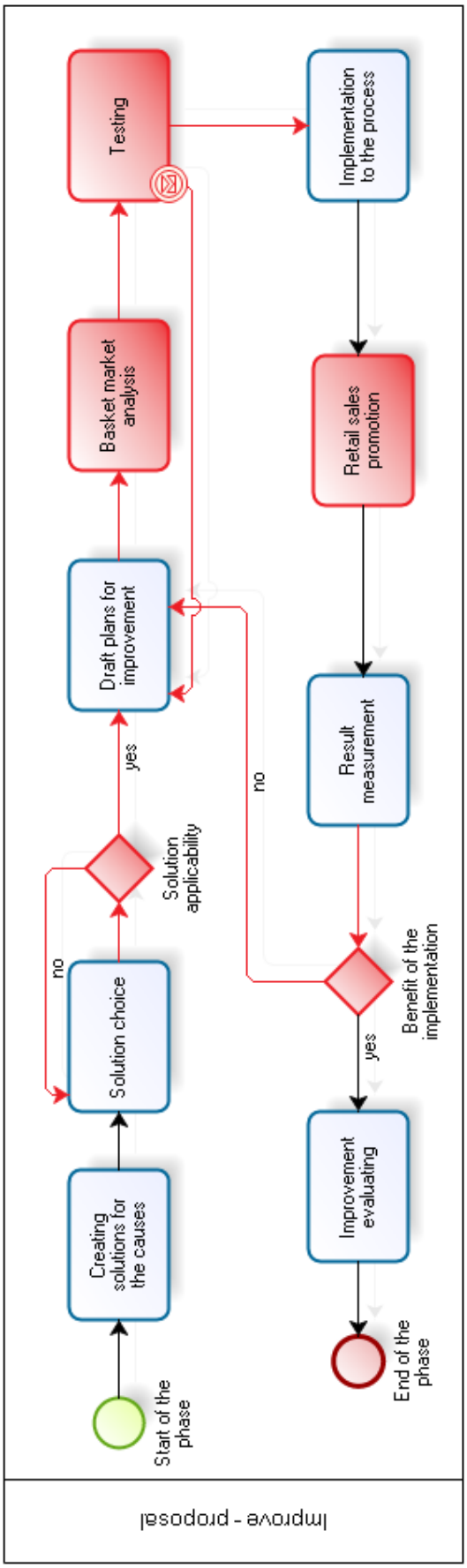

Fig. 8 Retail sales promotion in Improve phase

ISBN: 978-988-17012-0-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

ACKNOWLEDGMENT

Grateful acknowledgment for translating the English edition goes to Juraj Mistina.

REFERENCES

[1] P. Giudici, S.Figini. Applied Data Mining for Business and Industry. Second Edition. John Wiley & Sons Ltd, 2009, pp. 71. ISBN 978-0-470-05886-2
[2] Information on See5/C5.0 http://www.rulequest.com/see5-info.html [01/05/2010]
[3] Introduction to Data Mining and Knowledge Discovery. Third Edition. Two Crows; 2005, pp. 35, ISBN 1-892095-02-5
[4] M. Kebisek, M. Elias. The possibility of utilization of knowledge discovery in databases in the industry. In: Annals of MTeM for 2009 & Proceedings of the 9th International Conference Modern Technologies in Manufacturing; 2009 October 8-10; Cluj-Napoca, Romania. Cluj-Napoca: Technical University of Cluj-Napoca, 2009. ISBN 973-7937-07-04. p. 139-142
[5] PASW Modeler 13 – Algorithm Guide. SPSS; 2009
[6] T. Pyzdek, T. Keller. The Six Sigma Handbook. Third Edition. The McGraw-Hill Companies, 2010. ISBN 978-0-07-162337-7
[7] J. Zeman, P. Tanuska, M. Kebisek. The Utilization of Metrics Usability To Evaluate The Software Quality. In: ICCTD 2009 : International Conference on Computer Technology and Development. 13-15 November 2009, Kota Kinabalu, Malaysia. IEEE Computer Society, 2009. ISBN 978-0-7695-3892-1
[8] Six Sigma and Lean Resources - DMAIC Cycle. http://6sixsigma.com/index.php/DMAIC-Cycle.html [01/04/2010]
[9] L. Breiman et.al. Classification and Regression Trees. Chapman and Hall, 1984. ISBN 978-0412048418