

# Data Mining for Decision Support of the Quality Improvement Process

**Bashar Al-Salim**

Industrial and Management Systems Engineering  
University of Nebraska-Lincoln  
Bashar@bigred.unl.edu

**Mansour Abdoli**

Industrial and Management Systems Engineering  
University of Nebraska-Lincoln  
mansour@engrs.unl.edu

## ABSTRACT

A two-stage methodology is presented for enhancing the process of assigning quality problems to quality improvement teams in industrial firms. The method advances the decision support system of the quality improvement process by grouping the related quality problems in two steps: First, a soft grouping is performed using association rules as a data mining technique, and then, resulted groups are finalized by employing a costs minimization model. Moreover, to find the optimal groups, a mathematical programming language is used. Results show that this methodology is beneficial and attractive in making the quality improvement process more efficient and in providing support to managerial decisions for creating quality improvement teams. As a practical illustration, the implementation of this methodology is investigated for an EDM fast hole drilling process.

## Keywords

Quality Improvement, Quality Improvement Teams, Data Mining, Association Rules.

## INTRODUCTION

Traditionally, a products' quality improvement process could take one of two different approaches. The first approach is Juran's approach (Mitra, 1998) for quality improvement in which the quality problems or defects are eliminated one at a time. This approach is based on the identification of one specific quality problem as a project, and concentrating available resources to eliminate the problem. In the literature, this approach is also called a project-by-project approach. The other approach is Deming's approach (Mitra, 1998), which is called a bottom-up approach. In this approach more than one problem could be investigated at the same time provided that they are at the same levels of the organizational structure. Both approaches are based on establishing quality improvement teams (QIT) and then assigning quality problems to the teams. Typically, each team consists of personnel from different disciplines that are led by a team leader. Working as a team motivates the ease and the efficacy of the flow of information between the involved disciplinary departments. Furthermore, teamwork enhances the quality improvement process since it is more likely that there will be more skills and knowledge available among numerous team members as opposed to a single person. However, the success of the quality improvement process depends on the construction of these types of teams and what approach is used to solve quality problems. The importance of these issues are magnified when the members of the QIT are selected from different departments and participate on a part-time basis in addition to their departmental tasks. Hence selecting one of two possible approaches and the efficient allocation of human resource become important matters of decision support system of quality improvement.

Selection of the proper approach and efficient resource allocation are highly dependent on the type and the number of the assigned quality problems. Solving one quality problem (in project-by-project approach) will more likely require fewer employees compared to solving two or more individual quality problems simultaneously. On the other hand, and based on the economic of scale, solving few related quality problems simultaneously (in bottom-up approach) can be more efficient. Both mentioned above approaches have their own advantages and disadvantages. Hence, combining both approaches has the potential of offering advantages of both individual approaches. To use both approaches successfully, quality problems need to be grouped based on their relevance. Quality problems that are not significantly related to others can be investigated separately with a project-by-project approach, and each group with more than one quality problem can be approached using bottom-up approach. In this paper, we will present a two-stage clustering method that groups quality problems based on relevance and optimal cost for tem assignment.

Our method for the optimal clustering of quality problems consists of two steps based on the relevance of quality problems and the cost of forming QIT's. If some of the quality problems are related and occur mutually with high probability among

production batches, then they will be given a high chance of being investigated as a group by one specialized quality improvement team. This is due to the fact that the causes of the associated quality problems are common and/or related, and it will take the same skills and procedure to explore and develop a solution for the poor quality. Moreover, delivering a methodology for categorizing the quality problems in groups where each group has the correlated or related problems will support the idea of having specialized QIT. Hence, the output from this step is a set of quality problem groups, where a problem may be shared in more than one group and some groups include only one problem. For example, consider a EDM fast hole process with quality problems listed in table 1. Followings are two instances of possible groups of related quality problems: 1) a surface defect in a product due to nonconforming weight variable; and 2) a group of quality problems consists of defects in height, width, and weight variables which occur often associated with each other in production batches. The second step in our method is to minimize the quality improvement costs. In this step, an optimal set of clusters is selected from overlapping groups generated in the first step. The cost factors used in this step are obtained from the manufacturing information system of the firm. The final result would be a non-overlapping set of clusters. The application of this step in our example above may result to considering defects associated to weight as an individual project despite the relevance of height, width, and weight variables from the operational point of view. The final recommendation of our method could be based on the following rationale: there is a possibility that investigating and solving the height, width, and weight quality problems individually are cheaper than handling three of them together.

To obtain a soft set of clusters (overlapping groups) of quality problems we suggest mining the past tests records stored in the firm's quality control database using the *association rules*. We also consider using a mathematical programming language to solve our cost optimization problem in the second step. Moreover, we use the example mentioned above to illustrate a practical application of our proposed method.

## PREVIOUS WORKS

Kusiak and Kurasek (Kusiak and Kurasek, 2001) studies the causes of solder defects in the printed circuit boards using the data mining approach; more specifically, they presented rough set theory approach to identify the causes of the poor quality in electronics assembly. Their results categorize three separate rule sets of conditions for occurrence of defects. In another study (Steckenrider, Guha, Sethuraman, Ra and Kim, 2001), the automated defect classification system was presented. The study assessed, with high accuracy, the defect classification performance by a single estimate of the overall defects.

Lee and Park (Lee and Park, 2001) proposed an optimal measurement sampling method using the data mining approach to maintain the high quality in the semiconductor manufacturing process. The sampling method is based on applying Self-Organizing Map Neural Networks as a data mining technique on the post-process measurements of the wafer bin map data. They argued that the proposed sampling method is an efficient method even with a small sampling size. Another implementation of data mining for the quality control in the semiconductor manufacturing is presented by Last and Kandel (Last and Kandel, 2001). They use Information-Fuzzy Networks methodology to build a classification models for the throughput of the semiconductor production line and for the flow times of batches.

An integrated data mining procedure using Neural-Networks/Partial Least Squares method is suggested by (Oh, Han and Cho, 2001) to monitor, diagnose, and optimize the quality system's improvement. The systematic procedure consists of three data mining modules, which are applied in the process industry: processing, modeling and knowledge identification. A reduction in the investigation costs of quality defects in a case study for the shadow mask manufacturing process is reported. Finally, the procedure for developing our methodology is similar to Agrad and Kusiak approach (Agrad and Kusiak, 2004) in investigating the subassembly problem in the manufacturing systems. Also this approach is analogous to Al-salim (AL-salim, 2005) study for designing the optimal travel package for travel agencies using data mining technique.

## THE MODEL

The procedure for finding the optimal set of clusters of quality problems consists of two steps. The first step is to determine the related quality problems using data mining techniques based on their likelihood to appear together in the production batches. We choose association rules for this purpose since effectively reduces the number of possible combinations of the groups of quality problems, but it allows some flexibility for cost considerations in the second step. In the second step, an optimization model is employed for selecting the best set of clusters (non-overlapping groups) of quality problems that its total costs associated with adopting a quality improvement system including forming QIT's is minimum.

## Notations

Before we continue developing the model, a list of the notations that will be used in this paper is presented:

- $i$ : Group of quality problems index
- $j$ : Quality problem index
- $n$ : Total number of resulted groups of quality problems after the first stage
- $m$ : Total number of quality problems
- $Q_i$ : Group  $i$  of quality problems
- $P_i$ : Prevention costs for the group quality problem  $i$
- $A_i$ : Appraisal costs for the group quality problem  $i$
- $I_i$ : Internal failure costs due to for the group quality problem  $i$
- $E_i$ : External failure costs for the group quality problem  $i$
- $T(i)$ : Total quality costs

$q_{ij}$ : Indicators function for having the quality problem  $j$  in the group of quality problems  $Q_i$ . Mathematically;

$$q_{ij} = \begin{cases} 1 & \text{If quality problem } j \text{ is in the quality problem group } Q_i \\ 0 & \text{If component } j \text{ is not in the quality problem group } Q_i \end{cases}$$

Finally, the decision variables are

$I_{Q_i}$ : Indicator function for selecting or not selecting the group  $Q_i$  of quality problem. Mathematically:

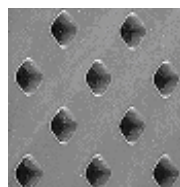
$$I_{Q_i} = \begin{cases} 1, & \text{If quality problem group } Q_i \text{ is selcted} \\ 0, & \text{if quality problem group } Q_i \text{ is not selcted} \end{cases}$$

**The first stage: Association Rules Technique**

Association Rules - or what is well-known as basket market (Post, 2005) - is considered one of the most popular data mining techniques. It is a data mining technique where rules are extracted out of the database and then ordered based on the percentage of times they are correct and how often they apply. It captures the relation among events in a database and it takes the form of: if event  $A$  (called antecedent event) occurred then event  $B$  (called consequent event) will happen with a certain probability. Symbolically,  $A \Rightarrow B$  and it is read  $A$  implies  $B$ . The two main concepts that determine the outline of the association rules are: 1) *Confidence* - confidence refers to the probability that if the antecedent is true then the precedent will be true. High confidence means that this is a rule that is highly dependable. 2) *Support* - support refers to the number of records in the database that the rule applies to. High support means that the rule can be used very often. In some literature, confidence is called the accuracy of the rule and support is called the coverage.

The set of the suggested groups of quality problems,  $\{Q_i\}_{i=1}^n$  resulted from the first stage of analysis will be determined for a sample of synthetic data set in Table 1. However, the same procedure is identical for any set of data with different recodes size. This data are the quality problems (defects) in EDM Fast Hole Drilling process. This process aims to drill an array of holes in a hard metal material within certain dimensional specification. Figure 1 shows a sample product that has been manufactured by using this process. The EDM Fast Hole Drilling is a fast and economical way to produce a hole in hard material. However, this non-traditional manufacturing process is associated with many complexities that lead to a high likelihood for product defects to occur. A list of possible quality problems is as follows: 1) Out of size diameter, 2) out of size thickness, 3) out of tolerances, 4) poor surface finish, 5) lack of circularity, 6) lack of irregularities, 7) lack of cylindricity, 8) poor contours, 9) poor edge conditions, 10) lack of perpendicularity, 11) incorrect number of holes, 12) lack of flatness, 13) off spacing, and 14) lack of parallelism.

It is clear that it is infeasible to assign a single quality problem to be solved by QIT - in the above case there will be approximately 14 QIT's. Therefore, grouping the related quality problems and assigning them as a group to the QIT is more desirable. Then an enhancement of the assignment process in from of an optimization model for costs minimization will be implemented.



**Figure 1. A part manufactured by the EDM fast hole drilling**

We have records of 15 production batches from the quality control department database. In each record, we use 1 to indicate that a quality problem has been detected and 0 to indicate that the batch is defects free. There are a large number of possible association rules that could be generated out of the data set in Table.1. Examples of some the association rules extracted from the database are:

**Rule1:** - If an out of size diameter defects was occurred in the production batches, then Out of Tolerances defect was also occurred. This happens with an confidence =71% and a support = 50%

**Rule2:** - If a lack of parallelism defect is found, then a lack of perpendicularly defect occurred. This happens with an confidence =100% and a support = 20%

**Rule3:** - If a poor edge conditions and poor contours are founded in a production batch, then an off Spacing problem will not occur. This happens with an confidence =71% and a support = 50%.

Defect \ Batch No.	Out of Size Diameter	Out of Size Thickness	Out of Tolerances	Poor Surface Finish	Lack of Circularity	Lack of Irregularities	Lack of Cylindricity	Poor Contours	Poor Edge Conditions	Lack of Perpendicularity	Incorrect Number Of Holes	Lack of Flatness	Off Spacing	Lack of Parallelism
1111	1	1	1	0	0	0	0	0	0	1	0	0	0	1
1112	0	0	0	1	0	1	0	0	0	1	0	0	0	1
1113	0	0	0	0	0	0	0	1	1	0	0	0	0	0
1114	0	0	0	1	0	0	0	0	0	0	1	0	0	0
1115	0	0	0	0	0	0	0	0	0	0	0	1	1	0
1116	0	0	0	0	1	0	1	0	0	0	0	0	0	0
1117	0	0	0	0	1	0	1	0	0	0	0	0	0	0
1118	0	0	0	0	1	0	0	1	1	0	0	1	1	0
1119	1	1	1	0	0	0	0	0	0	0	0	0	0	0
1120	1	1	1	0	0	0	0	0	0	0	0	0	0	0
1121	1	1	0	0	0	0	0	0	0	0	0	0	0	0
1122	1	1	0	0	0	0	0	0	0	0	0	1	1	0
1123	1	1	1	0	0	0	0	0	0	1	0	0	0	1
1124	0	0	0	1	0	1	0	1	1	1	0	0	0	0
1125	0	0	0	1	1	1	1	0	0	1	0	0	0	1

Table 1. Synthetic data set

In order to make the analysis more efficient, the following guidelines are suggested for selecting the most significant association rules:

- 1) A significant association rule should have a confidence percentage higher than 65%,
- 2) a significant association should have support percentage higher than 20%,
- 3) we assume that each single quality problem is a suggested group of quality problems that has one element in the group,
- 4) eliminate the association rules that have statement about a non occurring of quality problems, i.e. all quality problems in the association rule should have indicator functions equal to 1. For example, we eliminate Rule3 in the above list since off spacing indicator function value is 0.

When the synthetic data set is used and the above guidelines are applied then the significant association rules obtained. These rules are

- Rule1:** {out of size diameter, out of tolerances} with [coverage =66.7%, support = 26.7%]
- Rule2:** {lack of parallelism, lack of perpendicularity} with [coverage =80%, support = 26.7%]
- Rule4:** {out of size diameter, out of size thickness, out of tolerances} [coverage =80%, support = 26.7%]
- Rule5:** {out of size diameter, out of size thickness} [coverage =100%, support = 40%]
- Rule6:** {out of size thickness, out of tolerances} [coverage =66.7%, support = 26.7%]
- Rule7:** {poor surface finish, lack of irregularities} [coverage =77%, support = 20%]
- Rule8:** {lack of circularity, lack of cylindricity} [coverage =77%, support = 20%]
- Rule9:** {poor contours, poor edge conditions} [coverage =100%, support = 20%]
- Rule10:** {lack of flatness, off spacing} [coverage =100%, support = 20%]

After we determine the significant association rules, we added for each corresponding to these association rules a group of quality problems that has the elements of the association rules. For instance, since Rule1 suggests that if an out of size diameter defects occurs, then an out of tolerance also occurs, then we generate a group of quality problems that has two defects: out of size diameter and out of tolerances. Finally, we assigned  $Q_{15}$  for this group. Table 2 shows all the suggested groups of quality problems. In the next section, a refinement for the suggested quality problem groups is presented.

#### Other data mining techniques

The association rules technique works efficiently when enough resources can be allocated for creating a large number of the QIT's. In the above example, with 65% confidence and 20% support, there are 10 significant rules. However, in the case where there is a limitation on the resources or where there is a large number of quality problems with a vast number of database recodes, introducing more efficient and complex data mining techniques becomes essential in reducing the number of suggested groups. For example, the top management in the industrial firm may inform the quality insurance department that the maximum allowable number of QIT's is 5; hence association rules technique does not have full control to suggest a decision support system for grouping within this constraint. One possible data mining technique under these circumstances could be data clustering or data classification. This technique attempts to find groups of items that are similar. Since we are dealing with binary data, special data clustering techniques are involved. Potential methods are listed here:

- i. Small-Large ratios based algorithm for market basket data (Yun, Chuang and Chen, 2001).
- ii. Multi-variate Bernoulli model for naive Bayes text classification (McCallum and Nigam, 1998).
- iii. K-means approach for clustering binary data (Ordonez, 2003).

The second and third approaches are more suitable in the case where the number of QIT's is predetermined by the management.

#### The Second Stage: The Optimization Model

It is known that there are costs involved in finding, preventing and failing to catch poor quality. According to the American Society for Quality, there are four types of quality costs: *prevention costs*, *appraisal costs*, *internal failure costs*, and *external failure costs*.

Prevention costs are the costs of planning and implementing a quality system in an industrial firm to prevent poor quality. These costs may include: product and process redesign, human resources, quality control software, training, auditing, and performing cause-effect analysis. In addition, there are appraisal costs that are acquired in the process of finding and evaluating quality problems. Examples of the appraisal costs are incoming raw materials testing, in-process product testing, and providing instrumentation for measuring. In this paper we consider that these two types of costs are lower when the QIT manage a group of quality problems compared to managing each quality problem individually. This is true due to many reasons such as:

- Buying quality control software and training one QIT for analyzing and solving more than one quality problem is more cost effective than training multiple QIT for each quality problem. This is supported by the fact that each assigned group of quality problems for the QIT are related and most likely similar problems.

- Less paperwork and administration activities since there are fewer QIT's.
- Sampling and testing will likely be more efficient if the QIT investigates a group of quality problems. A worker who takes a reading for one variable could take another variable reading at the same time or place without the need for another trip for the second variable.

Internal failure costs are the costs that result from poor quality which leads to the failure of products to meet the requirements before they transfer to customers. Examples are the reworking and scrapping of defective products. Usually, the reworking or scrapping of defected products will be based on the quality control sampling technique. A specific sampling procedure has a particular set-up that reflects the quality control theme. We assume that the internal failure costs when a quality management team investigates a group of quality problems is less expensive compared to the internal failure costs when the quality management team investigates the same quality problems individually. The rationale behind this is as follows: for two quality problems  $A$  and  $B$  that are being investigated at the same time by one QIT, the scrap/rework rate will be lower than the sum of the scrap/rework rates of  $A$  and  $B$  when they are investigated individually. In the first case, both nonconforming qualities may occur simultaneously in one product, so the quality assurance people firm will scrap or rework only one product. However, in the second case, there are two products which will be scrapped or reworked each time a nonconforming product occurs for quality problem  $A$  and quality problem  $B$ .

*External failure costs* are the costs that arise after products are transferred to customers. Examples of such costs are customer service costs, repair costs, replacement costs, and warranty costs. The external failure costs will be used to assess the performance of the quality improvement process in this paper. We assume that assigning more than one quality problem for each QIT will lower the performance of the quality improvement process compared to assigning only one quality problem at a time. The rationale behind this is that the QIT's efforts are not as focused on the first case as compared to the second case. As a consequence, the external failure rate in the forms of repair costs and warranty costs will be higher in the first case.

The external failure costs are estimated out of the information system of the industrial firm that keep database for all previously acquired costs due to repair, return, warranty... etc. On the other hand, the other types are estimated based on the management experience. A summary of the quality problems' parameters is listed in Table 2.

Finally, the binary integer optimization model is a costs minimization problem where the total costs associated with having the group  $Q_i$  of quality problem is the sum of the four quality costs mentioned above, i.e. total costs:  $T(i) = P_i + A_i + I_i + E_i$ . Mathematically, the optimization model is formulated as follows:

$$\text{Minimize } \sum_{i=1}^n (P_i + A_i + I_i + E_i) \times I_{Q_i}$$

Subject to:

$$I_{Q_i} = 0 \text{ or } 1 \quad i=1, 2, \dots, n$$

$$\sum_{i=1}^n (q_{ij} \times I_{Q_i}) \geq 1 \quad j=1, 2, \dots, m$$

The second constraint ensures that each quality problem is a candidate problem to be solved by the QIT's.

## THE MODEL SOLUTION

The optimization problem in the above section is called a set covering problem. A heuristic solution is available in (Chvatal, 1979), in addition, (Agard and Kusiak, 2004) used the same algorithm to solve their subassembly optimization problem. However, we will implement the Modeling Language for Mathematical Programming (AMPL) as a mathematical modeling language technique to solve this problem and to originate the optimal quality problem improvement scheme. Furthermore, we used MINOS as a nonlinear optimization solver for the model. The AMPL program is presented in Figure 2 and the results are presented in Table 3. It is clear that required number of QIT's is reduced from 14 to 9 QIT's. Moreover, the expected quality costs are reduced from \$216,000 when each quality problem is being investigated individually to \$180,000 when some related problems are being investigated together (a 16.67% reduction).

Candidate Quality Problems Group	Out of Size Diameter	Out of Size Thickness	Out of Tolerances	Poor Surface Finish	Lack of Circularity	Lack of Irregularities	Lack of Cylindricity	Poor Contours	Poor Edge Conditions	Lack of Perpendicularity	Incorrect Number of Holes	Lack of Flatness	Off Spacing	Lack of Parallelism	Out of Size Diameter and Size Thickness	Out of Size Diameter+ Out of Tolerances	Out of Size Thickness+ Out of Tolerances	Out of Size Diameter, Thickness, and Out of Tolerances	Poor Surface Finish+ Lack of Irregularities	Lack of Circularity+ Lack of Cylindricity	Poor Contours + Poor Edge Conditions	Lack of Parallelism+ Lack of Perpendicularity	Lack of Flatness+ Off Spacing
$Q_i$	$Q_1$	$Q_2$	$Q_3$	$Q_4$	$Q_5$	$Q_6$	$Q_7$	$Q_8$	$Q_9$	$Q_{10}$	$Q_{11}$	$Q_{12}$	$Q_{13}$	$Q_{14}$	$Q_{15}$	$Q_{16}$	$Q_{17}$	$Q_{18}$	$Q_{19}$	$Q_{20}$	$Q_{21}$	$Q_{22}$	$Q_{23}$
$P_i^*$	10	9	9	5	5	4	6	8	6	3	6	2	2	2	15	14	12	10	5	6	11	4	3
$A_i^*$	6	5	4	3	3	3	5	5	4	2	3	2	1	2	7	6	10	10	3	5	5	2	2
$I_i^*$	3	3	3	2	3	2	3	3	2	1	2	2	1	2	5	5	4	4	2	3	3	2	2
$E_i^*$	4	4	5	4	3	5	4	4	3	6	7	3	2	5	21	23	24	16	17	13	18	10	15
$T(i)^*$	23	21	21	14	14	14	18	20	15	12	18	9	6	11	46	48	50	40	27	27	37	18	22

\*In thousands

Table 2. Parameters for each group of quality problems.

```

AMPL program
Set I; %set of group of problems
Set J; %set of quality problems
Param P{I}>=0; %prevention costs
param A{I}>=0; %appraisal costs
param I{I}>=0; %internal costs
param E{I}>=0; %external costs
param q{I,J}>=0; %indicator function of I in J
var IQ{I} binary; %the decision variable
Minimize cost: sum{i in I} (P[i]+A[i]+I[i]+E[i])* IQ[i];
Subject to c{j in J}: sum {i in I} (q[i,j]*IQ[i])>=1;
    
```

Figure 2. AMPL program

Variable	Value	Variable	Value
$I_{Q1}$	0	$I_{Q13}$	1
$I_{Q2}$	0	$I_{Q14}$	0
$I_{Q3}$	0	$I_{Q15}$	0
$I_{Q4}$	0	$I_{Q16}$	0
$I_{Q5}$	0	$I_{Q17}$	0
$I_{Q6}$	0	$I_{Q18}$	1
$I_{Q7}$	0	$I_{Q19}$	1
$I_{Q8}$	1	$I_{Q20}$	1
$I_{Q9}$	1	$I_{Q21}$	0
$I_{Q10}$	0	$I_{Q22}$	1
$I_{Q11}$	1	$I_{Q23}$	0
$I_{Q12}$	1		

Table 3. AMPL decision variables output

**CONCLUSION**

To sum up, the methodology in this paper aims to optimize the quality improvement process and to investigate the relations among poor quality causes. Particularly, the methodology assigns the right QIT's with the accurate number of specialized members to solve the right type and right number of related quality problems with minimum quality costs. Association rules technique works well in classifying the quality problems that occurs together. However, the non-significant rules could be eliminated by using elimination guidelines rather than what we proposed. This will depend on the type of the quality problem and on the goals of the quality improvement process. Moreover, *On-line quality control charts* produce a massive database;

hence a more advance and complex data mining technique becomes necessary to find the related quality problem. We suggest in this case a cluster (segmentation) analysis as an alternative to the association rule technique to be implemented for quality problems grouping. Finally, manufacturing information system could be implemented more efficiently in the second stage of the proposed methodology to obtain a better estimation of the quality costs.

#### ACKNOWLEDGEMENT

The authors acknowledge the support of NSF grant ESP-0091900 (Professor F. Fred Choobineh, P.I.).

#### REFERENCES

1. Agard, B. and Kusiak, A. (2004) Data mining for subassembly selection, *Journal of Manufacturing Science and Engineering*, 126, 627-631.
2. Al-Salim, B. (2005) Mass customization of travel packages: data mining approach, *working paper*, University of Nebraska-Lincoln.
3. Chvatal, V. (1979) A greedy heuristic for the set-covering problem, *Mathematical Operation Research*, 4, 3, 233–235.
4. Kusiak, A. and Kurasek, C. (2001) Data mining of printed-circuit board defects, *IEEE Transactions on Robotics and Automation*, 17, 2, 191-196.
5. Last, M. and Kandel, A. (2001) Data mining for process and quality control in the semiconductor industry, in Dan Braha (Ed.) *Data Mining for Design and Manufacturing: Methods and Applications*, The Netherlands, Kluwer Academic Publishers, 207-234.
6. Lee, J. and Park, S. (2001) Data mining for high quality and quick response manufacturing, in Dan Braha (Ed.) *Data Mining for Design and Manufacturing: Methods and Applications*, The Netherlands, Kluwer Academic Publishers, 179-205.
7. McCallum, A., Nigam, K.(1998) A comparison of event models for naive bayes text classification, *AAAI-98 Workshop on Learning for Text Categorization*.
8. Mitra, A. (1998) *Fundamentals of quality control and improvement*, Prentice Hall, Upper Saddle River, New Jersey.
9. Oh, S., Han, J. and Cho, H. (2001) Intelligent process control system for quality improvement by data mining in the process industry, in Dan Braha (Ed.) *Data Mining for Design and Manufacturing: Methods and Applications*, The Netherlands, Kluwer Academic Publishers, 289-309.
10. Ordonez, C. (2003) Clustering binary data streams with K-means, *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, June 13, San Diego, CA, USA, ACM Press, 12-19.
11. Post, G. (2005) *Database management systems: designing and building business applications*, McGraw-Hill/Irwin.
12. Steckenrider, j., Guha, S., Sethuraman, A., Ra, Y. and Kim H. (2001) Classifying defects for copper CMP process modules, *Micro Magazine*. Available online at <http://www.micromagazine.com/archive/01/09/brave.html>.
13. Yun, C., Chuang, K. and Chen, M.(2001) An efficient clustering algorithm for market basket data based on small large ratios, *Proceedings of the 25th International Computer Software and Applications Conference on Invigorating Software Development*, October 08–12, Chicago, IL, USA, IEEE Computer Society, 505–510.