Data mining methodology and its application

Tripti Upadhyay
akshara820@gmail.com
Suyash Institute of Information Technology, Gorakhpur, Uttar Pradesh

Sumit Verma
aadi.007.we@gmail.com
Suyash Institute of Information Technology, Gorakhpur, Uttar Pradesh

ABSTRACT

Data mining is the process of discovering correlations, patterns, trends or relationships by probing through a substantial amount of data stored in repositories, corporate databases, and data warehouses. Industrial engineering is a broad field and has many implements and techniques in its quandary-solving arsenal. The purport of this study is to ameliorate the efficacy of industrial engineering solutions through the application of data mining. To achieve this objective, an adaptation of the engineering design process is utilized to develop a methodology for efficacious application of data mining to databases and data repositories concretely designed for industrial engineering operations. This paper concludes by describing some of the advantages and disadvantages of the application of data mining techniques and implements to industrial engineering; it mentions some possible quandaries or issues in its implementation; and conclusively, it provides recommendations for future research in the application of data mining to facilitate decisions pertinent to industrial engineering.

Keywords — Data preprocessing, Pattern analysis, Pattern discovery, Web usage mining

1. INTRODUCTION

Data mining has recently become one of the most progressive and promising fields for the extraction and manipulation of data to engender utilizable information. Thousands of businesses are utilizing data mining applications every day in order to manipulate, identify, and extract subsidiary information from the records stored in their databases, data repositories, and data warehouses. With this kind of information, companies have been able to ameliorate their businesses by applying the patterns, relationships, and trends that have lain obnubilated or undiscovered within colossal amounts of data. For example, data mining has engendered information that enables companies to engender profiles of current and prospective customers to avail in gaining and retaining their customers. Other utilizations of data mining include the development of cross-selling and marketing strategies, exposure of possible malefactions or frauds, finding patterns in the access of users to their web sites, and process amelioration.

The puissance of data mining is yet to be planarity exploited by industry. Manufacturing, for example, is one of the incipient fields in which data mining implements and technique are commencing to be used prosperously. Process optimization, job shop scheduling, quality control, and human.

Factors are some of the areas in which data mining implements such as neural networks, genetic algorithms, decision trees, and data visualization can be implemented with great results.

Data mining utilizes a series of pattern apperception technologies and statistical and mathematical techniques to discover the possible rules or relationships that govern the data in the databases. Data mining must additionally be considered as an iterative process that requires goals and objectives to be designated [1]. Once the intended goals are thoroughly defined, it is obligatory to determine what data is available or can be amassed Sometimes the data.

Data mining additionally involves a methodology for implementation. The methodology, or structured approach, customarily 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. Each methodology strives to avail users to obtain the best data to provide the most responsive information to address their desiderata.

The apperception that efficacious decisions are predicated on felicitous information from precise and current data is not incipient. The evolution and development of finding the right data for decision-making commenced 30 years ago, and it has progressed through several stages of development [7]. These are shown in Figure 1, and their descriptions follow.

© 2019, www.IJARIIT.com All Rights Reserved
1.1 Data collection
During the late '60s, simple reports of pre-formatted information were created from data stored in databases. These databases stored the data, while applications retrieved and manipulated it to produce structured reports containing information to meet specific decision-making needs.

1.2 Data access
Factors are some of the areas in which data mining implements such as neural networks, genetic algorithms, decision trees, and data visualization can be implemented with great results.

Data mining utilizes a series of pattern apperception technologies and statistical and mathematical techniques to discover the possible rules or relationships that govern the data in the databases. Data mining must supplementally be considered as an iterative process that requires goals and objectives to be designated [1]. Once the intended goals are exhaustively defined, it is obligatory to determine what data is available or can be amassed. Sometime the data.

Data mining supplementally involves a methodology for implementation. The methodology, or structured approach, customarily 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. Each methodology strives to avail users to obtain the best data to provide the most responsive information to address their desiderata.

The apperception that efficacious decisions are predicated on felicitous information from precise and current data is not incipient. The evolution and development of finding the right data for decision-making commenced 30 years ago, and it has progressed through several stages of development [7]. These are shown in Figure 1, and their descriptions follow.

1.3 Data queries
Later, in the 1990s, users required immediate access to more detailed information that responded to “on the fly” questions. They wanted the information to be “just-in-time” to correlate with their engenderment and decision-making processes. That betokened that not all of the users’ informational needs could be preprogrammed into the system. At this stage, users commenced to indite their own queries to extract the information that they needed from the database.

1.4 Data mining
In the last few years, users commenced realizing the desideratum for more implements and techniques in order to identify and find relationships in data so that the information obtained was more paramount for their applications. Supplementally, companies apperceived that they had accumulated volumes of data; and, as a result, they needed incipient implements to sort through it all and meet their informational needs. Such implements enabled the system to probe for possible obnubilated relationships in the data, without the direct intervention of the cessation users. Data mining implements were first developed to avail scientists find paramount relationships or patterns from sizably voluminous amounts of data that, if done in a traditional way, would require much time and many resources to find. The next step is to exploit these implements for paramount applications.

2. LITERATURE REVIEW
The application of data mining techniques to industrial engineering is an area that holds promise, but that is currently underdeveloped. Data mining, can, however, be strategically applied to industrial engineering processes such as scheduling, quality control, cost reduction, safety, and others. This chapter outlines some of the data mining techniques and applications that can be utilized by industrial engineers, as well as some of the subsisting ways that industrial engineers employ data mining.

2.1 Data mining techniques
There are a number of techniques utilized in data mining, but not all of them can be applied to all types of data. Neural network algorithms, for example, can be habituated to quantify data (numerical data), but they cannot qualify data precisely (categorical data); consequently, categorical data is conventionally broken up into multiple dichotomous variables, each of them with values of 1 ("yes") or 0 ("no") [34]. For that reason, one single technique cannot be acclimated to perform a consummate data mining study and each technique has its own scope of applications. Some of the techniques applied in data mining include traditional statistics, induction, neural networks, and data visualization. These are described in the following sections.

2.1.1 Traditional Statistics
Some of the traditional statistical methods that can be used for data mining are the following [18]:
Cluster analysis (or segmentation) is one of the most frequently used data mining techniques; it involves disuniting sets of data into groups that include a series of consistent patterns. After the data reveals a consistent pattern, it is then sorted into subsets that are more facile to analyze. This information is withal used to identify subgroups of a population for supplementary studies, as well as to engender profiles for target marketing. Kohonen feature maps and K-designates are some of the most paramount algorithms applied for cluster analysis [34].

Discriminant analysis is one of the oldest relegation techniques. It 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. Discriminant analysis has constraints, however. It postulates that all soothsayer variables are mundanely distributed—but this is not always true. Moreover, unordered categorical values cannot be relegated, and boundaries are restricted to linear forms. Revisions of discriminant analysis are being developed to handle these inhibitions by utilizing quadratic boundaries, estimates of authentic distributions, and bins defined by the categorical variables [34].

Logistic regression is a generalization of linear regression. It is primarily utilized for presaging binary variables and, less frequently, multi-class variables. Models of logistic regression soothsay the logarithm of the odds of the occurrences of discrete variables. The main posit of the logistic regression model is that the logarithm of the odds is linear in the coefficients of the presager variables [34]. Analysts utilizing this technique require experience and adeptness in order to cull the right variables, optate the functional relationship with the replication variable, and account for possible interactions.

Conclusively, time series forecasting soothsays “unknown future values, predicated on time varying series of predictors” [34]. Time series databases contain a series of sequences of values and events that change over time. The trends of those values can be habituated to construct functions of the form Y=f(t), so attributes can be presaged in time or predicated on other process values. For example, downtimes can be soothsaid as a function of setups. With this information, preventive maintenance programs can then be implemented, scheduled, and adjusted in authentic time. Yet with this technique, consequential factors such as the hierarchy of periods, seasonality, calendar effects, and date arithmetic may influence the results. Thus, these factors should be accounted for when time series forecasting is utilized.

2.1.2 Induction and Decision Trees: Induction 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. The specific occurrences of the events in the data are then applied to establish a “confidence factor of the rule” [1]. 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.

Using decision trees, unknown events such as types of defects can be classified, testing the values of important attributes against the values of each node. By following this process, a path can be traced from the root node to the leaf node that identifies the class prediction for that event. Rules can be constructed very easily using decision trees, and they usually follow a form such as “If x =’y’ and z=’d’ and p =’0,5’, then defect = ‘yes’.”

2.1.3 Neural Networks: The neural networks approach includes a series of mathematical models that have the ability to “learn” and thus adapt their actions according to results that have been previously obtained. This technique is based on research in neurophysiology, which studies how the human brain works, replicating it with computers. Neural networks can analyze imprecise, incomplete and complex information and deduct or find important relationships or patterns from this information. Usually, the patterns involved in this kind of analysis are so intricate that they are not easily detected by humans or by other types of computer-based analysis.

2.2 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 [23].

2.3 Data mining applications in industrial engineering
Because data mining techniques search through a 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 over the internet. Yet these are not the only fields in which data mining can be applied. Industrial engineers can indeed use data mining to understand complex systems. While the use of data mining in industrial engineering is not widespread, several successful applications of data mining in fields related to industrial engineering have been reported. Examples of data mining applications in manufacturing processes are the computation of job shop schedules [31], process improvement in circuit and semiconductor manufacturing [24], and the design and selection of new materials [10]. These and other examples are explained in more detail in the following section.
3. RESEARCH METHODOLOGY

3.1 Introduction
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 [23], was used in the execution of this study. This engineering design process is depicted in Figure 3 and its six steps are detailed below.

3.2 Identification of a need or opportunity
The first step in problem-solving is the identification of a need or opportunity. For industrial engineering, these needs and opportunities are extensive and varied. Industrial engineering is a broad area of specialty among the engineering disciplines. Those sitting for its exam for the Fundamentals of Engineering (FE) Exam need mastery in twenty topical areas. This is at least twice as many as other disciplines. Thus, industrial engineers have a larger-than-average demand for information in doing their jobs. Much of this information is available; the challenge is getting to it. The ideal tool to assist in that effort is data mining.

3.3 Problem definition
There are so many options, tasks, techniques, tools, formats, and approaches to data mining that industrial engineers 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 engineering applications. Thus, a data mining methodology to meet the specific requirements of industrial engineering is needed. Such a methodology should assist industrial engineers in selecting appropriate data mining tools and implementing data mining projects from a systems perspective.

3.4 Data and information collection
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 (see Appendix). 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. The most relevant results of this survey are shown in figure 2.

Fig. 2: Application of data mining software to IE areas

Fig. 3: Data mining software price distribution, the year 2002
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 engineering. This structured approach should lead analysts through the steps required in obtaining the data needed to provide the information required for problem-solving. This approach has a number of steps.

- Analyze the organization.
- Structure the work.
- Develop the data model.
- Implement the model.
- Establish on-going support.

The present [33] worth for the costs can be calculated by the following equation:

\[
PW(Cost) = \sum_{k=0}^{n} Cost_k (1 + i)^{-k}
\]

where \(i\) = effective interest rate or MARR, per compounding period;

\(k\) = index for each compounding period;
\(Cost_k\) = amount estimated expenses at the end of period K;
\(n\) = number of compounding periods in the planning horizon.

The best alternative, in this case, will be the one with a lower present worth value—in other words, the lower cost.

If benefits are not equal for all the alternatives, present worth values should be calculated for the difference between the benefits and costs of each alternative. If the benefits are greater than the cost, the alternative may be viable and the best alternative will be the one with a greater net value. If costs are greater than the benefits, a detailed economic feasibility analysis should be conducted in order to determine if the project’s value is enough to justify its completion.

\[
PW(Net) = \sum_{k=0}^{n} (Benefits_k - Cost_k) (1 + i)^{-k}
\]

Where \(i\) = effective interest rate or MARR, per compounding period;

\(k\) = index for each compounding period;
\(Benefits_k\) = amount estimated of benefits or incomes at the end of period K;
\(Cost_k\) = amount estimated of expenses at the end of period K; \(n\) = number of compounding periods in the planning horizon.

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 a market value at the end of the useful life. Thus, the co-terminated 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 [33].

The future worth method is then calculated using the equation [33]:

\[
PW(Net) = \sum_{k=0}^{d} (Benefits_k - Cost_k) d^{-k}(1 + i)(1 + i)^{n-d}
\]

Where \(i\) = effective interest rate or MARR, per compounding period;

\(k\) = index for each compounding period;
\(Benefits_k\) = amount estimated of benefits or incomes at the end of period K.
\(Cost_k\) = amount estimated expenses at the end of period K.
\(d\) = useful life of the alternative.
\(n\) = number of compounding periods in the study period.

Thus, the alternative with the greater positive net future value will be the best from an economic point of view. However, sometimes a useful life is not easy to determine for all data mining projects. In those cases, a probabilistic approach can be used. A discrete probability distribution can be estimated for a useful life on a given alternative. Probabilities of various useful life values can be projected; keeping in mind that the summation of the probability values for all the possible useful lives must be equal to 1. With these probabilities, expected values, variance, and standard deviation of the present worth values can be found by using the following equations [33]:
Expected (PW)

\[ E(PW) = \sum_{j=0}^{l} \left[ \left( \sum_{k=0}^{n} (Benefits_k - Cost_k)(1+i)^{-k} \right) \times p_j \right] \text{ where } \sum_{j=0}^{l} p_j = 1; \]

i = effective interest rate or MARR, per compounding period;

l = number of the possible useful lives for a given alternative;

j = index for each useful life;

p_j = probability associate with a useful life value; k = index for each compounding period;

Benefits_k = amount estimated of benefits or incomes at the end of period K;

Cost_k = amount estimated expenses at the end of period K; n = number of compounding periods in the study period.

Variance (PW)

\[ V(PW) = E[(PW)^2] - [E(PW)]^2 \]

\[ = \sum_{j=0}^{l} \left( \left[ \sum_{k=0}^{n} \left( Benefits_k - Cost_k \right)(1+i)^{-k} \right] \times p_j \right) - \left[ \sum_{j=0}^{l} \left( \sum_{k=0}^{n} \left( Benefits_k - Cost_k \right)(1+i)^{-k} \right) \times p_j \right]^2 \]

And the standard deviation:

\[ SD(PW) = [V(PW)]^{1/2} \]

In order to select the best alternative, in this case, the expected values, variance, and standard deviation of all the alternatives should be compared, the best candidate is the one with the greatest positive expected value and very low variance and standard deviation values.

A major difficulty in these methods is that for several alternatives, several scenarios, and several factors, the computational analysis required may comprehend a considerable amount of effort. Consequently, if useful life is considered as the main factor for data mining projects, a sensitivity analysis can be obtained by comparing the FW values of the alternatives. Thus, it is possible to compare the values obtained using different estimates of d, with the same value of n such as:

\[ FW_{p, \text{(Net)}} = \left( \sum_{k=0}^{d_p} \left( Benefits_k - Cost_k \right)(1+i)^{-k} \right)(1+i)^{n-d_p} \]

Where \( d_p \leq n \), \( d_p \) is the values estimated of useful lives for all the

Alternatives n is the number of compounding periods, and there are p different values of useful lives in the study. FW values then can be compared within each alternative, and the best alternative with the greater FW values should be selected.

Finally, but not least, support is an essential issue of data mining applications, one that must be studied in detail. 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. This condition may influence not only the execution and completion of the project’s schedule but also may involve a far greater additional investment of assets.

5. CONCLUSIONS

Using the relationships, patterns, and rules found by data mining tools, industrial engineers may discover unexpected and useful information that can lead to a better understanding of systems and processes. This information can then be used to design new processes and new products or to create modules and expert systems capable of controlling and optimizing systems. Industrial engineers can also use these modules to obtain better performance and resource utilization. The methodology developed in this research can help in these efforts. There are many possible areas for the application of data mining in industrial engineering. Although many of them have been listed in this research many others can also be found. The methodology proposed in this research provides a structural basis from which additional studies can be developed.

6. REFERENCES


