2010 International Conference on Networking and Information Technology

Market Basket Analysis with Data Mining Methods

Six Sigma methodology improvement

Andrej Trnka
Department of Applied Informatics
University of SS. Cyril and Methodius
Trnava, Slovak Republic
andrej.trnka@ucm.sk

***Abstract*— This paper describes the way of Market Basket Analysis implementation to Six Sigma methodology. Data Mining methods provide a lot of opportunities in the market sector. Basket Market Analysis is one of them. Six Sigma methodology uses several statistical methods. With implementation of Market Basket Analysis (as a part of Data Mining) to Six Sigma (to one of its phase), we can improve the results and change the Sigma performance level of the process. In our research we used GRI (General Rule Induction) algorithm to produce association rules between products in the market basket. These associations show a variety between the products. To show the dependence between the products we used a Web plot. The last algorithm in analysis was C5.0. This algorithm was used to build rule-based profiles.**

Keywords-Data Mining; Six Sigma; Market Basket; CRISP-DM

I. INTRODUCTION

In our research we tried to implement a few Data Mining methods to Six Sigma methodology and improve results with them. Reference [7] describe possibility of utilization Data Mining methods in industry, generally. However, Six Sigma is specific with its approach to capability of the process. Six Sigma methodology is used to improve the business processes.

We can use DMAIC cycle for the existing process or DMADV cycle for a new process. The letters in the acronym mean activities in Six Sigma methodology.

DMAIC – Define, Measure, Analyze, Improve, Control

DMADV – Define, Measure, Analyze, Design, Verify

In our research we improve the existing process, so we focus to the DMAIC cycle of Six Sigma methodology.

The **Define phase** is where we start each project. It's helpful to think of the outputs of each stage to know what the stage does. In this stage we establish the goal of the project. We should know who our customer is, which process we will be working on, which part of the process we will work on, which of the process variables are important, what the goal is for those process variables, how much money we expect to save or other benefits we will get, when the project will be finished, who the project team is, and who the stakeholders are. It is in the Define phase where we first see evidence of Six Sigma thinking. The goal of a project always has to be stated as a measurement, the value of which will be improved. It might be stated as several measurements and goals for each of them. Each of these measurements is an important process output that matters to the customer. All six similar projects must identify who the customer is for each process. The customer sets the acceptable parameters for each of the key process outputs. These are usually expressed as specifications, but might take other forms. The key is that no project can proceed unless a quantifiable goal is stated. At the end of the define stage you will have a well-defined project with well-defined and quantifiable goals.

The next stage is the **Measure phase**. At the output of this phase we should have a thorough understanding of how the process is behaving right now. By understanding we mean a quantitative description. This description will include current process variable averages, standard deviations, behavior over time, and histograms. In addition, we will know whether the process is stable or not. Another thing we will know at the end of this stage is whether or not the process has a capability and what the value is. The Sigma performance level, for which Six Sigma is named, is an example of a capability measure. To do all of this we have to collect data, and to know what data to collect, we need to know which characteristics are critical. Sometimes this is determined in the Define phase, but sometimes what we are given in the Define phase is not sufficient for us to collect data. The first thing we would then do is determine the precise measurements necessary to determine process capability.

The third phase in a DMAIC cycle is the **Analyze phase**. The overall approach in Six Sigma problem solving is to carefully define a problem, find the root causes for the problem, and then attack the root causes. The purpose of the Analyze phase is to correctly identify what those root causes are and prove it with data. In this disciplined problem-solving method, opinions are not worth much. We might have historical data that we can analyze using some basic statistics or we might have to conduct some experiments and collect some new data.

In the **Improve phase**, we come up with solutions. This phase also often starts with a brainstorming session. But now we know what the root causes are. A root cause is something that is controllable and has a direct effect on the characteristic we are trying to improve. This problem-solving session again can be short and simple or long and complex, depending on process. At the end of this phase, the process improvement is installed and the process is now running per goals. Some Six Sigma programs split this phase into two,

978-1-4244-7578-0/$26.00 © 2010 IEEE

where the improve stage involves figuring out what the solution is and the implement stage actually implements the solution. This is because there are different skills needed to problem solve (improve) and to build (implement).

The final DMAIC phase is **Control**. This is unique to the Six Sigma approach. Most managers have heard of the Hawthorne effect. The Hawthorne effect is a temporary change of behavior or performance in response to a change in the environmental conditions, with the response being typically an improvement. The term was coined in 1955 by Henry A. Landsberger. Landsberger defined the Hawthorne effect as a short-term improvement caused by observing worker performance. In other words, it's not what we did that improved performance, it's the fact that we paid attention to it that caused everyone around you to behave better than normal. Even if the Hawthorne effect isn't operating on project, because problems are more technical and equipment oriented and less reliant on people's behavior, there is still a tendency for systems that have undergone an improvement to degrade back to where they were. In this phase the team makes sure that appropriate things are done so that the process will continue to perform at its new level. This can include administrative things like making sure that training materials or written procedures are modified, making sure that new specifications are transmitted and understood by everyone who needs them, including suppliers, and making sure that critical control points are monitored and that procedures exist to react quickly when something goes out of balance.

Fig. 1 shows the DMAIC phases in Six Sigma methodology. [2]

Defects per Million Opportunities (DPMO) is a major metric in Six Sigma methodology. DPMO is a measure of the process performance. DPMO is the average number of defects per unit observed during an average production process divided by the number of opportunities to make a defect on the product under study during that process normalized to one million. Tab. I shows Six Sigma performance level with DPMO and corresponding yield and defects of the process. [3]

TABLE I. SIX SIGMA PERFORMANCE LEVEL

Sigma performance level	DPMO	Percent defective	Percentage yield
1	691 462	69%	31%
2	308 538	31%	69%
3	66 807	6,7%	93,3%
4	6 210	0,62%	99,38%
5	233	0,023%	99,977%
6	3,4	0,00034%	99,99966%

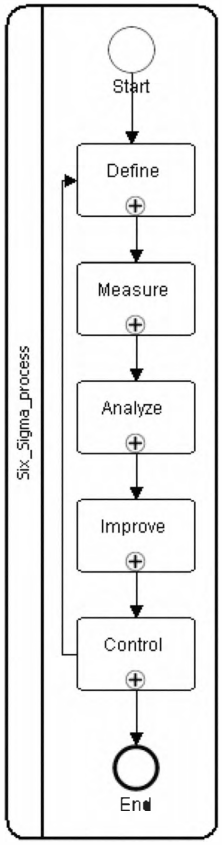

Figure 1. DMAIC cycle (in BPMN notification)

We can use a lot of Data Mining methods in Six Sigma methodology. This implementation can improve results of the process by increasing the yield and decreasing the DPMO value. This has a direct impact to the Sigma performance level. The aim is to achieve the highest possible value of the Sigma performance level.

Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques. [4]

There exist a few Data Mining models. Nowadays Data Mining processes are based on "CRoss-Industry Standard Process for Data Mining" (CRISP-DM). This model consists of six phases intended as a cyclical process.

Business Understanding includes determining business objectives, assessing the current situation, establishing data mining goals, and developing a project plan.

Data Understanding considers data requirements. This step can include initial data collection, data description, data exploration, and the verification of data quality. Data exploration such as viewing summary statistics (which includes the visual display of categorical variables) can occur at the end of this phase. Models such as cluster analysis can also be applied during this phase, with the intent of identifying patterns in the data.

Data Preparation prepares the data. Once the data resources available are identified, they need to be selected, cleaned, built into the form desired, and formatted. Data cleaning and data transformation in preparation of data modeling needs to occur in this phase. Data exploration at a greater depth can be applied during this phase, and additional models utilized, again providing the opportunity to see patterns based on business understanding.

Modeling in Data Mining software tools such as visualization (plotting data and establishing relationships) and cluster analysis (to identify which variables go well together) are useful for initial analysis. Tools such as generalized rule induction can develop initial association rules. Once greater data understanding is gained (often through pattern recognition triggered by viewing model output), more detailed models appropriate to the data type can be applied. The division of data into training and test sets is also needed for modeling.

Evaluation of the results. Model results should be evaluated in the context of the business objectives established in the first phase (business understanding). This will lead to the identification of other needs (often through pattern recognition), frequently reverting to prior phases of CRISP-DM. Gaining business understanding is an iterative procedure in data mining, where the results of various visualization, statistical, and artificial intelligence tools show the user new relationships that provide a deeper understanding of organizational operations.

Deployment of the model. Data mining can be used to both verify previously held hypotheses, or for knowledge discovery (identification of unexpected and useful relationships). Through the knowledge discovered in the earlier phases of the CRISP-DM process, sound models can be obtained that may then be applied to business operations for many purposes, including prediction or identification of key situations. These models need to be monitored for changes in operating conditions, because what might be true today may not be true a year from now. If significant changes do occur, the model should be redone. It's also wise to record the results of data mining projects so documented evidence is available for future studies. [5]

Fig. 2 shows the CRISP-DM process.

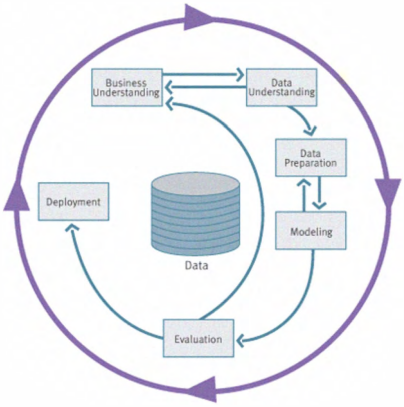

Figure 2. CRISP-DM process

The use of Data Mining methods in Six Sigma methodology requires joining of CRISP-DM process to DMAIC phases.

II. Market Basket Analysis

Market basket analysis has the objective of indentifying products, or groups of products, which tend to occur together (are associated) in buying transactions (baskets). The knowledge obtained from a market basket analysis can be very valuable; for instance, it can be employed by a supermarket to reorganize its layout, taking products frequently sold together and locating them in close proximity. But it can also be used to improve the efficiency of a promotional campaign: products that are associated should not be put on promotion at the same time. By promoting just one of the associated products, it should be possible to increase the sales of that product and get accompanying sales increases for the associated products.

The databases usually considered in a market basket analysis consist of all the transactions made in a certain sale period (e.g. one year) and in certain sale locations (e.g. a chain of supermarkets). Consumers can appear more than once in the database. In fact, consumers will appear in the database whenever they carry out a transaction at a sales location. The objective of the analysis is to find the most frequent combinations of products bought by the customers. The association rules in Section 4.8 represent the most natural methodology here; indeed they were actually developed for this purpose. Analyzing the combinations of products bought by the customers, and the number of times these combinations are repeated, leads to a rule of the type 'if condition, then result' with a corresponding interestingness measurement. Each rule of this type describes a particular local pattern. The set of association rules can be easily interpreted and communicated. Possible disadvantages are locality and lack of probability modeling.

In shop the recorded transactions are all the transactions made by someone holding one of the chain's loyalty cards. Each card carries a code that identifies features about the owner, including important personal characteristics such as sex, date of birth, partner's date of birth, number of children, profession and education. The card allows the analyst to follow the buying behavior of its owner: how many times they go to the supermarket in a given period, what they buy, whether they follow the promotions, etc. [1]

Similar analyses can be found in [1].

III. Our Research

Our research is focused on implementation the Data Mining methods to the Six Sigma methodology (to its phases). We decided to implement market basket analysis to the Improve phase, with this allows us to predict behavior of customer. The prediction can determine the main products of manufacturing process and profile the customer groups.

We suggest creating the Data Warehouse, because integrity of the data from process is variable. [6]

To provide anonymity of companies, we use common products label terms.

Our offer consists of ten products. We store basket contents, basket summary and personal information. For the sake of personal data we will use only fictive loyalty card numbers.

The first step we made was the acquirement of overall pictures of association between products in the basket. We used Generalized Rule Induction (GRI) to produce association rules. The dataset contains 634 records. Tab. II shows generated association rules between products. Data in the table are sorted by confidence.

TABLE II. ASSOCIATION RULES (A PART)

Consequent	Antecedent	Support %	Confidence %
product G	product B product D product F	3,47	95,45
product D	product B product F product G	3,47	95,45
product D	product A product F product G	4,42	92,86
product F	product C product D product G	2,21	92,86
product D	product E product F product G	3,79	91,67
product F	product B product D product G	3,63	91,3
product F	product D product G	16,56	88,57
product G	product A product D product F	4,73	86,67
product F	product A product D product G	4,73	86,67
product G	product C product D product F	2,37	86,67
product G	product D product F	17,03	86,11
product D	product F product G	17,03	86,11
product D	product E product G	5,68	72,22

These association rules (two-way) show a variety between products G, products D and products F.

Next step was a graphical view to associated products. We used a Web plot to show the dependence between the products. The results are showed in Fig. 3.

Bold lines in this plot show the groups of customers suggested by the GRI model. In the resulting display, two groups of customers stand out:

- those who buy product D, product F and product G (group A),
- those who buy product A and product J (group B).

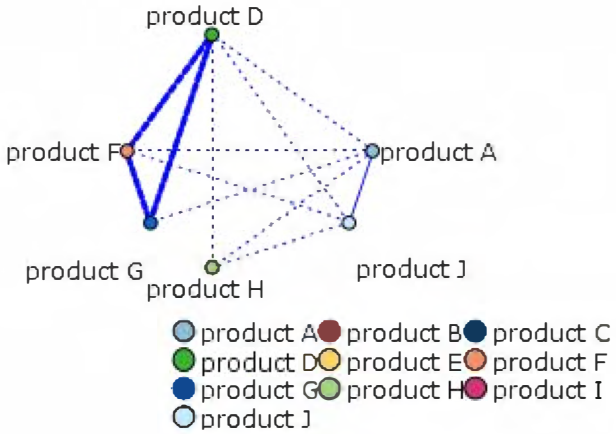

Figure 3. Web plot of associations

In the third step we profiled the customer groups. We needed to know who these customers are. This could be achieved by tagging each customer with a flag for each of these groups. To build rule-based profiles of these flags we used rule induction (C5.0). Consequently we received the following rules:

Generated rule for group A:

Rule 1 for T (true)
if income <= 16 900
and sex = M
then T

Generated rules for group B:

Rule 1 for T
if age <= 19
and income > 26 800
then T

Rule 2 for T
if age > 22
and age <= 24
and income <= 20 900
and payment_method = CASH
then T

Rule 3 for T
if age > 16
and age <= 22
and income > 12 200
and income <= 15 500
and payment_method = CARD
and sex = F
then T

Rule 4 for T
if age <= 24
and income > 26 800
and payment_method = CHEQUE
then T

Rule 5 for T

if age > 17
and age <= 24
and income > 17 700
and income <= 26 800
and payment_method = CARD
and value > 28,741
and value <= 49,158
then T

Rule 6 for T

if age <= 16
then T

Rule 7 for T

if age <= 24
and income > 17 700
and income <= 26 800
and payment_method = CARD
and value > 28,741
then T

Fig. 4 shows model built in IBM SPSS Modeler.

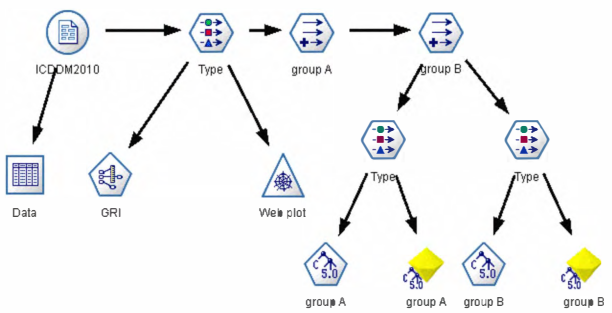

Figure 4. Model built with GRI, Web plot and C5.0 algorithm

With this analysis we made profiling products in market basket. This implementation of Data Mining methods to Improve phase of Six Sigma methodology might be used to target special offers. These special offers might improve the Six Sigma performance level (indirectly), because we can spend money with targeting a specific customer group.

Each implementation of Data Mining methods to Six Sigma methodology should be evaluated. [8]

REFERENCES

[1] P. Giudici, S. Figini, "Applied Data Mining for Business and Industry. Second Edition". John Wiley & Sons Ltd; 2009. ISBN 978-0-470-05886-2

[2] W. Bentley, P. T. Davis, "Lean Six Sigma: Secrets for the CEO". CRC Press; 2010. ISBN 978-1-4398-0379-0

[3] C. Gygi, N. DeCarlo, B. Williams, "Six Sigma for Dummies". Wiley Publishing, Inc.; 2005. ISBN 0-7645-6798-5

[4] D. Larose, "Discovering Knowledge in Data: An Introduction to Data Mining". John Wiley; 2005. ISBN 0-471-66657-2

[5] D. Olson, D. Dursun, "Advanced Data Mining Techniques". Springer; 2008. ISBN 978-3-540-76916-3

Article in a journal:

[6] R. Halenar, "Loading data into data warehouse and their testing – Zavádzanie údajov do dátového skladu a ich testovanie" In: Journal of Information Technologies, vol. 2 (2009), pp. 7-14. ISSN 1337-7469.

Article in a conference proceedings:

[7] 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. pp. 139-142 ISBN 973-7937-07-04.

[8] 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