Scientia Iranica, Vol. 15, No. 4, pp 469–479

SCIENTIA
IRANICA

Essential Components of an Integrated Data Mining Tool for the Oil and Gas Industry with an Example Application in the DJ Basin

Sh. D. Mohaghegh[1]

Data mining seems to be the new buzz word. During the past several years many industries other than the oil and gas industry have realized the potential benefits of da5ta mining and have established sophisticated operations in order to implement this exciting technology in their respective organizations. Data mining is not new. It has been around for many years. What is new about its current implementation is the incorporation of machine learning techniques. The oil and gas industry has become familiar with machine learning techniques since the early 1990s. Neural networks, genetic optimization and fuzzy logic have been used in numerous applications, from well log interpretations to hydraulic fracturing optimization. Therefore, the new interest in data mining in this industry is not surprising. The industry is at its peak state for benefiting from what data mining has to offer, thanks to an abundance of digital data. A word of caution is in order, which is the main motivation behind writing this paper. As with many other new tools and technologies, the term "Data Mining" can be, and is currently being, misused on several occasions. In this paper, an attempt has been made to answer questions such as; what is Data Mining? How can it be accomplished? What are the essential components of an integrated data mining process and what would be the benefits of such a process? In addition to answering questions such as those mentioned above, this paper will provide a road map (a set of guidelines) for a successful data mining project. Finally, the paper concludes by applying the presented guidelines to a hydraulic fracturing data set in the DJ basin of the United States Rockies for a data mining study.

INTRODUCTION

In the past two decades, oil and gas companies have spent millions of dollars to collect digital data or to convert the existing data into digital form. This is due to the fact that they have realized the value of data and the potential it possesses in enhancing their operations. IT departments in larger oil and gas companies, major service companies and other vendors have developed sophisticated software tools that allow operators to organize their data, currently existing in different databases, into a cohesive data warehouse and make it available to engineers. Furthermore, more software applications have been developed to put all that information at the fingertips of engineers and geologists, so that they can look at all sorts of data pertaining to a reservoir, a field or a well. Although these are absolutely essential for the successful operation of a large company, it has created a new monster. There are far more data that can be effectively analyzed. The human brain, although being the most remarkable information processing entity, can only work simultaneously in so many dimensions and is incapable of processing very large volumes of data. Data mining and knowledge discovery, as an integrated process that is shown in Figure 1, can come to rescue in such occasions.

The data mining market size started at about \$540 million in 2002 and grew to about \$1.5 billion in 2005 [1]. Many industries have realized data mining's value and are jumping on the bandwagon of implementation and integration into their operations. Today's data mining uses machine learning, statistical

1. *Department of Petroleum and Natural Gas Engineering, 347 MR Building, West Virginia University, P.O. Box: 6070, Morgantown WV 26506. E-Mail: shahab.Mohaghegh@mail.wvu.edu*

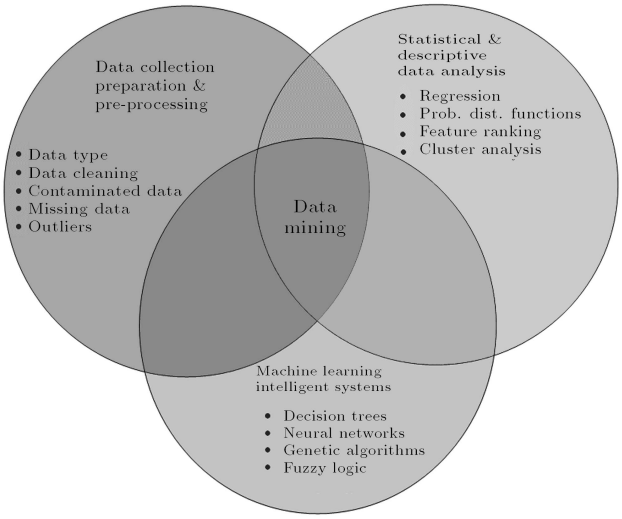

Figure 1. Data mining as an integrated analytical process.

and visualization techniques to discover and present knowledge in a form which is easily comprehensible to humans [2].

By sifting through large volumes of data already available in corporate data warehouses and extracting patterns, information and knowledge from these databases, data mining allows managers to be proactive. It helps them become prospective in company operations, rather than retrospective [3]. It must be noted that company managers are not the only beneficiaries of data mining. In the oil and gas industry there are many field related operations that can benefit from the tools and capabilities that data mining has to offer. One of the natural applications of data mining and knowledge discovery processes in drilling, reservoir and production operations would be identification of the best practices [4]. Data mining processes can have as many applications in this industry as engineers can dream of. It can be applied to the identification of new infill drilling locations, the optimization of hydraulic fracturing results, candidate selection for stimulations using hydraulic fracturing versus chemical treatments in primary or enhanced production operations and/or storage fields, the anticipation of well operation anomalies from real-time down-hole data, formation evaluation by integrating logs and seismic data, to name but a few.

There can be different approaches in using data mining processes. Many times, the operation is exploratory in nature. One may just want to explore the potential of finding valuable information from an existing database. Other times, data mining is after some particular objective. This is when the process is more guided and has the potential to result in immediate benefits for the company. It is important that the goals and objectives of the project be identified in advance and some metrics for measuring its success are determined.

DATA MINING CLASSIFICATIONS

Data mining has recently been enjoying a renewed interest from many different industries. The new interest in data mining has its roots in the large amount of digital data that is being collected and stored in databases and data warehouses. The data owners have allocated an enormous amount of resources in the collection and preservation of the data and would like to utilize this asset by turning data into information and, ultimately, into knowledge. Data mining has been defined as the major tool for knowledge discovery from data. A more formal definition of data mining would be: "The nontrivial extraction of implicit, previously unknown, and potentially useful information from data" [5].

The new interest in data mining may be attributed to the fact that the new set of processes called data mining are a superset of the processes that were previously known by the same name. The original data mining processes were summarized as a collection of statistical analyses. The new data mining processes include several machine learning techniques, as well as statistical analysis. The addition of the recently popularized machine learning and intelligent processes, such as artificial neural networks, genetic algorithms, fuzzy logic and modified cluster analyses, have considerably increased the capabilities and utilities offered by data mining.

Many authors have offered different classifications of the processes that are collectively known as data mining. The most appropriate of these definitions (one that suits the gas and oil industry most) seems to be the one that identifies two classes of data mining processes. These are descriptive and predictive data mining. It is the belief of the author that descriptive data mining is, essentially, a subset of predictive data mining. In other words, in order to perform predictive data mining successfully, one will have to first perform a descriptive data mining and then use the information and the results of this process to complete the predictive data mining. Descriptive and predictive data mining share several common processes as shown in Figure 2.

Descriptive Data Mining

Descriptive data mining is very useful for getting an initial understanding of the presented data. It is an exploratory process that attempts to discover patterns and relationships between different features present in the database. During the descriptive data mining process, the data miner must keep in mind that relevance is an important issue. In other words, the

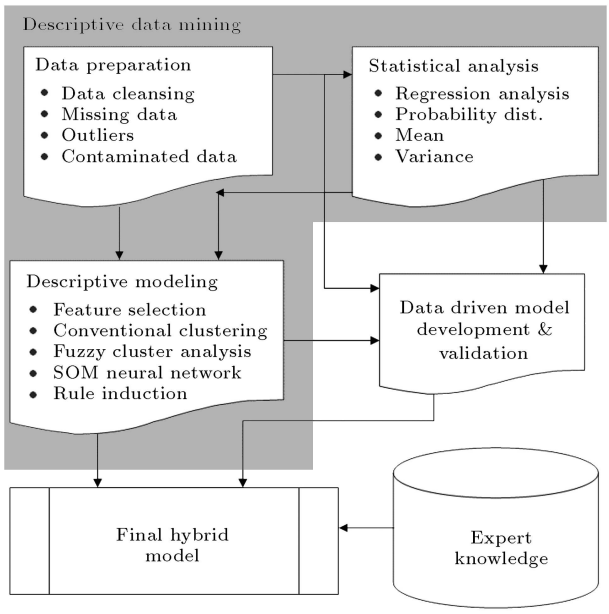

Figure 2. Two classes of data mining with details.

relationships discovered by the miner must be those that are important to users. During this process, many non-obvious patterns may pop out of interest to the data owners. Examples for the gas and oil industry will be presented in future sections of this article.

Tools used during the descriptive data mining process usually consist of an innovative use of the fuzzy set theory [6] and different types of cluster analysis, such as hierarchical clustering, k-mean clustering and fuzzy c-mean clustering. Other popular descriptive data mining tools are rule induction techniques and self-organizing maps. Self-organizing maps utilize unsupervised neural networks, such as Kohonene networks.

Predictive Data Mining

As it was mentioned previously, predictive data mining is a superset that should include descriptive data mining as part of its processes, or, at least, that is how the author wishes to define it, based on past experience. During predictive data mining, the descriptive data mining processes are used as a prelude to the development of a predictive model. The predictive model can then be used in order to answer questions and assist the data miner in identifying trends in the data. What is most interesting about predictive data mining that distinguishes it from descriptive data mining is that it can identify the type of patterns that might not yet exist in the dataset but which has the potential of developing.

Unlike the descriptive data mining that is an unsupervised process, predictive data mining is very much a supervised process. Predictive data mining not only discovers present patterns and information in the data, it attempts to solve problems. Through the existence of modeling processes in the analysis, predictive data mining can answer questions that cannot be answered via other techniques.

Tools that are used in the predictive data mining process include decision trees, neural networks, genetic algorithms and fuzzy systems. Decision trees are ideal for solving problems that can be dissected into a logical progression of events. The existing types of decision tree include Chi-Square Automatic Interaction Detection, developed by Kass [7], Classification and Regression Trees, CART, originally developed by Breimam and Friedman and enhanced by Olshen and Stone [8], and C4.5 developed by Quinlan [9].

Neural networks include several types that can be used to solve different kinds of problems. The most popular neural network (actually a learning algorithm) is backpropagation. Most neural networks in this and most other industries use the backpropagation algorithm. They are easy to use and understand and are capable of solving many complex problems. Other neural networks that are also used to develop models based on historical data are radial basis functions, general regression neural networks and probabilistic neural networks. The nature of the data that is being studied, as well as the complexity of the problem, usually dictates the type of neural network that should be used during the model building process. Genetic algorithm (as one of the many analytical tools known as evolutionary computation) is an intelligent search and optimization tool that has proven to be indispensable for any data mining study.

The role of fuzzy logic in data mining and knowledge discovery analyses cannot be over-emphasized. Fuzzy logic should form the data miner's basic platform and be used as the fundamental approach to studying complex problems. Fuzzy logic plays an important role in descriptive, as well as predictive, data mining. If one agrees that the data one deals with are instances of reality and nature, and that the complexity of reality and nature cannot sufficiently be explained using the binary system of belief, then, fuzzy logic becomes the most important tool in data mining studies. A more complete look at the role of neural networks, genetic algorithms and fuzzy logic has been provided by the author in [10-12] and there are many books that provide a good background for those interested in this exciting technology [13,14].

COMPONENTS OF AN INTEGRATED DATA MINING TOOL

An integrated data mining tool must include the following components. Furthermore, these components

must allow such interaction that the result of each component can be used in other components.

1. It must include a module that allows the user to import data from different sources and combine them into a table that can be used during the analysis;
2. It must have a data cleansing module, which is one of the most important modules of an integrated data mining tool. The quality of data being used in the analysis determines the degree of success of a data mining project. Essential algorithms required for a data cleansing module include:
   - 2.1 Identification and remediation of missing data,
   - 2.2 Identification and remediation of contaminated or erroneous data,
   - 2.3 Identification and remediation of outliers.
3. It must have a module for the identification of Key Performance Indicators (KPIs). KPI identification is a process during which all independent variables are analyzed and ranked, based on their degree of influence on the dependent variable;
4. It must have a clustering module, which must at least include the following clustering algorithms:
   - 4.1 Hierarchical cluster analysis,
   - 4.2 k-mean cluster analysis,
   - 4.3 Fuzzy c-mean cluster analysis.

   There are two issues regarding Cluster Analysis:
   - A. Two important pieces of information should be known (or it is usually guessed) prior to performing the analysis. These are the number of clusters and variables (dimensionality) that would provide the best separation of the data;
   - B. Cluster analysis essentially is an unsupervised process. It would be ideal to have access to a tool that could provide supervised cluster analysis. In some petroleum engineering problems, supervised cluster analysis can be of great help, such as in the prediction of high permeability streaks in carbonate reservoirs from well-log and the processing of 4D seismic data. More details about these issues are outside the scope of this article and will be discussed in future papers.
5. It must have an integrated neural network module. The integration of results from KPI and cluster analysis modules in the neural network module is of immense importance. The integration of KPI in the analysis allows identification of the lowest number variables to be used in the predictive modeling process. Furthermore; integration of cluster analysis algorithms can play a key role in the predictive neural network model by optimizing the data partitioning process. The neural network module should include different algorithms for training. The unsupervised algorithms are useful for descriptive data mining, while supervised algorithms are essential for predictive data mining processes. The following neural networks are suggested for an ideal integrated data mining tool:
   - 5.1 Kohonen Self-Organizing maps;
   - 5.2 Backpropagation neural networks. This algorithm has several variations that have proven to be very useful in several applications. It is strongly recommended that all the variations of this popular algorithm be present in such a tool;
   - 5.3 General regression neural network. There is a variation of this algorithm that can help data miners when the number of data records is limited. This variation can be of enormous importance to some oil and gas problems;
   - 5.4 Radial basis function neural networks;
   - 5.5 Probability neural networks.
6. It must have an integrated genetic algorithms module. It is very important that the genetic algorithm module can communicate with the neural network module and be able to use available neural network models as its fitness function;
7. It must have an integrated fuzzy system module. The fuzzy system has to be integrated such that it can use the results of the cluster analysis, neural networks and genetic algorithms during the development process. The fuzzy module should provide means for automatic and user defined fuzzy set definitions and rule identification.

DATA PRE-PROCESSING

Data pre-processing is one of the most important components of a data mining process. It usually consumes more than 50% of the project. During detailed and thorough data pre-processing, the data miner has to study the data very carefully and identify the missing cells (data element) in a data record. Sometimes, especially when the number of data records is limited, a few missing cells can result in eliminating entire data records from the analysis. Furthermore, it is not very easy to detect missing data elements in a large database and, if they go undetected, their damaging effects will be noticed far into the analysis, at which time the issue must be addressed and the analysis repeated from the beginning.

There are several ways of patching the missing data. The most commonly used method is the conventional statistical method that would substitute an average value of the parameter for the missing data

element. This is not the most appropriate way of solving this problem. In many cases this method can over simplify the problem and result in erroneous outcomes, especially during predictive data mining. There are other methods that can be used, in order to preserve the integrity of the data record, while substituting the missing data element with the most appropriate (which means the least damaging) values. The objective of these new techniques is not to magically find the missing data element and substitute it in the database. These techniques, using a combination of neural networks and genetic algorithms, identify the best value that can be substituted for the missing value, while the information content of the rest of the data records remains valid and usable.

The identification and handling of outliers is another data pre-processing task that is of the utmost importance in a data mining study. It is important to identify whether or not a data record is truly an outlier or if it carries information about the behavior of the system under specific conditions. Domain expertise can become very important in judging correctly. Furthermore, it should be identified which data element in a particular data record contributes to it becoming an outlier. The consistency of a particular data element in making a data record to appear as an outlier can provide important information in handling that data element. This can be identified by plotting all the parameters involved in a database against one another and by studying the behavior of the data record in question. If it is concluded that certain data elements are causing a data record to appear as an outlier, then those elements may be treated as erroneous and can be dealt with as missing data, as mentioned previously.

Another important pre-processing problem is the identification of contaminated or erroneous data records. Sometime, for whatever reason, some data records appear in certain databases where they do not belong, by mistake (due to human or machine error), which can cause havoc. Such data records may look very much like other data records but carry fundamentally different kinds of information. Combining different databases from different sources, which are usually a result of recent mergers and acquisitions, can be one of the sources of such problems, since different companies use different conventions for record keeping. Sometimes such problems can be solved (of course they have to be identified first and, if unnoticed, they can modify the results of the analysis) by referring to those involved in the data collection process or those familiar with the data. Otherwise, the detection and handling of such issues can become very time consuming. A simple example of such a case took place in a company that merged data from a recently purchased property into its main database. The depth reference in two databases were not consistent (one was from KB, while the other was from sea level), which was causing issues during the analysis. Fortunately, an engineer from the newly purchased company was on the staff and was able to address the problem quickly. In the case of the sample field study presented in this paper, the contaminated data records were an important problem that took the research team many months to resolve. The team developed a new and novel technique [15], using neural networks and a fuzzy cluster analysis, to solve this problem. This technique is applicable to any dataset. Figure 3 demonstrates the components involved in the data pre-processing of a data mining analysis.

STATISTICAL ANALYSIS

During the past decades, petroleum engineers, geophysicists and geologists have come to realize the importance of geostatistics [16] in day to day operations when dealing with hydrocarbon producing reservoirs. The statistics of data being used in a data mining study provide the analyst with valuable information. Figure 4 shows the different components and measures of a statistical analysis that should be used in a data mining study. The linear regression between different parameters in a database can reveal any visible and readily detectable relationship that might exist between different parameters. As the relationship between parameters becomes more complex, tools, like linear regression, will no longer be useful and the data miner will need to use more sophisticated techniques. These techniques include principal component analysis [17], fuzzy curves [18] and fuzzy combinatorial analysis [19].

One of the most common and basic statisti-

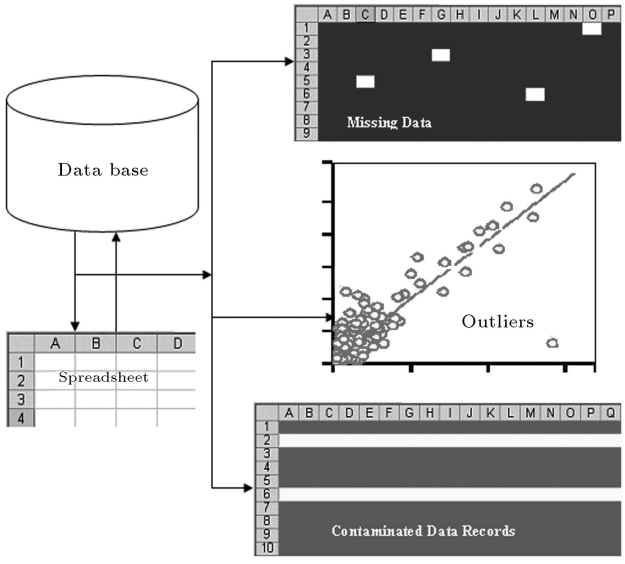

Figure 3. Important components of data pre-processing.

cal analyses that are performed on all parameters in the database is the identification of their probability distribution function, along with minimum, maximum, mean and variance. Based on the nature and type of the parameter (continuous versus categorical data types), the mean, median or mode of the parameter is calculated. Chances are that the distribution for most of the parameters does not follow the characteristics of a normal distribution. In such cases, the calculation of distribution kurtosis can help the analysts in their analysis as shown in Figure 4.

DESCRIPTIVE ANALYSIS

Two of the most common methods for descriptive analysis of the data are cluster analysis and KPI identification, as shown in Figure 5. Clustering describes a collection of unsupervised methods, whose aim is to partition an overall data set into a significantly smaller number of "clusters". These methods, in general, require some kind of distance measure among the data entities, in order to group them together and identify each data entity with one cluster.

Most clustering algorithms partition the data based on how similar individual records are; the more similar, the more likely they belong to the same cluster. Their main purpose is to identify clusters which maximize the inter-cluster distance and minimize the intra- cluster distance, so that clearly distinct groups of similar entities are obtained. This grouping introduces a "natural" unsupervised classification schema, based on similarities, according to the given distance measure [20].

There are several types of clustering technique, such as hierarchical clustering, k-mean clustering and fuzzy c-mean clustering. Hierarchical clustering does not partition data into a particular number of clusters in a single step. Instead, a series of partitions take place, which may run from a single cluster containing all objects to "n" clusters, each containing a single object. k-mean clustering divides the database into k clusters identified by the user, such that the distances between the objects in a cluster are minimized, while

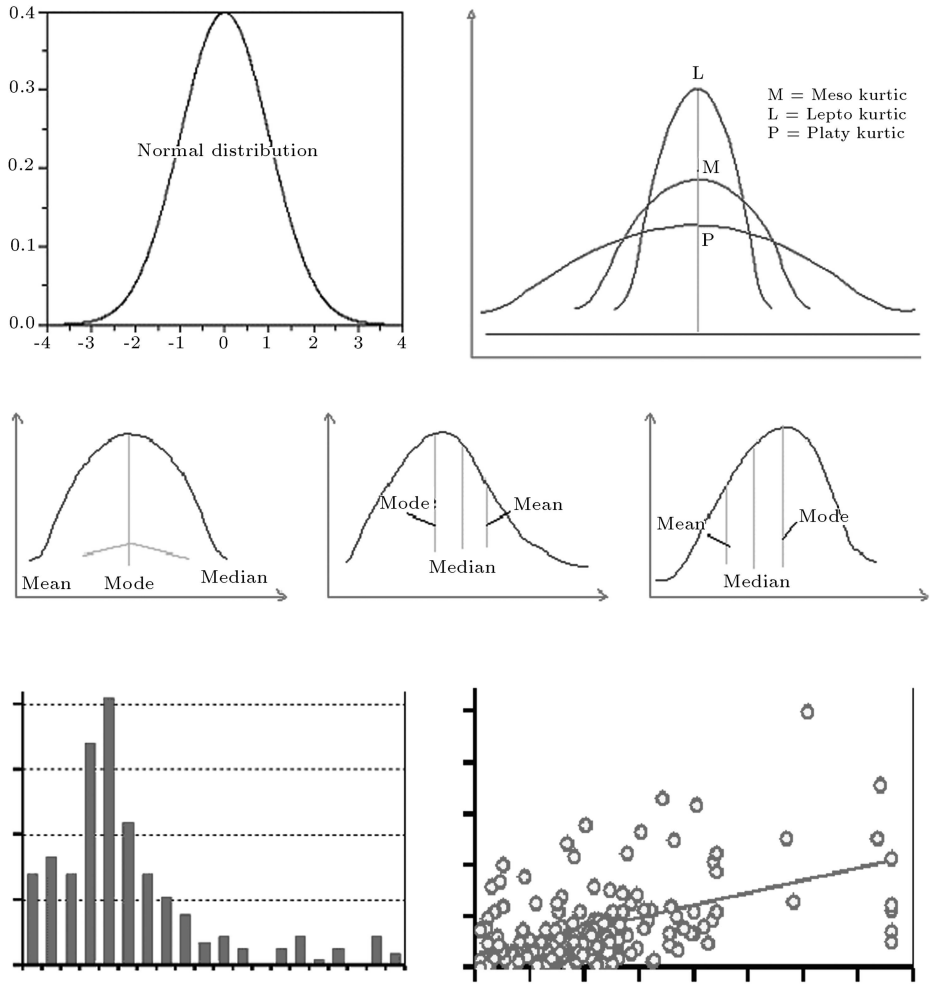

Figure 4. Statistical analysis as part of data mining.

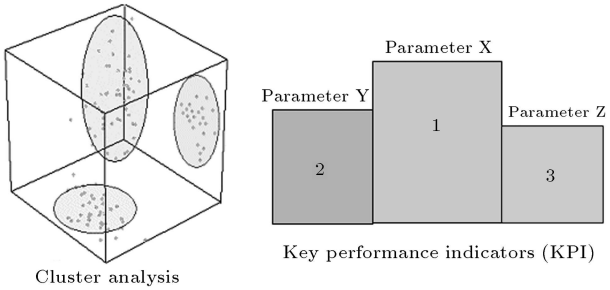

Figure 5. Cluster analysis and key performance indicator are an important part of descriptive analysis.

the distance between clusters are maximized. In k-mean cluster analysis, each object (a data record) will fully belong to only one of the clusters. In a fuzzy c-mean clustering, data records are assigned to different cluster centers to a degree. In this case, a data record may have a membership of 0.7 in one cluster while having memberships of 0.05 and 0.25 in two other clusters [21].

The other descriptive analysis is identification of Key Performance Indicators. It is also known as the identification of performance drivers or the identification of parameter influence. It is the belief of the author that a reliable technique in ranking the importance of parameters in an oil and gas related database (due to the nature and complexity of the problems in this field) must take into account the influence of the parameters on one another, as they collectively influence the outcome. As an example, in a hydraulic fracture treatment, several additives are used throughout the treatment. The ultimate objective of these additives is to enhance the hydraulic fracture treatment outcome (fracture length or conductivity). The additives will have an effect on the fracturing fluid but, in many cases, they will have an interfering effect on each other. This interfering effect must be taken into account when feature ranking analysis is to take place. Another industry that can benefit from such analysis is the pharmaceutical industry, as in testing the effect of different drugs on patients. Fuzzy Pattern Recognition [19] takes such interfering effects into account.

PREDICTIVE ANALYSIS

In order to perform predictive data mining analysis, the analyst must develop a predictive model. There are many techniques for developing predictive modeling. First and foremost are deterministic models, which are familiar. These are models that have their basis in physics and are developed through rigorous mathematical manipulation of physical concepts. If one were involved in a data mining project, the chances are that developing a deterministic, physics-based model would be impractical, either due to the complexity of the problem at hand or for the unavailability of needed information. This is when data-driven modeling becomes the best alternative.

Neural networks have proven to be a great tool for data-driven model development. A successful neural network modeling process will require all the information that is generated during the descriptive data mining process. This information must be integrated into the statistically representative partitioning of the database in the training, calibration (testing) and verification (validation) of the developed model, in order to increase the possibility of developing a representative model.

Most predictive data mining activities take place upon completion of the predictive model development process. This is when many "what if" scenarios can be played out and detailed sensitivity and parametric analyses based on multiple parameters can be performed. During such analysis, two dimensional plots can show the sensitivity of the outcome as a function of all possible values of a parameter. Using three dimensional plots, sensitivity analysis can be performed on two parameters at a time studying the changes in the outcome as a function of two parameters. When it is desired to study the effect of several parameters (more than two) on the outcome simultaneously, Monte Carlo simulation can be used to show the potential probability distribution of the outcome as a function of several parameters, each of which must be assigned a probability distribution function.

The next potential step in a predictive data mining study is to use the data driven model in an optimization study. Employing an integrated genetic algorithm routine and identifying the parameters that are to be optimized, the data driven model can be used as the fitness function of the genetic algorithm, in order to find the optimum combination of values for the parameters being studied, so as to achieve the objectives of the optimization study. Genetic algorithms can be constrained through pre-determined rules, which would make sure that the solutions suggested by the optimization process pertain to the realities on the ground.

Another approach would be the development of hybrid robust models that would augment the data-driven models with expert knowledge. This data-knowledge fusion technique can markedly enhance the performance of predictive models [22]. Using fuzzy logic, the expert knowledge can be coded into semantics and used in fuzzy systems that can combine the results of data-driven models with expert knowledge, resulting in superior predictive models. This would prove to be valuable in cases where the available data does not necessarily cover all the possible cases that may exist in a process.

EXAMPLE IN DJ BASIN, NIOBRARA/CODELL FORMATIONS

The example application discussed here is a short summary of the results of a study that has been presented in a previous SPE paper [4]. Its summarized presentation is mainly for the completeness of this paper. The study was conducted in the DJ basin, Niobrara/Codell formations. The Patina Oil and Gas Company has been one of the most active operators in the United States in the re-stimulation of tight gas sand wells. Patina has over 3,400 producing wells in the basin and has, so far, re-stimulated over 230 Niobrara/Codell completions.

The original database in the DJ basin needed considerable quality control in order to remove erroneous records [15]. Once the quality control of the data set was completed, Fuzzy Pattern Recognition was performed. The analysis was performed for a combination of up to five features, meaning that the influence of the parameters (up to five parameters) on each other, as well as on the outcome, was examined.

Table 1 shows the parameters that ranked as the top 10 in the database. This table shows that seven out of the top ten parameters that seem to be controlling the hydraulic fracturing outcome in this field are operational parameters that can be controlled by the operator. The other three (Lat., Long. and Codell Gas-ft) are indicators of geology and reservoir quality. This provides important information to the operator as to which parameters should be concentrated on, in order to get the most out of the stimulation jobs.

Upon completion of a cluster analysis on the database, a predictive model was built for this field. The predictive model was then used to study several "What If" scenarios. Two, three and multi-dimensional (using Monte Carlo simulation) sensitivity analyses were performed on the predictive model for each of the wells involved in the analysis. Then, the cluster analysis results (which identified three categories of similar wells in the field), were used to perform an analysis to find the best practices for the cluster of wells. Finally, another set of best practice studies were performed for the entire field.

Table 1. Most influential parameters in the DJ basing hydraulic fracturing program.

Rank	Parameter
1	Flow back volume, (bbl)
2	Codell gas-ft
3	Bicarbonate (ppm)
4	Peak viscosity
5	Latitude
6	Amount of sand (Mlbs)
7	Longitude
8	Date of refrac
9	Viscosity shear 100-30 min.
10	Total hardness (ppm)

The full field analysis revealed that higher amounts of proppant (after a certain concentration has been reached) do not contribute to the success of the frac jobs and, therefore, are not recommended. Another conclusion, based on the full field analysis, was the use of low viscosity fluids in the frac jobs. The source of the water used in the operation was also shown to play an important role in the success of the frac jobs. Once these items were identified as parameters that controlled the success of the stimulation jobs, the predictive model was used to design the frac jobs for new wells in the field. The importance of the above conclusions are realized when one looks at the type of data that such conclusions have been drawn from. Figure 6 shows the scatter plot for the amount of proppant and peak viscosity versus the post-stimulation deliverability. These figures show that there are no visually detectable trends in this data and without tools such as those used in this study (an integrated data mining software application specifically for the oil and gas industry currently under development) identification of such patterns and making such recommendations would have not been possible.

A NOTE ON APPLICATION OF DATA MINING TO RESERVOIR STUDIES

There are other issues that may arise when performing data mining during reservoir studies that are above and beyond the topics covered in the previous sections. As an example, two of such issues will be briefly addressed in this section. The first issue is up-scaling, which is often encountered by geo-scientists and reservoir engineers, mostly during the reservoir modeling process and the second issue is the handling of soft data, such as geological knowledge and interpretations. Data mining, in its conventional and traditional realm, does not offer any specific tools for addressing these issues, but reservoir engineers and geo-scientists who use data mining approaches as part of their solution toolbox, may create new techniques inspired by existing data mining paradigms in order to tackle problems such as those mentioned here.

For example, in the case of up-scaling, the conventional approaches that are currently being used vary from a simple averaging of the reservoir parameters to the development of pseudo-functions [23,24], in order to achieve the objectives of the reservoir studies. In order to address up-scaling using data mining tools, one may use neural networks (in the case of having

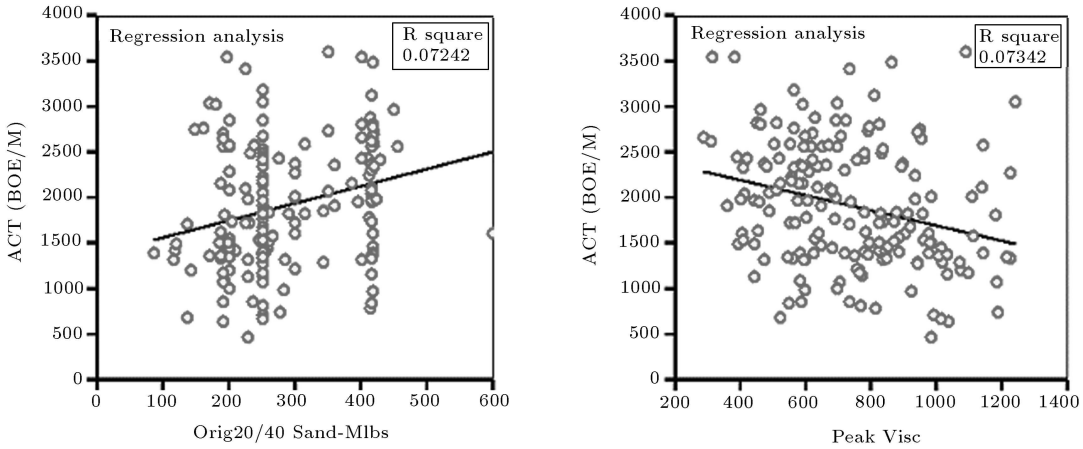

Figure 6. Proppant amount and peak viscosity data versus post-frac deliverability.

sufficient data [23]) to develop customized up-scaling models. Such models may then be tested to evaluate their applicability in different parts of the reservoir. Following is a conceptual example of developing a neural network-based up-scaling model for relative permeability, in cases where sufficient data can be compiled. Relative permeability has been selected as the topic of this example, since its up-scaling is a complex task.

The process would start with performing a relative permeability analysis on a whole core or a block of outcrop. The result of such measurements would be the up-scaled relative permeability function and the output of the neural network model. Then, multiple samples of small core plugs could be obtained from the whole core or the outcrop. The core plugs are then individually tested to measure the relative permeability curves for each sample. The relative permeability curves of each core plug, which result from this exercise, provide the input to the neural network model. Therefore, the neural network model will be trained to correlate the relative permeability characteristics of core plugs (high resolution, small scale) to the relative permeability characteristics of a large sample (low resolution, large scale). Of course, this is not the only way that up-scaling may be addressed using data mining tools, rather, only one example of many possible solution strategies. The conceptual design of such an up-scaling model is shown in Figure 7. Training such a neural network model presents several challenges. For example, in the conceptual design presented in Figure 7, input and output are in the form of relative permeability profiles rather than in real numbers. There are two ways to address this issue:

1. Use Linear Vector Quantization (LVQ) [25] in order to represent the relative permeability profiles, both input and output;

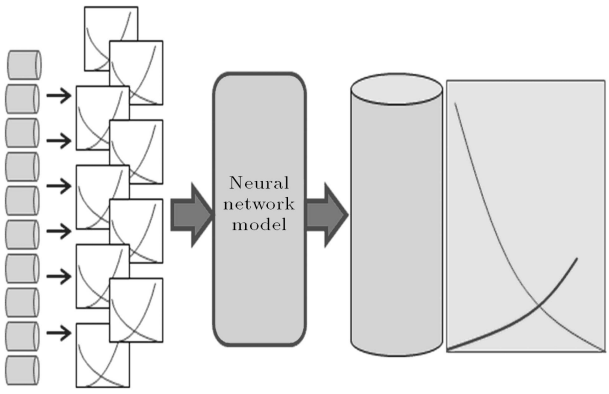

Figure 7. Conceptual design of an up-scaling model for relative permeability.

2. Present relative permeability values at predetermined water saturation points as real numbers.

Furthermore, one might want to consider using a recurrent neural network, in order to accommodate the fact that relative permeability is a path function rather than a point function. Since recurrent neural network architecture has been developed to handle time series, it would be the most appropriate architecture for problems such as up-scaling relative permeability functions.

As for processing soft data, such as geological knowledge and interpretations, the most appropriate tool is the fuzzy set theory [10]. Using a fuzzy set theory, one can define variables with semantics and develop fuzzy expert systems to perform calculations, using his/her observations in terms of fuzzy rules. An example of such a system was presented in a paper dealing with well log interpretations [26]. In this paper, the author developed a fuzzy expert system, using parameters such as shale, shaley-sand, sandy-shale and

sand, to describe the petro-physicist's observation of a particular rock quality.

CONCLUSIONS

Data mining and knowledge discovery has much to offer the oil and gas industry. The technology is fairly new in this industry and can, and will be misused, as part of a natural growth. Many processes that are not necessarily data mining in nature or include only small components of data mining will be called data mining in an effort to ride the waves and the hype surrounding the technology. This should not deter the industry from aggressively pursuing this exciting technology, which could bring about a new way of looking at many different aspects of this industry, from technical problem solving to management level decision-making.

The next step in the evolution of this technology is the development of specialized and integrated software tools that can help engineers and scientists get the most out of several analytical techniques being offered by this technology. The essential components of such a tool were outlined in this paper. Any application that lacks several of these components cannot be considered an integrated data mining tool. On the other hand, there might be other techniques that might have been omitted in the list of components provided in this document. As new techniques prove themselves in other industries and show the promise of having the potential to contribute to a data mining process, they can be added to the list of essential components mentioned in this document.

REFERENCES

1. *According to IDC, World's Leading Provider of Technology Intelligence, PC AI magazine*, **17**, pp 23-25 (January 2003).
2. Berry, M.J.A. and Linoff, G.S., *Mastering Data Mining*, Wiley Computer Publishing, John Wiley & Sons, Inc., New York, NY (2000).
3. Westphal, C. and Blaxton, T., *Data Mining Solutions*, Wiley Computer Publishing, John, Wiley & Sons, Inc. New York, NY (1998).
4. Mohaghegh, S., Popa, A., Gaskari, R., Ameri, S. and Wolhart, S. "Identifying successful practices in hydraulic fracturing using intelligence data mining tools; application to the Codell formation in the DJ basin", *SPE 77597, Proceedings, SPE Annual Conference and Exhibition*, September 29-October 2, San Antonio, Texas (2002).
5. Frawley, W., Piatetsky-Shapiro, G. and Matheus, C., *Knowledge Discovery in Databases: An Overview. AI Magazine*, pp 213-228 (Fall 1992).
6. Mohaghegh, S.D. "A new methodology for the identification of best practices in the oil and gas industry, using intelligent systems", *Journal of Petroleum Science and Engineering*, pp 239-260 (Dec. 2005).
7. Kass, G.V., *An Exploratory Technique for Investigating Large Quantities of Categorical Data Applied Statistics*, **29**, pp 119-127 (1980).
8. Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J., *Classification and Regression Trees*, Wadsworth International, Belmont, Ca (1984).
9. Quinlan, J.R., *C4.5: Programs For Machine Learning*, Morgan Kaufmann, Los Altos (1993).
10. Virtual Intelligence Applications in Petroleum Engineering: Part 3 - Fuzzy Logic, *Journal of Petroleum Technology, Distinguished Author Series*, pp 82-87 (November 2000).
11. Virtual Intelligence Applications in Petroleum Engineering: Part 2 - Evolutionary Computing, *Journal of Petroleum Technology, Distinguished Author Series*, pp 40-46 (October 2000).
12. Virtual Intelligence Applications in Petroleum Engineering: Part 1 - Artificial Neural Networks, *Journal of Petroleum Technology, Distinguished Author Series*, pp 64-73 (September 2000).
13. Berthold, M. and Hand, D.J., *Intelligent Data Analysis: An Introduction*, Springer Verlag, 2nd Ed., April 15 (2003).
14. Hastie, T., Tibshirani, R. and Friedman, J., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Verlag, August 9 (2001).
15. Popa, A., Mohaghegh, S.D., Gaskari, R. and Ameri, S. "Identification of contaminated data in hydraulic fracturing databases: Application to the codell formation in the DJ Basin", *SPE 83446, Proceedings, SPE Western Regional Conference and Exhibition*, May 20-24, Long Beach, California (2003).
16. Jensen, J.L., Lake, L.W., Corbett, P.W.M. and Goggin, D.J., *Statisitcs for Petroleum Engineers and Geoscientists*, 2nd Ed., Elsevier, Amsterdam, The Netherlands (2000).
17. Jolliffe, I.T., *Principal Component Analysis*, Springer Verlag, New York (1986).
18. Lin, Y. and Coningham, G. "A fuzzy approach to input variable identification", *Proceedings of the Third IEEE International Conference on Fuzzy Systems*, June 26-July 2, Lake Buena Vista, Florida (1994).
19. Mohaghegh, S., Gaskari, R., Popa, A., Ameri, S. and Wolhart, S. "Identifying best practices in hydraulic fracturing using virtual intelligence techniques", *SPE 72385, Proceedings, SPE Eastern Regional Conference and Exhibition*, October 17-19, North Canton, Ohio (2001).
20. *Similarity Clustering*, Thomas Prang 1998. http://www-lehre.informatik.uni-osnabrueck.de/~ftprang/papers /tproject/node27.html
21. Kuncheva, L.I., *Fuzzy Classifier Design, Studies in Fuzziness and Soft Computing*, Physica-Verlag, New York, NY (2000).

22. Mohaghegh, S., Reeves, S. and Hill, D. "Development of an intelligent systems approach to restimulation candidate selection", *SPE 59767, Proceedings, SPE Gas Technology Symposium*, Calgary, Alberta (April 2000).

23. Tidwell, V.C. "Heterogeneity, permeability patterns, and permeability upscaling: Physical characterization of a block of massillon sandstone exhibiting nested scales of heterogeneity", *SPE Reservoir Engineering & Evaluation*, **3**(4), pp 283-291 (2000).

24. Baker, R.O. and Moor, R.G. "Effect of reservoir heterogeneities on flow simulation requirements", *SPE 35413, SPE/DOE Improved Oil Recovery Symposium*, Tulsa, Oklahoma (21-24 April 1996).

25. Etemoglu, C.O. and Cuperman, V. "Structured vector quantization using linear transforms", *IEEE Transactions on Signal Processing*, **51**(6), pp 1625-1631 (June 2003).

26. Mohaghegh, S.D., Popa, A., Gaskari, R., Wolhart, S., Siegfried, R. and Ameri, S. "Determining in-situ stress profiles from logs", *SPE 90070, Proceedings, SPE Annual Conference and Exhibition*, September 26-September 29, Houston, Texas (2004).