Journal of Industrial Engineering 4 (2009) 37-50

A Proposed Data Mining Methodology and its Application to Industrial Procedures

Seyyed Soroush Rohanizadeh[a,*], Mohammad Bameni Moghadam[a]

[b]*Department of management and accounting, Islamic Azad University, Qazvin Branch, Qazvin, Iran*

Received 2 May., 2009; Revised 18 May., 2009; Accepted 15 Jun., 2009

Abstract

Data mining is the process of discovering correlations, patterns, trends or relationships by searching through a large amount of data stored in repositories, corporate databases, and data warehouses. Industrial procedures with the help of engineers, managers, and other specialists, comprise a broad field and have many tools and techniques in their problem-solving arsenal. The purpose of this study is to improve the effectiveness of these solutions through the application of data mining. To achieve this objective, an adaptation of the engineering design process is used to develop a methodology, specifically designed for industrial procedures' operations. This paper concludes by describing some of the advantages and disadvantages of the application of data mining techniques and tools to industrial procedures; it mentions some possible problems or issues in its implementation; and finally, it provides recommendations for future research in the application of data mining to facilitate decisions relevant to industrial procedures.

Keywords: Data Mining, Industrial Procedures, Methodology, Database, Artificial Intelligence, Neural Networks, Decision Tree.

1. Introduction

Thousands of businesses are using data mining applications every day in order to manipulate, identify, and extract useful information from the records stored in their databases, data repositories and data warehouses. With this kind of information, companies have been able to improve their businesses by applying the patterns, relationships, and trends that have lain hidden or undiscovered within colossal amounts of data [1]. Manufacturing, also, is one of the new fields in which data mining tools and techniques are beginning to be used successfully; Process optimization, job shop scheduling, quality control, and human factors are some of the areas in which data mining tools such as Neural Networks, Genetic Algorithms, Decision Trees, and Data Visualization can be implemented with great results [10, 13]. However, implementation of these data mining techniques is inconsistent in practice; because software vendors propose different and proprietary approaches that focus on specific business applications. These approaches even use different sets of analysis tools [7]. To develop good data mining strategies, industrial specialists require an application-neutral methodology. Moreover, they should keep in mind that What is needed is a guide through the maze of tools and approaches to the myriad of applications.

This Paper Contains: First, in section 2, revision of some important cases of data mining projects in some industrial procedures. Section 3, gives a general description of the research methodology. Section 4, presents the proposed methodology; and finally section 5, summarizes the major conclusions of this document and also states possible areas of further researches.

1.1. Background

Data mining is often described as the process of discovering correlations, patterns, trends or relationships by searching through a large amount of data stored in repositories, corporate databases, and data warehouses. Humans, in that sense, are limited by information overload; thus, new tools and techniques are being developed to solve this problem through automation. Data mining uses a series of pattern recognition technologies and statistical and mathematical techniques to discover the possible rules or relationships that govern the data in the databases. Data mining must also be considered as an iterative process that requires goals and objectives to be specified [9]. Once the intended goals are completely

*Corresponding author E-mail: soroosh.rohany@gmail.com

defined, it is necessary to determine what data is available or can be collected. Data mining also involves a methodology for implementation. The methodology, or structured approach, usually varies from vendor to vendor. SAS Institute [2], for example, promotes SEMMA (sample, explore, modify, model and assess). Another methodology is CRISP-DM by SPSS, Inc [5]. Each methodology strives to help users obtain the best data to provide the most responsive information to address their needs. The evolution and development of finding the right data for decision-making are shown [2] in Figure 1, and their descriptions follow.

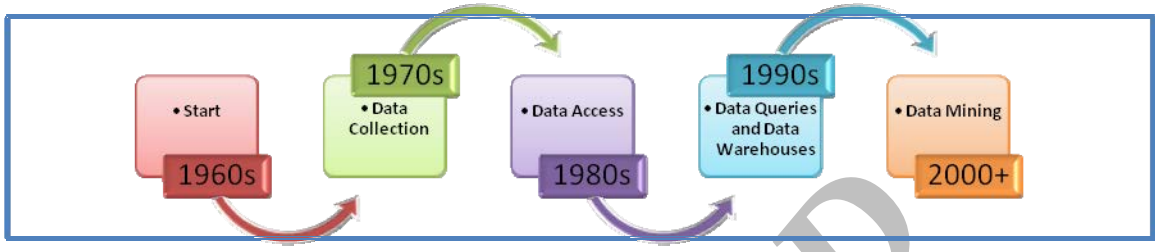

Fig .1. Data Mining Evolution

Data Collection: During the late '60's, simple reports of pre-formatted information were created from data stored in databases.

Data Access: In the 1980s, users began to want information more frequently and they wanted it to be more individualized. Thus, they began to make queries, or informational requests, of the databases.

Data Queries. Later, in the 1990s, users required immediate access to more detailed information that responded to "on the fly" questions. They wanted information to be "just-in-time" to correlate with their production and decision-making processes. At this stage, users began to write their own queries to extract the information that they needed from the database.

Data Mining. In the last few years, users began to realize the need for more tools and techniques in order to identify and find relationships in data so that the information obtained was more meaningful for their applications. Additionally, companies recognized that they had accumulated volumes of data; and, as a result, they needed new tools to sort through it all and meet their informational needs. The next step is to exploit these tools for meaningful applications.

1-1- Limitations of the Study

This study focuses specifically on applying data mining to problems generally addressed in industrial procedures. Only the information necessary to illustrate the concepts described in this document, have been included. The methodology proposed in this study is an abstract and functional framework which is a conceptual model, and it has not been implemented yet; therefore it has not been executed or tested. That task remains for the future.

1.2. Problem Statement

Data mining not only involves a collection of systems, solutions or technologies, but also includes a structured process in which human interaction is important. Humans decide if the patterns discovered have some relevance to the problem at hand or if they justify further study and exploration. With this in mind, data mining approaches have been integrated with the needs and interests of specific businesses. Data mining techniques can be used in many different fields and have many applications. In order for data mining techniques to provide the intended results—full exploitation of all available data--it is very important that the data is correctly prepared and collected for its specific applications. With so many choices on the market, users need assistance in deciding the various tools offered by the many vendors in the market. Thus, applications of data mining in areas such as quality control, process control, human factors, material handling, maintenance and reliability in production systems should be studied and addressed in more detail. Consequently, this research proposes to develop a convenient methodology for the application of data mining in industrial procedures base on analyzing and comparing the different tasks and techniques used in data mining and identifying the main advantages and disadvantages of data mining techniques and tools in industrial procedures' application.

2. Literature Review

The application of data mining techniques to industrial procedures is an area that holds promise, but that is currently underdeveloped [19]. Data mining can, however, be strategically applied to industrial procedures processes such as scheduling, quality control, cost reduction, safety, and others. This section outlines some of the data mining techniques and applications that can be utilized by industrial specialists.

2.1. Data Mining Techniques

There are a number of techniques used in data mining, but not all of them can be applied to all types of data. Neural network algorithms, for example, can be used to quantify data (numerical data), but they cannot qualify data precisely (categorical data); [10, 19]. For that reason, one single technique cannot be used to perform a complete data mining study and each technique has its own scope of applications. Some of the techniques applied in data mining are:

Traditional Statistics: Some of the traditional statistical methods that can be used for data mining are [9]: *Cluster analysis* (or segmentation), which involves separating sets of data into groups that include a series of consistent patterns. *Discriminant analysis*, finds hyper planes that separate classes so that users can then apply them to determine the side of the hyper plane in which to catalogue the data. *Logistic regression* is primarily used for predicting binary variables and, less frequently, multi-class variables. Finally, *time series forecasting*, predicts "unknown future values, based on time varying series of predictors".

Induction and Decision Trees: These techniques try to uncover associations in the data. They search for similarities within the existing records and try to infer the rules that express those relationships [9]. Decision trees are flow charts--tree structures in which nodes represent tests or attributes, branches represent test outcomes, and leaf nodes represent classes or class distributions.

Neural Networks: Neural networks can analyze imprecise, incomplete, and complex information and find important relationships or patterns from this information; by their special ability to "learn". Usually the patterns involved in this kind of analysis are so complicated that they are not easily detected by humans or by other types of computer-based analysis.

Data Visualization: Through using visual tools, analysts can reach a better understanding of the data because they can focus their attention on some of the patterns found by other method. Using variations of color dimensions, and depth, may lead to find new associations and improve the differentiation between them [19].

2.2. Data Mining Tasks

Data mining can be used in many different ways. Some of the tasks most commonly found are [9]:

Description and summarization which involves the study of data in order to find its major and most important characteristics. The most common techniques applied in this task are the basic descriptive statistical models and data visualization (histograms, box plots, scatter plots). The main goal of *concept descriptions* is to describe data classes or subgroups and to point out important concepts, characteristics and parts that may facilitate the process of understanding them. *Clustering and induction methods* are usually employed in concept description. *Segmentation* is mainly used for sorting data into a series of unknown different classes or subgroups that share the same characteristics, but that are different from each other. The techniques frequently used in segmentation include clustering, neural networks, and data visualization. *Classification* is a task that is very similar to segmentation, and the major difference between them is that classification assumes classes and subgroups which are known. It employs techniques such as discriminant analysis, induction and decision trees, neural networks, and genetic algorithms. *Prediction models* try to find or forecast an unknown continuous value corresponding to a specific class. Prediction models are usually built using techniques such as neural networks, regression analyses, regression trees, and genetic algorithms. *Dependency analysis* describes all the important and significant dependencies among the data elements. Two special cases of dependency analysis are particularly valuable for data mining: association and sequential patterns [9].

2.3. Data Mining Applications in Industrial Procedures

Because data mining techniques search through large amount of data in order to discover correlations, patterns, rules or relationships, they can be applied in many different fields [8]. While the use of data mining in industrial procedures is not widespread, several successful applications of data mining in fields related to industrial procedures have been reported as follow.

Quality Control: Data mining has been applied in some Statistical Quality Control (SQC) software packages as an integral part of decision support tools used in the analysis of process behavior. Additionally, data mining techniques have been applied to analyze and detect possible defects and their corresponding causes in the fabrication of semiconductors [3]. Data mining has also been applied in predicting defects for the papermaking industry [15]. Furthermore, companies such as Daimler Chrysler have used data mining to evaluate warranty claims in order to identify high quality patterns, as well as the key factors that give rise to claims, improving customer satisfaction and the reliability of its products [13].

Scheduling: This area has been an important area for applying data mining techniques. For example, schedules for job shop operations have been created using rules extracted with data mining analysis over schedules generated by genetic algorithms [12]. Additionally, quality tests have also been scheduled with a data mining approach. Using decision tree models and mining the data provided by a MRP system in a factory of hydraulic pumps, new and improved schedules have been generated [18]. In general, the number of operators and work stations assigned to a specific order or task could be improved through the use of rules and models generated

by data mining applied to historical data such as throughput, operations performance, and completed orders.

Process Optimization: In integrated circuit manufacturing, yield improvement has been considered a suitable application for data mining techniques to address the problem of low yield analysis [14]. Data mining can also be use to reduce rework in process. For example, in exploring data of a specific production process or product line, it is possible to find rules identifying the best settings and conditions to achieve more throughputs, to reduce cost, or to reduce waste.

Process Control: Process control, monitoring, and diagnosis are other important areas in which data mining analysis can be effectively applied. For example, long performance deterioration in processes can be studied using historical data to identify its major factors [21]. Historical process logs can be analyzed to monitor the process at different stages.

Safety: Regarding safety, data mining studies in road traffic accidents have already created classification models and identified influential factors for accident severity [16]. Hazardous elements such gasses or radiation levels have also been monitored to protect and extend human lives. Moreover, data mining techniques used to analyze occupational accidents and disease records have revealed important patterns that can be used to reduce occupational risks [19].

Cost Reduction: Data mining can be effectively be applied to cost reduction. A good example of an application of data mining techniques to reduce cost in products with high customization, for example, is analyzing sales and product options to identify the ones with greater demand [13]. The products with same and most common options can be manufactured together to reduce cost and inventories.

Maintenance and Reliability: Data mining techniques can also be applied to identify combinations of plants, machines, workstations and products that have higher breakdown or malfunction rates, or to find repairs that are likely to occur together or in close time proximity, or to report problems that often precede specific repairs [2]. With this information, preventive maintenance can be performed in parts or components that are identified as having a similar time between failures, reducing downtimes for repairs and their corresponding costs.

Product Development: Other applications of data mining include product development and design [4]. Data mining can be used to extract relationships between design requirements and manufacturing specifications to explore the different tradeoffs between overlapping activities and coordinating costs. Data mining can also be performed in historical data to reduce inventory costs for new products, to analyze suppliers and delivery times, and to select materials for the manufacturing process.

2.4. Problems in Making Effective decisions in Data Mining

Some authors, such as Koonce and Fang [11], acknowledge that industrial specialists can use data mining to explain the behavior of complex systems. They also have established the need for further studies related to applications of data mining to job shop scheduling systems [12]. Furthermore, Bertina and Catania also found in their study of data mining applications for wafer manufacturing [3] that it is difficult to select specific data mining techniques. They recognized the need for general methodologies and guidelines that could support users in the development of data mining applications. Selection is not the only difficulty with applying data mining. Other problems in making effective decisions are that companies and organizations store great amounts of data and information that are very difficult and time-consuming to analyze by traditional means. Moreover, there are many elements to consider in selecting data mining tools [6]. Finally, data mining is the result of the confluence of multiple disciplines [9]. It thus requires much specialized knowledge to make the right decisions. Many different fields (e.g., statistics, databases, artificial intelligence, and software development) contribute to data mining (See Figure 2).

Fig. 2. Different Fields affecting the Data Mining Process

3. Research Methodology

The engineering design process is based on the scientific approach to problem solving. By this approach it is possible to implement corrective solutions that take the form of new or improved systems. The engineering design process, as described by Landis [13], was used in the execution of this study and is depicted in Figure 3, which its six steps are detailed below.

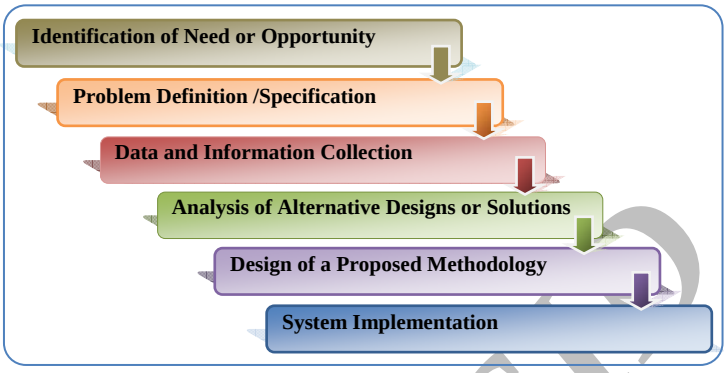

Fig. 3. The Engineering Design Process Applied

3.1. Identification of a need or opportunity

This step is the first step in problem, which is so extensive and varied for industrial procedures. But because of the variety of data and information, deciding on the most effective data mining techniques and systems can be complicated. As noted above, there are many different software vendors with many different data mining software applications; each promotes its own data mining methodology. Data mining, like industrial procedures, is the result of the confluence of multiple disciplines and for that reason, the process of implementing data mining process in industrial procedures is difficult and requires much specialized knowledge.

3.2. Problem Definition

There are so many options, tasks, techniques, tools, formats, and approaches to data mining that industrial specialists find it very difficult to design and implement projects. Although methodologies already exist, they are designed for specific software packages. Most of these methodologies use a traditional statistical approach. It is still not clear that this approach to data mining is sufficient for obtaining the vast array of data needed for industrial procedures applications. Thus, a data mining methodology to meet the specific requirements of industrial procedures is needed.

3.3. Data and information Collection

In order to accomplish this study, surveys, analysis, reviews, and comparisons of data mining applications were collected and studied. These were based on several vendors' information and case studies. One relevant survey was sent to more than 80 different companies of data mining software over the Internet. There were 30 responses. The survey asked companies whether their product had been or could be used in industrial procedures' applications. It also asked whether they had applied or sold their data mining products for the implementation of projects related to industrial procedures areas such as quality control, scheduling, manufacturing, safety, or ergonomics. Other questions were related to hardware requirements and prices. The most relevant results of this survey are shown in Figures 4 and 5. Figure 4 shows that approximately 60% of the companies have either sold their product for industrial procedures' applications or believe their product is applicable for industrial procedures. The survey also asked about costs, because the cost of data mining may prevent some companies from using it even though it could benefit them. Figure 5 show that the average cost is approximately $5,000, a feature that might be prohibitive for smaller companies. Thus, design restrictions and cost appear to be key factors that affect the use of data mining in industrial procedures.

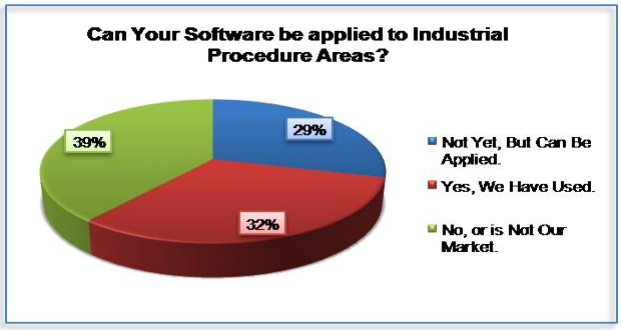

Fig. 4. Application of Data Mining Software to Industrial Procedure Areas

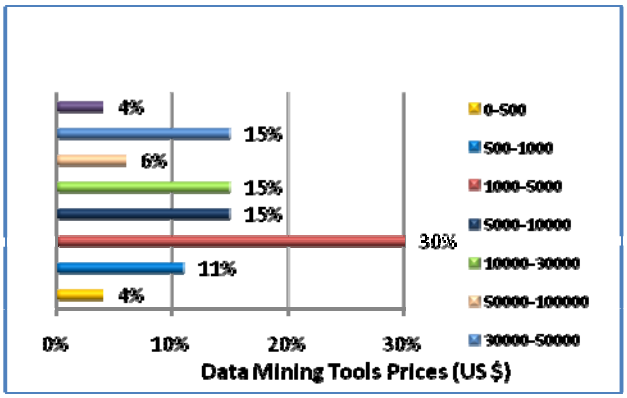

Fig. 5. Data Mining Software Price Distribution

3.4. Analysis of Alternatives

There are several different data mining methodologies, but there is no one standard methodology for applying data mining. Consequently, several vendors have created their own proprietary methodologies. Software vendors have designed approaches that are strongly correlated with the design of their own solutions and software packages. A related methodological issue is that data mining has been considered as a kind of art in which each analyst may follow his or her own "recipe" or form [7]. Two popular methodologies are SEMMA and CRISP-DM. They are described in the following sections.

SEMMA: SEMMA is the methodology for data mining processes proposed by the SAS Institute--one of the most important companies that develop statistical software applications--with the software package Enterprise Miner [2]. In SEMMA, SAS offers a data mining process that consists of five steps: sample, explore, modify, model, and assess. This methodology begins by analyzing a small portion of a large data set. The next step is to explore the data and the information by looking for trends and anomalies in the data with the purpose of gaining some information about the data. In the third phase, data is modified to create, select, and transform the variables for the study. A valid model is then created using the software tools, which search automatically for combinations of rules and patterns that reliably predict the observed results. Finally, the last step of the SEMMA methodology consists of evaluating the usefulness and reliability of the findings. Although the SEMMA methodology contains some of the essential elements of any data-mining project, it concerns only the statistical, the modeling, and the data manipulation parts of the data-mining process. It lacks some of the fundamental parts of any information systems project, including analysis, design, and implementation phases. But, the SEMMA methodology does not consider the roles of the organization and the stakeholders during the project; it does not see data mining as an integral element within a systems perspective. Also, SEMMA is specifically designed to work with the Enterprise Miner software, the data mining software of the SAS institute; therefore, it cannot be applied outside the limitations of that system.

CRISP-DM: Another data mining methodology is CRISP-DM [5] (cross-industry standard process for data mining). It was developed by a consortium of data mining vendors and companies through an effort founded by the European Commission. The CRISP-DM methodology describes a data-mining project as a 6-phase cycle in which the sequence of phases is not rigid. The phases that CRISP-DM considers are business understanding, data understanding, data preparation, modeling, evaluation, and deployment. This approach includes in its first phase very important elements such as the business's objectives, requirements, constraints, and resources available for the project, in order to establish the data mining goals. Good documentation is also promoted from the beginning of the plan. CRISP data preparation comprises the selection, cleaning, construction, integration, and formatting that data requires in order to create any model. However, it also assumes that all the information required is already available and continues to be valid, so new data should not be collected. CRISP-DM methodology also emphasize that data must sometimes be divided into training and validation sets. After building the models with the training data, a validation test is then used to ensure that the obtained model behaves with adequate fidelity to the real system. Finally, in CRISP methodology, after a satisfactory creation of models is done, the evaluation phase continues with an analysis of the results, a review of the process, and the final deployment phases. The CRISP deployment phase consists in the creation of a deployment plan, a monitoring and maintenance plan, a final report, and the final review of the project. Techniques under CRISP-DM may be applied because they are incorporated in the tools available for the organization and not because they are really needed. Therefore, results from this approach may not properly correspond to the organization's main objectives, and the models generated this way may not truly represent the behavior of the entities for which the study was intended in the first place. Another problem with the approach suggested by CRISP is that the selection of the technique is delayed until the modeling phase; if the data required is not available or is in the wrong format, the model has to return to the data analysis phase again. CRISP-DM, indeed, stresses in the importance of assessing tools and techniques early in the process but also affirms that the selection of tools may influence the entire project [5]. Furthermore, Techniques should be selected according to an organization's goals and requirements and should not depend only on the data available. Besides the difficulties that the CRISP-DM

methodology presents, it is a good approach to the general process of data mining.

3.5. Design of a Proposed Methodology

While SEMMA and CRISP-DM are still useful methodologies, but they have some deficiencies, and may not be suitable methodologies for industrial procedures' purposes. Therefore, by understanding the needs of industrial procedures, studying the common methodologies in this section, and also the application of information system analysis and design structure, a proposed methodology for the application of data mining in industrial procedures, composed and described in the next section.

4. A PROPOSED METHODOLOGY

Engineers follow a structured approach to problem-Solving. This enables them to duplicate results or determine where errors have occurred in the process. As a result, they may have confidence in the solutions they recommend. For that reason, this study offers a methodology for using data mining in solving problems related to industrial procedures. This structured approach should lead analysts through the steps required in obtaining the data needed to provide information required for problem-solving. This approach has a number of steps:

a. Analyze the organization.
b. Structure the work.
c. Develop the data model.
d. Implement the model.
e. Establish on-going support.

These steps are shown in Figure 6 and described in detail in the following.

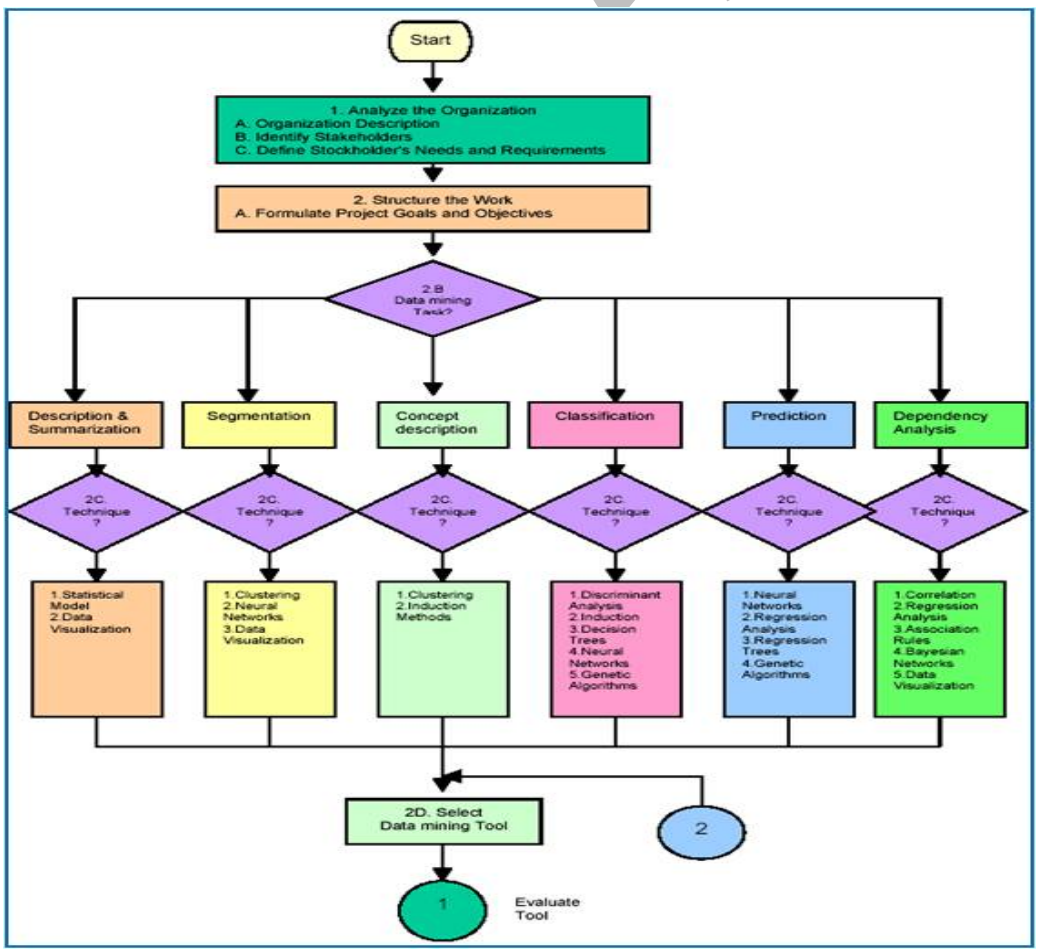

Fig. 6. Proposed Methodology

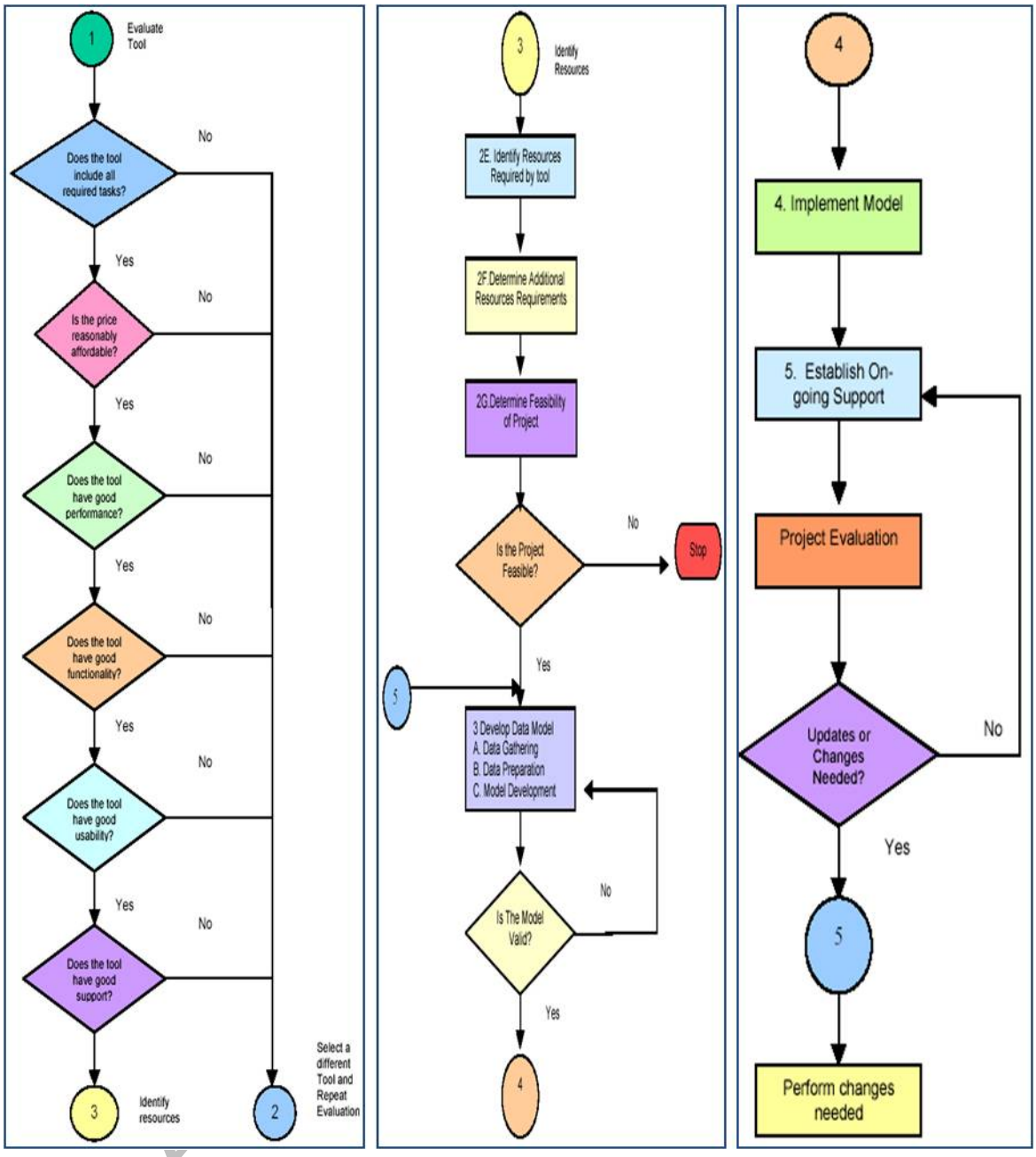

Figure 6 Continued.

a. Analyze the Organization

Organization Description: The first step of any data mining project is to understand the purpose of its existence. When a data mining project is conducted in an organization, a study of the organization's goals, objectives, and strategies is required in order to understand the purpose of the project. This enables the analyst to determine the best way to execute the project so that it will empower and facilitate the achievement of the business's targets. Failing to understand the organization's needs before implementing the project may cause its results to be incompatible with or of no use at all to the organization.

Define Stakeholders' Requirements and Expectations: Before identifying the requirements and defining the goals and objectives of a complete data mining project, the requirements and expectations of the stakeholders must be recognized. It's an important fact that the successful implementation of any information project depends on the direct involvement of the staff and stakeholders, the compromises that they develop for the project, and the satisfactions of their own expectations. If

users and stakeholders do not believe in the project's results, it is likely that models, patterns, or relationships will not be applied or implemented.

b. Structure the Work

Formulate Project Goals and Objectives: The goals and objectives of a data mining project must be clear and specific; they must be completely understood by all the participants involved. These goals and objectives are also dynamic because they must correspond to the needs and requirements of the business, factory, or organization. Goals and objectives should be periodically revised, updated, and they must be defined within a timeframe that corresponds to the business's or organization's perspective; otherwise, the data, rules, models and relationships structured and defined by the project will be outdated and useless. The necessity for clear but dynamic goals corresponds to the way that products, business, organizations, and processes evolve. New markets, new customers, new products, and new processes may require different data, different tasks, or different tools. Any data mining project that tries to produce useful results under these conditions must account for change (see Figure 7).

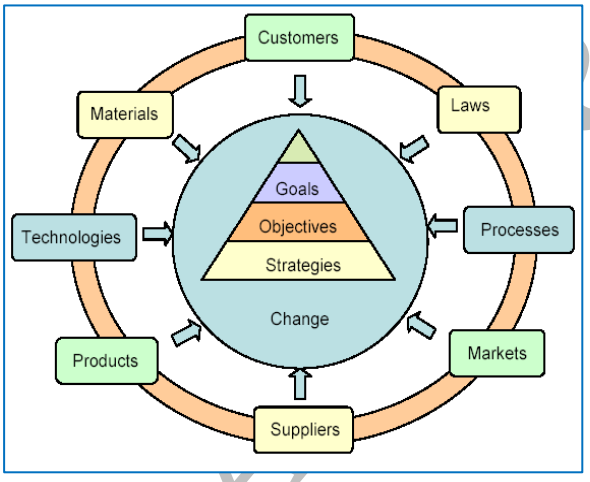

Fig.7. Factors of Change in Data Mining Projects

One way to account for change is to maintain a Gantt chart, which is a valuable tool that can be used to plan and schedule data mining projects. A Gantt chart represents project tasks as horizontal bars, using a calendar time line [20]. They can be easily prepared; they are easy to update, easy to understand, and are very useful in evaluating a project's progress.

Select Task, Techniques and Tools for the Project: The selection of tasks must depend primarily on the goals of the project, rather than solely for techniques and tools available. It is unwise to select the tasks after the selection of the techniques and tools because both tools and techniques may influence or limit the data mining tasks. Data mining techniques must be chosen before tools are chosen, to avoid applying techniques that do not correspond with.

As discussed above, certain guidelines can be used for selecting appropriate techniques in data mining projects. For example, decision trees are useful because of the following characteristics: they are easy to manage and understand, they can work with categorical and numeric data, they are not affected by extremes values, they can work with missing data, and they can process large data sets. However, decision trees also present some disadvantages. For continuous variables or multiple regressions, the use of many rules is usually required, and small changes in the data may generate considerably different tree structures. In order to select the best data mining tool for the given conditions of a specific project or study, important features should be evaluated to determine whether hey correspond with the requirements of the project. Although several authors have suggested a number of features that should be analyzed when selecting a data mining package, they also have concluded, based in their own experience that "there is no one best data mining tool for all purposes " [6]. Some of the most important characteristics to keep in mind when selecting data mining tools are presented in Figure 8. First, it is very important to determine whether the software will be used for a specific type of project or used in a variety of different studies with multiple characteristics and requirements. Purchasing a tool that has unused features will be a waste of resources. Price is another important element, especially with the wide range of options available in the market. One of the issues of data mining software is that it can be very expensive, while the returns on projects may be difficult to quantify or require a considerable amount of time before they can be precisely measured. When determining the cost of a data mining solution, it is essential to include the potential costs of elements such as software for all applications and licenses; all the equipment required; installation; personnel and staff training; maintenance and support; the future cost of upgrading software and hardware; and any other cost that may be incurred during the project; But most of them usually ignored. After estimating all the possible costs, the alternatives, which are in most cases mutually exclusive, evaluated using equivalent worth methods such as present worth (PW), annual worth, (AW) or future worth (FW) [17]. If the benefits are determined to be equal for all the alternatives, only costs should be compared. Performance tries to measure the ability of data mining tool to handle and perform efficiently all of its correspondent tasks. It can be established by the amount of time required for a specific task given a selected set of conditions, but it can also include such elements as robustness, important and can also include such elements as robustness, capacity, formats, software architecture, platforms of operation, and compatibility with other software packages. One way to evaluate and

compare the features corresponding to each of the available data mining tools is a decision matrix [6], in which all the most important and relevant characteristics are selected and then weighed and measured corresponding with the preferences for and requirements of the project, the stakeholders, and the organization. Performance benchmarking analysis among different software solutions can also be included in this approach (see Figure 9). Functionality measures the ability of software to work under different sets of circumstances. It involves elements such as the number of different data mining techniques included, the degree of customization of models and algorithms that the package supports, and the different types of data that can be utilized. Reporting capabilities, sampling techniques, model validation and model exporting are also other features that can be analyzed under the category of functionality [6]. Usability; All data mining tools must be easy to learn, understand, and use, so that they may be applied effectively. Selecting an otherwise excellent package that is very difficult to use or Figure out may risk acceptance of the results, may create more resistance from the point of view of the users, and may cause the project to fail. Finally, remember that a good data mining application without convenient service support may cause considerable amounts of time and resources to be diverted to solve unexpected problems, conflicts or misunderstandings. Support can be measured by several different factors, such as the documentation provided by the vendors, the time available for inquiries and conflict solutions, the vendors' services and resources available for customer support, locations, the training available and offered to the users, and consulting services for future projects or expansions. The importance and weight of each factor should be determined and measured according to the specific conditions of each project; moreover, they must periodically be revised as a response to the dynamic conditions that influence the software market.

Fig. 8. Proposed Factors for Selection of Data Mining Tools

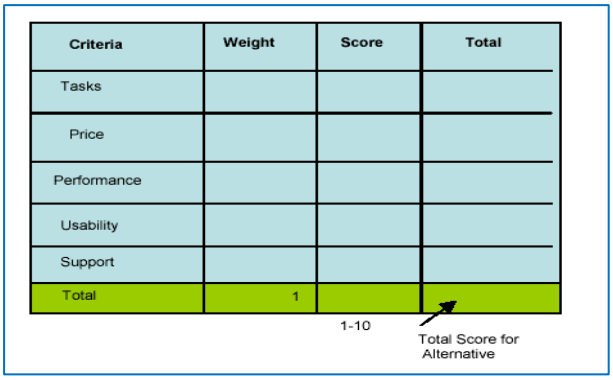

Fig. 9. Decision Matrix for the Selection of Tools

Identify Resources Required: Hardware, Software, Data, and Personnel: Data mining projects may involve many resources, which may be classified into four types: software, hardware, data and personnel. Identifying all these resources is vitally important to determine their accessibility, functions, and involvement in new data mining projects. Unfortunately, although institutions may have already acquired these types of resources, they may be currently assigned to other different projects in execution or they can be unavailable during the implementation of the project. A careful evaluation of resource capacity and availability should be conducted in order to determine possible involvement in the project.

Identify Additional Resources Requirements: Once it has been established what resources are available, an estimation of the remaining resources required for the project should be conducted. This study must include all the software, hardware, personnel and data that are required and are not already available in the organization. We should notice that the quality of model built with a data mining study depends on the quality of the information on which it has been based.

Determine Feasibility of Project: Feasibility analyses can be divided into four major sections: operational, technical, schedule, and economic. The operational feasibility analysis determines whether the project can work, as well as whether it would be accepted in the organization. Technical feasibility, in contrast, concerns the availability of the technology required to implement the project. The schedule feasibility analysis determines whether the project can be successfully completed within a desirable or required timeframe. The schedule feasibility analysis determines whether the project can be successfully completed within a desirable or required timeframe. An economic feasibility study involves determining whether the benefits generated by the project are economically attractive enough to make it worth implementing the project [20]. In order to perform an economic feasibility study, both the benefits and the cost should be estimated as well as possible. If the benefit cost ratio is equal to or greater than 1, the project is economically attractive [17]. It is also important to remember that data mining costs can include several different categories like: Software, Hardware, Installation, Training, Consulting and outsourcing, Maintenance and support costs. Benefits from a data mining project vary according to the goal, the strategies, and the type of study. Some of the benefits that a data mining project may represent are: Increase in productivity, Reduction of cost, Increase in product quality, Increase in personnel safety, Process improvement, Waste reduction, Reduction in production times, Increase in sales, and Improvement in design of new products.

c. Develop Data Model

Data mining models can be automatically produced by data mining tools or programmed using the rules, patterns, or relationships that the tool discovers. Not all data mining projects require the creation of a model; in some cases, the information provided by a data mining tool is good enough to be used alone, to implement changes in a manufacturing process for example, or to select a specific combination of variables and materials. The following section describes the major phases that must be performed for the development of a data mining model.

Data Gathering: Information is a dynamic asset which changes in time. Products, processes, operators, regulations, services, customers, suppliers and materials are dynamic factors that frequently change. And so does the information concerning them; but many data mining studies assume that required information is already available. Another important fact is considering essential aspects of information such as owners, persons responsible, formats available, cost of retrieval, size, security requirements, and privacy [19].

Data Preparation: In many cases, data must be cleaned and integrated in order to correct possible inaccuracies, remove irregularities, eliminate duplicated data, detect and correct missing values, and check for any possible inconsistencies, before the analysis can began. Data mining tools can effectively create valid and insightful models only when the information provided is free of noise factors. In the case of missing values, the approaches that are usually considered are: ignore records, fill in values manually, create a special value or category, use the mean value of the distribution, use the mean value of the same class, and use the most probable value [9].

Model Development: Unfortunately, the rules, patterns or relationships identified by the different algorithms do not always have a significant meaning or use. Human experts are then required to identify, choose, and decide which are the most important rules and

significant models. For this aspect, training plays a very important role. Users must understand not only how to manage the software packages, but also what the data really represents. Additionally, for specific tasks such as process monitoring, quality control or product design, users must also have an authentic understanding about the process, tasks, materials and conditions involved.

Model Validation: The purpose of the model validation phase is to determine whether or not the models created by the data mining tools can correctly predict the behavior of the variables represented by the data. As mentioned above, a validation data set can be used to verify whether the predicted values of the model are close enough to the behavior expressed by the data in the validation data set. In order to perform this task, thresholds can be assigned according to the specific needs and conditions of each project.

d. Implement Model

Once a model is validated, it can be implemented according to the goals and objectives initially established for the project. Implementation is an important phase and also requires analyzing and interpreting the results generated by the models. Not all data mining projects require the implementation of a specific model. However, the information gathered during the process, and the rules, relationships, or patterns discovered, can be used to solve specific problems, give recommendations, make decisions, or identify the necessity of further studies. In this step, evaluation of the project can also be measured using the decision matrix in Figure 10.

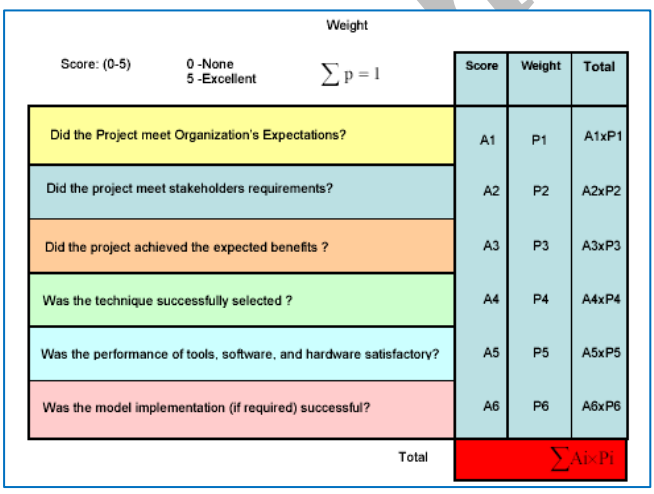

Fig. 10. Decision Matrix for Project Evaluation

e. Establish On-going Support

Finally, in many cases, maintenance operations must be periodically conducted for the equipment; moreover, the data and information residing in data marts, data repositories, and data warehouses must be protected by performing periodic back-ups. Back-ups can be full, differential, or incremental, according to the requirements of any given case. Additionally, new types or sources of information, new versions of software packets, new operational systems, or new equipment may be available. The support phase must ensure that both the model and the corresponded applications are working appropriately and correspond with the specifications of the project.

5. CONCLUSIONS AND RECOMMENDATIONS

By using a system analysis approach, this paper presents a proposed methodology for using data mining in solving problems related to industrial procedures (See Figure 11). The proposed methodology encompasses five major phases: analyze the organization, structure the work; develop data model, implement model, and on-going support. Each of these phases has been described in detail and covers the major steps that any data mining project in industrial procedures must sustain from the origin of the project to its final implementation and support phases. The proposed methodology presents a solid framework capable of enabling industrial specialists to apply data mining in a consistent and repeatable way, which would enable them to evaluate data mining projects, duplicate results, or determine where the errors have occurred in their data mining projects. This research presented a conceptual model to be applied in industrial procedures applications of data mining. This methodology, however, should be applicable to a variety of data mining projects. The next step for this research is to test and improve this conceptual model. Also, another research can conducted on the effectiveness of the other methodologies, rather than SEMMA and Crisp-DM. Data mining is a constantly evolving tool, so this research will endeavor to involve it dynamically in the industrial procedures' toolbox.

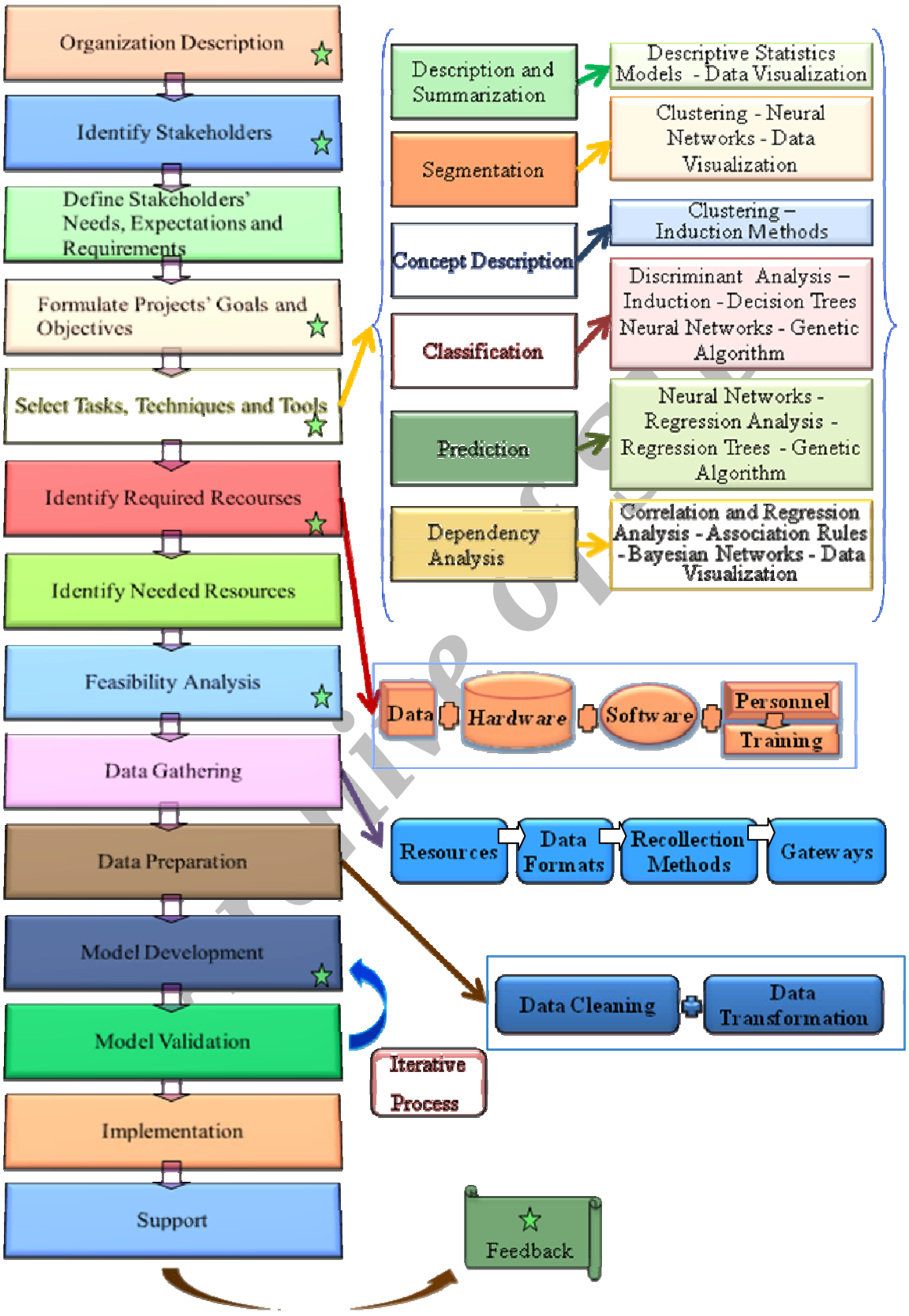

Fig. 11. Schematic Diagram of the Proposed Methodology

References

[1] Anonymous, Mining for a competitive Advantage in your Data warehouse. Techguide.com retrieved from the World Wide Web on January13, 2002. http://techguide.znet.com/html/datamine/.

[2] Anonymous, Uncover gems of information. SAS Institute. Retrieved from the World Wide Web on January 28, 2002. http://www.sas.com/products/miner/index.html

[3] E. Bertino, B. Catania, and E. Caglio, Applying Data mining Techniques to Wafer Manufacturing. 112-115, 2001.

[4] D. Braha, Data Mining for Design and Manufacturing: Methods and Applications. Kluwer Academic Publishers, 53-70, 2001.

[5] Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, And R. Wirth, CRISP-DM 1.0, Step by Step data mining guide. USA, SPSS Inc.,35-45, 2000.

[6] K. Collier, D. Sautter, M. Medidi, et al. Methodology for evaluating and selecting Data mining software. Center for Data Insight (CDI), Northern Arizona University, 83-112, 2002.

[7] W, Christopher. B. Tersa, Data Mining Solutions, Methods for Solving Real-World Problems. John Wiley & Sons, Inc.,30-59, 1998.

[8] R. L.Grossman, C. Kamath, P. Kegelmeyer, V. Kumar, and R. Namburu, Data Mining For Scientific and Engineering Applications. Kluwer Academic Publishers, 45-78, 2001, Netherlands.

[9] J. Han, M. Kamber, Data Mining, Concepts and Techniques, Morgan Kaufmann Publishers, 67-120, 2001.

[10] B. Stacey, D. Michelle, J. Cooch, and R. Paul, Data Mining. University of Iowa, 26-43, 2002.

[11] D.A. Koonce, S. Tsai, and C. H. Fang, A Data Mining Tool For Learning From Manufacturing Systems. Computer and Industrial procedures, V33, 50-57, 1997.

[12] D.A. Koonce, and S. Tsai, Using data mining to find patterns in genetic algorithms solutions to a job shop schedule. Computer and Industrial procedures, 63-68, 2000.

[13] R. B. Landis, Studying Engineering, A Road Map to a Rewarding Career. Discovery Press, Burbank California, USA, 35-50, 1995.

[14] C. J. McDonald, New tools for yield improvement in the integrated circuit manufacturing: can they be applied to reliability? Microelectronics Reliability, 25-29, 1999.

[15] R. Milne, M. Drummond, and P. Renoux, Predicting making defects online using data mining. Knowledge-Base Systems, V11, 33-38, 1998.

[16] S. Y. Sohn, and H. Shin, Pattern recognition for road traffic accident severity in Korea. Ergonomics. V44, Issue 1, January, 15-45-63, 2001.

[17] W. G. Sullivan, J. A. Bontadelli, and E. M. Wicks, Engineering Economy. Prentice- Hall Inc, USA, 45-57, 2000.

[18] N. Subramanian, Data mining approach to improvement and standardization of operations in test facility. Thesis presented for the Master of Science degree, Ohio State University. 53-70, 1999.

[19] Two crows corporation, Introduction to Data Mining and Knowledge Discovery. Third Edition, 112-124, 1999.

[20] J. Whitten, and L. Bentley, Systems Analysis and Design methods. Irwin McGraw-Hill, Forth Edition, 35-57, 1998.

[21] X. Z. Wang, Data Mining and Knowledge Discovery for Process Monitoring and Control. Springer – Verlang London Limited, Great Britain, 32-41, 1999.