Survey Data Mining and its Application in Industrial Engineering

Payam Mohebbi, Pirooz Jafari

MSc in Industrial Engineering, System Management and Productivity, Amirkabir University of Technology, Mahshahr Branch, Khuzestan Province, Iran
alborzmohebbi@yahoo.com

Abstract: Data mining is the computational process of discovering patterns in large data sets and is the process of discovering correlations, patterns, trends or relationships by searching through a large amount of data stored in repositories, corporate databases. Industrial engineering is a broad field and has many tools and techniques in its problem-solving arsenal. Each place of operation may generate large volumes of data. Corporate decision makers require access from all such sources and take strategic decisions. The data warehouse is used in the significant business value by improving the effectiveness of managerial decision-making. The purpose of this study is to improve the effectiveness of industrial engineering solutions through the application of data mining. To achieve this objective, an adaptation of the engineering design process is used to develop a methodology for effective application of data mining to databases and data repositories specifically designed for industrial engineering operations. This paper concludes by describing some of the advantages and disadvantages of the application of data mining techniques and tools to industrial engineering; it mentions some possible problems or issues in its implementation.
[Payam Mohebbi, Pirooz Jafari. **Survey Data Mining and its Application in Industrial Engineering.** *Nat Sci* 2016;14(9):70-75]. ISSN 1545-0740 (print); ISSN 2375-7167 (online). http://www.sciencepub.net/nature. 11. doi:10.7537/marsnsj140916.11.

Keywords: Data Mining, Industrial Management, System Management and Productivity

1. Introduction

Data mining has recently become one of the most progressive and promising fields for the extraction and manipulation of data to produce useful information. 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. For example, data mining has produced information that enables companies to create profiles of current and prospective customers to help in gaining and retaining their customers. Other uses of data mining include development of cross selling and marketing strategies, exposure of possible crimes or frauds, finding patterns in the access of users to their web sites, and process improvement. The power of data mining is yet to be fully exploited by industry. Manufacturing, for example, 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.

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. The kinds of relationships that exist are believed to be sometimes unclear to information analysts because the amounts of information are too large or the kinds of relationships are too difficult to imagine. 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 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.

Additionally, data mining applications continue to be developed. There are, however, few that support decision-making in industrial engineering. Thus, applications of data mining in areas such as quality control, process control, human factors, material handling and maintenance and reliability in production systems should be studied and addressed in more detail.

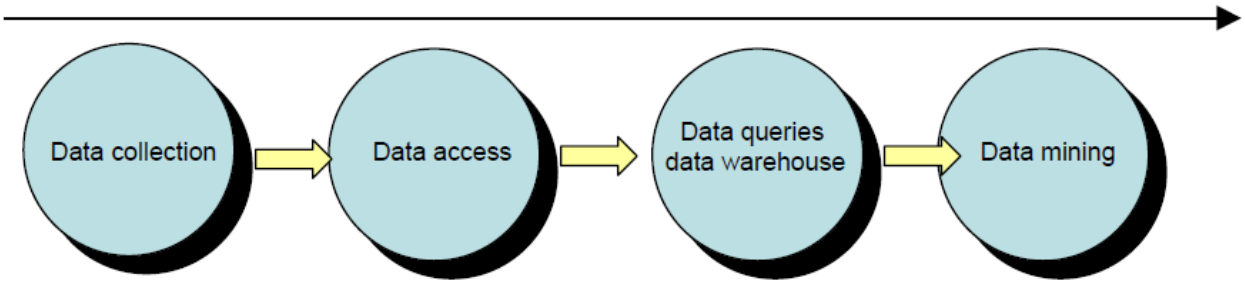

Fig. 1. Data Mining Evolution

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

- Description and summarization
- Concept descriptions
- Segmentation
- Classification and case-based reasoning
- Prediction
- Dependency analysis.

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. 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. More and more data mining applications are being discovered and implemented; they are helping many companies to manage and allocate their resources in a more effective and efficient way, reducing costs, and improving the quality of products and services that they offer. One of the more frequently used data mining applications is cross-selling or expanding the products that are being sold to customers [36]. Thanks to data mining, it is possible to identify groups of customers who have a special set of characteristics or preferences and are therefore more likely to respond to some kind of targeted publicity or marketing strategy. Data mining techniques have been used to predict and detect fraudulent transactions or claims, to combine medical procedures that may produce better treatments, and to implement quality control in the manufacture of a variety of products.

Data mining can also establish what motivations or factors influence customer behavior and which groups are more likely to change from one company to another in a given time. This approach has been widely applied, with very good results, to mailing lists, catalogues, and the distribution arrangement of products in stores and supermarkets. For that reason, data mining techniques are being introduced into applications such as customer relationship management (CRM) solutions, which are decision support systems (DSS) capable of managing and studying customers' data, behaviors, and preferences, in order to increase and maximize the profits in businesses.

Industrial Engineering Decisions

Industrial engineers focus on the design, improvement, and installation of integrated systems of people, processes, materials, and equipment. As a result, there are many possible applications for data mining techniques in industrial engineering. Industrial engineers must decide and select the most effective ways for an organization to apply the basic factors of production for example, machines, materials, people, processes, information and energy to make or generate products and services. Industrial engineers also plan, design, implement, and manage integrated production and service delivery systems, and make decisions that ensure performance, reliability, maintainability, schedule adherence, and cost control. 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. Data mining solutions have been focused thus far on applications such as customer retention, customer profile analysis, fraud detection, cross-selling, marketing expansion, medical treatments, and the creation of user access profiles.

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. These SQC systems are usually employed to analyze data collected by Statistical Process Control (SPC) systems, which monitor production processes in real time through the use of online sensors. SQC is usually applied using statistical techniques also included in data mining, but these techniques are also capable of analyzing parameters, with the same understandable effect on the process of the one given by SPC systems.

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. Historical process logs can be analyzed to monitor the process at different stages. When monitoring processes, the models created with data mining tools can determine whether or not the current process state can generate satisfactory outcomes; if corrections are needed, then programs can also notify operators or recommend changes.

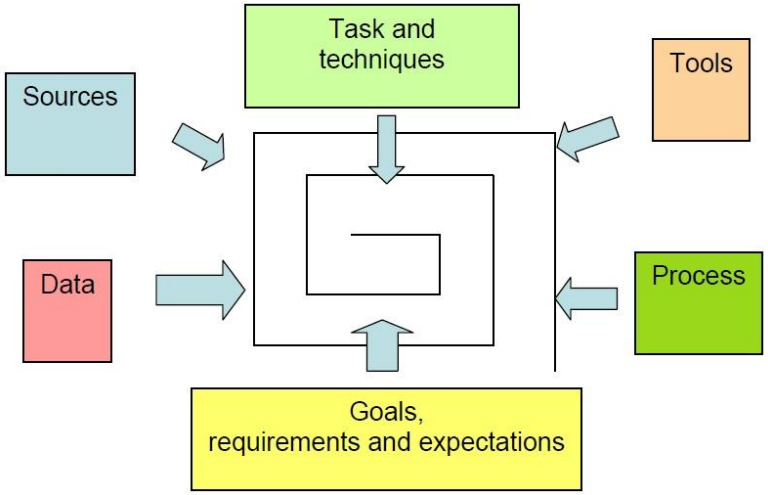

Fig. 2. Data Mining of Knowledge

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. They can also be used to create rules and models that identify the source of problems or to identify additional patterns in parts or equipment failures. 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.

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. There are many options, software vendors, and techniques; and is difficult to decide how to choose a data mining tool.

Methodology:

The engineering design process is based on the scientific approach to problem solving. The distinguishing characteristic of engineering, however, is that it uses a systems perspective; that is, it studies a problem environment in order to implement corrective solutions that take the form of new or improved systems. The engineering design process, as described by Landis, was used in the execution of this study. This engineering design process is depicted in Figure 3 and its six steps are detailed below.

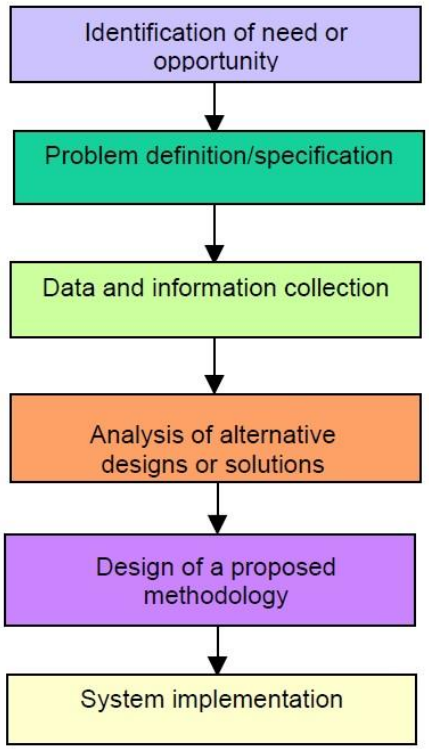

Fig. 3. Design Process

In order to accomplish this study, surveys, analysis, reviews, and comparisons of data mining applications were conducted. These were based on vendors' information and case studies available in literature and research publications. The 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 engineering applications. It also asked whether they had applied or sold their data mining products for the implementation of projects related to industrial engineering areas such as quality control, scheduling, manufacturing, safety, or ergonomics. Other questions were related to hardware requirements and prices.

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. These have some drawbacks. 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". However, this statement is only partially true. While individual preferences and intuition may contribute to ingenious data mining methodologies, at the same time, there are essential steps that data mining methodologies must include and elegant, efficient ways to combine methodological elements to obtain superior results.

While this approach has broader applications than SEMMA, one of its major drawbacks is that it combines tools (software packets) and techniques in the same category. If tools and techniques are combined and selected simultaneously, techniques may be chosen because they are supported by specific data mining tools, and not because they are the most relevant to the purpose of the study, or because they are needed. This also may cause the organization's goals and requirements to be under-analyzed and biases the study. In data mining projects, it is important to analyze the organization's needs, requirements, goals, and strategies. Organizations, for example, may need to extract information from their data warehouses either on a one-time basis or on a recurrent one.

Other essential elements that are considered are the collection of data during the data understanding phase, but only for analyzing available data (not necessarily the data required). Data is also analyzed and verified in order to ensure that the quality of data will allow the intended results of the modules. Techniques should be selected according to an organization's goals and requirements and should not depend only on the data available. If the data available is not enough for the organization to perform a data mining project, new data and information should be collected; otherwise, the selection of the technique required may be influenced by only the data in existence, so the resulting models may be biased and will not correspond with the organization's actual intentions. A related problem is that although assumptions are clearly declared, they may not be sufficiently revised. Changes in data from the past to the future may cause assumptions about data that were once valid to be incorrect.

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.

To successfully implement a data mining project, it is very important to identify all the key elements involved in it, and stakeholders are an essential element in any information system analysis and design task. Stakeholders include all the major owners, users, analysts, designers, and developers of an information system, as well as the essential personnel on which the successful implementation of the project will depend. Identifying the stakeholders and their requirements will allow analysts to completely recognize the critical elements of the project, together with its true intentions and expected results.

Develop Data Model

Data mining techniques, moreover, must be chosen before tools are chosen, to avoid applying techniques that do not correspond with the real goals of the study. The traditional data mining approach has been to consider several techniques that are usually applied in order to find the one that fits the best. 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, they are able to reveal complex interactions and a lack of linear relationships, they are good at handling noise in data, and they can processing 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.

Neural networks also have good advantages; their models, for example, are usually considered easy to use, and because they are universal approximations, it is possible to apply them to model a wide range of relationships or patterns. Nevertheless, because of the complexity of the neural network patterns they create, their models frequently are difficult to understand, they may require a lot of time to process large data sets, and they cannot be implemented in different software packages without difficulty.

Some of the most important characteristics to keep in mind when selecting data mining tools are presented in Figure 8. Considering the task involved in the project is an essential factor in selecting the best tool. 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. If a data mining tool will be applied to a specific set of conditions, the evaluation of the tool should concern those conditions; additional features may be desirable but not required. Purchasing a tool that has unused features will be a waste of resources.

Unfortunately, because the useful life of a given software package is difficult to estimate, care must be taken when selecting study periods. Moreover, in data mining projects, the repeatability assumption may not be an adequate method. Because software life cycles are becoming shorter [33], its value can be affected by new version releases, and most of the software assets do not have market value at the end of the useful life. Thus, the contaminated assumption can be used in which, for all the investment alternatives whose useful lives are less than the study period, all the cash flows are reinvested at the MARR until the end of the study period.

Additionally, sensitivity analysis is another suitable method that can be employed to select the best alternative when considering different useful lives for a project. Sensitivity analysis is a good method to apply when considering risk and uncertainty in decision-making activities for projects. Both risk and uncertainty are caused by the lack of precise knowledge about the future, and the main idea behind sensitivity analysis is to determine the degree to which changes in a given factor or estimate would affect capital investment decision.

Techniques and Tools for the Project

The usability of the software package is another important element that must be analyzed when selecting a data mining tool. 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. The users should be confident and understand what they are doing; otherwise the probability of errors dramatically increases, and nobody in the organization will believe in the results. Usability depends on several factors such as the graphic interfaces, access and navigation features, learning curves, experience required, help tools, reporting and visualization features, and predefined functions and models. All of these fundamentals must to be consistent with the main purpose of the project and should effectively contribute solving any difficulties that arise during its execution and implementation phases.

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. Although these elements are among the most important to consider when selecting data mining tools and applications, these are not the only ones, and in many cases not all of them are required.

Many of the companies and institutions willing to implement a data mining project may already have an existing infrastructure of resources available. Organizations may already own, for example, networks, servers, database management systems, data repositories, or data warehouses that can be employed in the project. Data analysts, server administrators, and technicians working in the company can also be very useful, even if they are not directly considered stakeholders themselves; they can provide an invaluable source of information thanks to their knowledge and experience with current systems.

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.

The quality of model built with a data mining study depends on the quality of the information on which it has been based. If this information contains large numbers of errors and inconsistencies, the inconsistencies and the errors may be also be shown in the results predicted by the model. Information also may become outdated because of changes in processes, workstations, operations and products, so

models may produce predictions that are no longer valid.

For that reason, although data may be already available for the project, this data may not be consistent enough to create adequate rules, patterns and relationships and new data should be collected. In those cases a data recollection process should be conducted before data preparation can be initiated.

The schedule feasibility analysis determines whether the project can be successfully completed within a desirable or required timeframe. For certain projects, for example, the amount of data available would not be sufficient, and more time would be required in order to collect more information before the models can be successfully developed.

In other cases, by the time the project has being successfully implemented; changes in processes, products, materials or workstations can cause the model to be no longer necessary or valid. For that reason, if unexpected changes occur during the development phases of the project, schedule feasibility should be updated in order to guarantee that the study will generate the expected benefits.

The creation and development of a data mining model is another important step in a data mining project. 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; the physical development of the model in many cases is optional and depends on the requirements of the organization.

Conclusion:

This research presented a conceptual model to be applied in industrial engineering 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. Data mining is a constantly evolving tool, so this research will endeavor to involve it dynamically in the industrial engineering toolbox. Although, many of these databases are not designed for data mining analysis and much of the information contained in them is in text format, they still can be used as a source of data for further analysis. Text data mining, which is presently under development by many different software companies, would be a suitable application for these cases. Text data mining is currently being used for email routing, document indexing, and document filtering; but in the future, it will be able to extract more detailed and comprehensive information from a wide variety of sources.

References:

1. Armstrong, Rob. Coffin, Tom. And Rolf Hanusa. “Secrets of the Best Data Warehouses in the world”. Coffin Data Warehousing .2015. USA.
2. Bertino, Elisa, Catania, Barbara and Caglio, Eleonora. “Applying Data mining Techniques to Wafer Manufacturing” ”, retrieved from the World Wide, 2015.
3. Braha, Dan. “Data Mining for Design and Manufacturing: Methods and Applications”. Kluwer Academic Publishers. 2015.
4. Chapman, Pete. Clinton, Julian. Kerber, Randy. Khabaza, Thomas. Reinartz, Thomas. Shearer, Colin. And Wirth, Rüdiger. “CRISP-DM 1.0, Step by Step data mining guide” USA .SPSS Inc. 2000, retrieved from the World Wide Web, 2013.
5. Chen, Nianyi., Zhu, Dongping Daniel, and Wang, Wenhua, “Intelligent material processing by hyperspace data mining”. Engineering Applications of Artificial Intelligence, V13. 2014.
6. Famili,A., Shen, Wei-Min., Weber.,and Richard., Simoudis, Evangelos. “Data Preprocessing and Intelligent Data Analysis”. Intelligent Data Analysis. V1. 2012.
7. Grossman, Robert L., Kamath, Chandrika., Kegelmeyer, Philip., Kumar, Vipin., and Namburu, Raju R. “Data Mining For Scientific and Engineering Applications”. Kluwer Academic Publishers. 2013.
8. Inmon, H. Zachman, John, and Geiger, Jonathan. “Data Stores Data Warehousing and the Zachman Framework, Managing Enterprise Knowledge”. McGraw-Hill.2012.USA.
9. Koonce, D.A., and Tsai , S.-C. “Using data mining to find patterns in genetic algorithms solutions to a job shop schedule”. Computer and Industrial Engineering .V38. 2012.
10. McDonald, Chris J. “New tools for yield improvement in the integrated circuit manufacturing: can they be applied to reliability?”. Microelectronics Reliability. V39. 2011.
11. Parsaye K., “Data Mines for Data Warehouses”, Information Discovery, retrieved from the World Wide Web, 2012
12. Shelly, Gary B. Cashman, Thomas J. Vermaat, Mysty E. and Walker, Tim J. “Discovering Computers 2001”. Shelly Cashman Series. 2012. USA.
13. Sohn, S Y., and Shin, H. “Pattern recognition for road traffic accident severity in Korea”. Ergonomics. V44. Issue 1. 2011.
14. Sullivan, William G., Bontadelli, James A., and Wicks Elin M. “Engineering Economy” Prentice- Hall Inc. 2011. USA.
15. Witten, IAN, and Frank, Eibe. “Data Mining, Practical Machine Learning Tools and Techniques with Java Implementations”. Morgan Kaufmann Publishers.2012. USA.