SPE-167839-MS

Advanced Machine Learning Methods for Production Data Pattern Recognition

Niranjan Subrahmanya, ExxonMobil Research and Engineering Company, Peng Xu, ExxonMobil Upstream Research Company, Amr El-Bakry, ExxonMobil Production Company, Carmon Reynolds, ExxonMobil Information Technology

This paper was prepared for presentation at the SPE Intelligent Energy Conference and Exhibition held in Utrecht, The Netherlands, 1–3 April 2014.

Abstract

An important challenge for asset management is to analyze large amounts of data in a short period of time to provide insightful information for decision making in a timely fashion. Analyzing all available data manually is impractical and inefficient. It is advantageous to develop pattern recognition algorithms to recognize events-of-interest to achieve effective asset management.

Conventional pattern recognition algorithms usually require a fairly large training set in which data points are carefully prepared and “labeled”. Examples include designating an equipment’s status as healthy or faulty by subject-matter experts. This can make the process time consuming and error-prone when the training set is large. For most applications, a small amount of data points are already labeled by the experts through their routine activities. While these data points are usually not enough to form a training set for conventional pattern recognition methods, some of the newer methods can take advantage of them along with the hidden manifold structures manifested by the unlabeled data. Moreover, subject matter experts may be willing to provide more input if a clear indication of the value of information and a manageable subset of data is pre-selected for their inspection. In fact, many other industries are facing the same challenge where the cost of acquiring labels is too expensive to be practical and large amounts of unlabeled data and limited expert time are available. A suite of advanced machine learning algorithms (e.g., semi-supervised learning, active learning) have been developed to tackle this challenge, and many of them have been successfully used for various applications in the past few years. In this paper, we will review the concepts and report our observations about the effectiveness of these methods in a real-world asset management scenario. We consider well test validation in an asset with a large number of tests as an example of a label-rich data set that can serve as the basis for our numerical review of existing methods. In this example we will specifically look at the task of

Exxon Mobil Corporation has numerous subsidiaries, many with names that include ExxonMobil, Exxon, Esso and Mobil. For convenience and simplicity in this paper, the parent company and its subsidiaries may be referenced separately or collectively as "ExxonMobil." Abbreviated references describing global or regional operational organizations and global or regional business lines are also sometimes used for convenience and simplicity. Nothing in this paper is intended to override the corporate separateness of these separate legal entities. Working relationships discussed in this paper do not necessarily represent a reporting connection, but may reflect a functional guidance, stewardship, or service relationship.

building a statistical model to recognize the validity of rate measurement tests in a test separator. In this case, through their daily activities, the operators have labeled most of these tests as valid or invalid. The extensive amount of well test validation data provides sufficient information to assess the newer approaches under review. The plan then is to apply a similar approach to tasks such as equipment health monitoring to identify pump failures with limited expert input.

Conceptually, reduction in labeled input can be achieved by combining the information from the labels and the statistical distribution of the data (e.g., clusters). As an extreme example, consider that the pump measurement data may show two distinct clusters and the operators have labeled a few data points in one cluster as pump failures when reports had to be made due to wells being shut in. This information is sufficient to label one of the clusters as healthy and the other one as faulty. For a new measurement, a prediction may be made by first determining the cluster to which the measurement belongs and then assigning it the corresponding label. While most real world problems are much more challenging than this example due to the number of data points, dimensionality of the data, lack of clear cluster structure and potential ambiguity of data structures, similar ideas can be used to develop highly accurate statistical models with a limited number of labels.

Introduction

Efficient management of upstream assets with near real-time decision making is a challenging task due to the number and complexity of the components and processes involved. The multitudes of technical challenges to be addressed include topics such as data measurement and storage, advanced data analysis for asset health monitoring and prognostics, and finally performance optimization. ExxonMobil has a long history of applying a disciplined and value-driven approach to the deployment of advanced technologies in asset management. In order to effectively implement and sustain these advancements, we are concurrently working on addressing both the technical challenges mentioned above as well as organizational and change management concerns, and this has been reported in a number of prior publications (Shyeh et al., 2008; Killian et al., 2012; El-Bakry et al., 2012).

The focus of this paper is the use of advanced machine learning methods to enable novel workflows that reduce the burden on engineers and operators during the development and sustainment of data-driven models and encourage the adoption of advanced analytical methods. The value of artificial intelligence (AI) and machine learning (ML) based statistical models in improving asset management decisions has been established in multiple case studies. One of the main challenges for the development and use of models created through conventional pattern recognition algorithms is the requirement of a fairly large training set in which data points are carefully prepared and labeled by subject-matter experts (e.g., an equipment condition would have to be labeled as healthy, faulty etc.). For most applications in well instrumented assets, a large amount of data is usually historized and available but only a small portion of this data is already labeled by the experts through their routine activities (e.g., through a report when a well had to be shut in due to pump failure). Hence, in order to initiate the development of a new data-driven model, it is important for an expert to first go through the historical data, get rid of outliers and bad measurements, and finally label it. This process is very time consuming as it may involve cross referencing the historical data with other available information sources, such as operator reports, and can be error-prone when the training set is large. Moreover, labeling may not be consistently done due to differing experience level or time availability. This last point may significantly impact the quality of such models.

This problem of acquiring labeled data is not unique to the oil and gas industry and has been recognized as a major bottleneck for the use of machine learning algorithms in many other industries. Hence there has been significant research in the last decade to address this problem. We are going to explore two types of methods; one is active learning and the other one is semi-supervised learning. An active learning method analyzes available data and selects a few unlabeled points that are distinct from each other and are able to provide significant value in terms of increasing model accuracy. The method then presents these data points to an expert, who will label them. Given the newly labeled data and existing labeled data, the active learning method analyzes the data again and presents more unlabeled data points to the expert. This process will iterate until the method has identified the structure in the data or a stop criterion is reached. By contrast, a semi-supervised learning method does not involve active user engagement. It assumes that some data points are already labeled by the expert and it tries to conduct classification or clustering analysis by incorporating information from these data points as well as the structure in the data set. In this paper, we propose a smart workflow that uses both methods in tandem. Our approach exploits

the strength of both methods to ensure the high quality of resulting models while interacting with the expert in the most efficient and effective way.

Concept

The rapid proliferation of sensing and data acquisition systems as well as digital storage systems has resulted in the acquisition of huge data sets in various sectors of the economy. Expert time to comb through the data and provide informative labels is highly limited, hence, methods that go beyond the classical supervised learning approach have become a key aspect of developing large-scale data-driven models for applications such as image recognition, speech recognition, document classification and drug discovery. The technology developed to address these challenges can be beneficial to applications in the oil and gas industry as well. Two significant complementary approaches to addressing this problem are "semi-supervised learning" and "active learning", and the scenarios under which they are applicable are shown in Figure 1.

Semi-supervised learning considers scenarios when the size (number of data points) of available inputs to a statistical model is larger than the size of available outputs. In this case, the classical approach would be to discard the part of the data without labels and reduce the size of the training set before applying a supervised learning approach. As shown in Figure 1b, this may result in wasting a lot of good data and reduce the accuracy of the developed data-driven model. On the other hand, unlabeled data may contain information about the structure of the data, for example, the data may contain clear clusters and in that case one only needs one label per cluster to learn the model accurately. This is of course an extremely favorable scenario and in general it is necessary to exploit more subtle structures in the data to extract the maximum use from unlabeled data. A popular hypothesis that is exploited by many semi-supervised learning algorithms is that the decision boundary should pass through regions of low data density. This allows discovery of more complicated decision boundaries using the unlabeled data as illustrated in Figure 2, which shows the results of semi-supervised learning with just 15 labeled points for a 3 class problem (5 labels available per class) vs. supervised learning using the same labels. It can be seen that the semi-supervised method is able to combine the manifold structure of the data (which is easily visible in this 2-dimensional data but is usually not apparent in higher dimensional real-world datasets) with the provided labels. In general, semi-supervised methods try to extract more information from unlabeled data by making strong assumptions about the nature of the underlying data. Therefore, it is important to make sure that the assumptions match the data structure to get the best results. For a detailed survey of various semi-supervised methods, the interested reader is referred to Zhu and Goldberg (2009) and references therein.

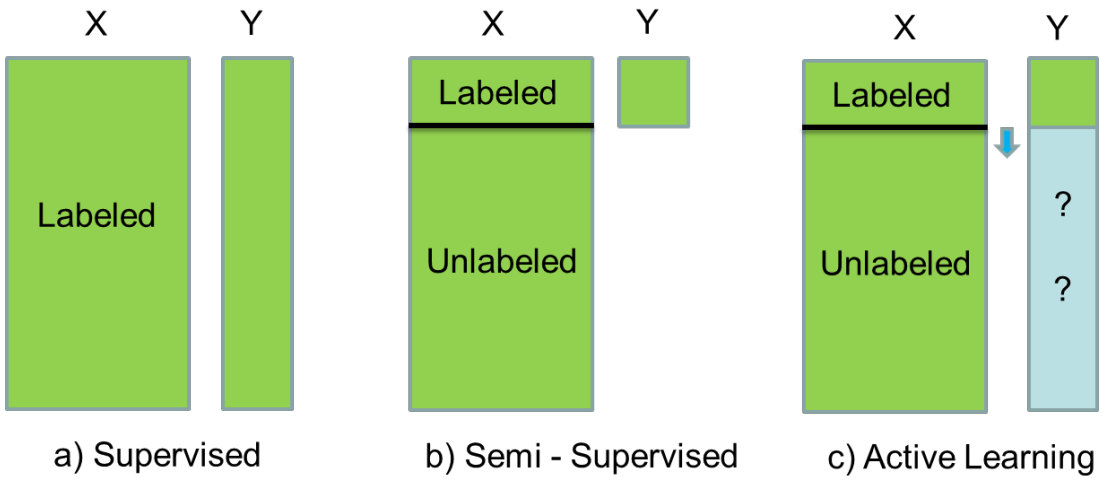
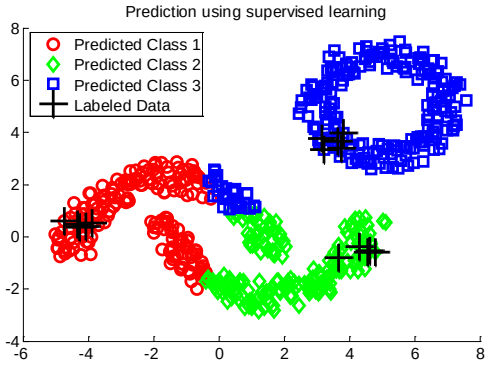
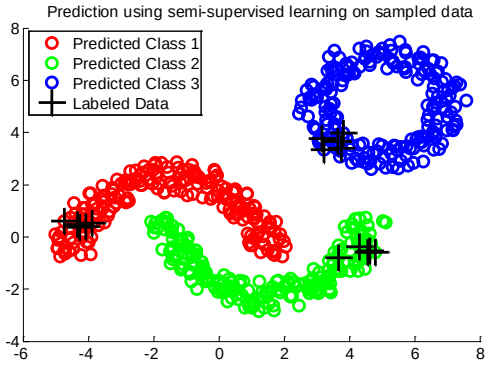


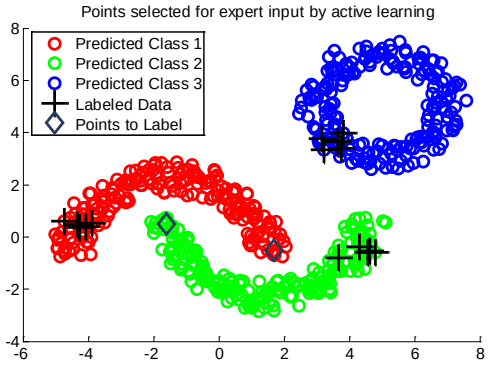
Figure 1. Various scenarios for available training data while learning a statistical model for Y as a function of X. Y is the label for the data (such as healthy, faulty etc.) and X contains the measurements (such as pressure, temperature etc.) a) Classical supervised learning setting (Y is available for all X). b) Semi-supervised learning setting (Y is available only for a fraction of the available X) c) Active learning setting (Y is available only for a fraction of the available X, but more Y values can be obtained if required at a certain cost)



a) Class prediction using supervised learning



b) Class prediction using semi-supervised learning



c) Actively selected points for expert input

Figure 2. Comparison of supervised (support vector machines with Gaussian kernel) vs. semi-supervised learning results on a toy data set. The black plus markers indicate the data points where labels are available (5 per class in a very local area). The semi-supervised method combines this information with the structure of the unlabeled data (all the shown data) and the hypothesis that the class boundary should pass through low data density regions. The active learning method then selects the points with highest uncertainty for expert input and verification.

Active learning is also applicable when the size (number of data points) of available inputs to a statistical model is larger than the size of available outputs (Figure 1c). However, unlike semi-supervised learning, active learning assumes that more labels can be obtained by an expert or “oracle” for any data point but each such query incurs a certain cost. Therefore, the objective for active learning is to obtain the same model quality as one would if all the data were labeled but with a much fewer number of actively queried samples. This is usually achieved by sequentially querying labels for points with the highest amount of uncertainty or the most information content (as shown in the example in Figure 2c). For a detailed survey of various active learning methods, the interested reader is referred to Settles (2009) and references therein.

Example Application

The application chosen to demonstrate this approach is an automated quality assessment module for rate measurements in test separators using advanced analytics and real-time data. The module is for an unconventional asset that records data from thousands of tests per day. The physical set-up of the system is as shown in Figure 3 and fluid production from each well is sequentially directed to the test separator for about a couple of hours. The high-frequency measurements made during the test include water rate, oil rate, fluid temperature and water content in oil. This high-frequency data is then put through some signal processing and feature extraction modules to extract the salient features of each test such as the average rates, water cut, time of test and so on. In all a total of 12 such features are extracted to characterize each test. A representative plot of the raw data that may be acquired during a test is given in Figure 4. Many of the tests may not be acceptable due to a number of reasons such as a problem with the sensors, a problem with fluid separation, malfunctioning of equipment such as the pump or dumping valves and so on. An experienced operator can detect these problems from the measured data and reject the tests accordingly. Our objective is to reproduce this pattern recognition capability in the analytics module. The underlying assumption is that the extracted features capture the important characteristics of a test and allow us to distinguish between good and bad tests. Since all the data is automatically collected and it is straight-forward to extract the raw data from the database and extract features, we can generate a large dataset of "unlabeled" test data with relatively little effort. In this case, the operator labels about test acceptance or rejection are also available from their daily tasks, but the labels required for a more advanced diagnostic system (for example to not only reject a test but to also diagnose the root cause of the problem) are not, hence, the labels for the development of a diagnostic system would have to be provided by an expert user who would have to spend considerable time to go through the historical data and diagnose each test.

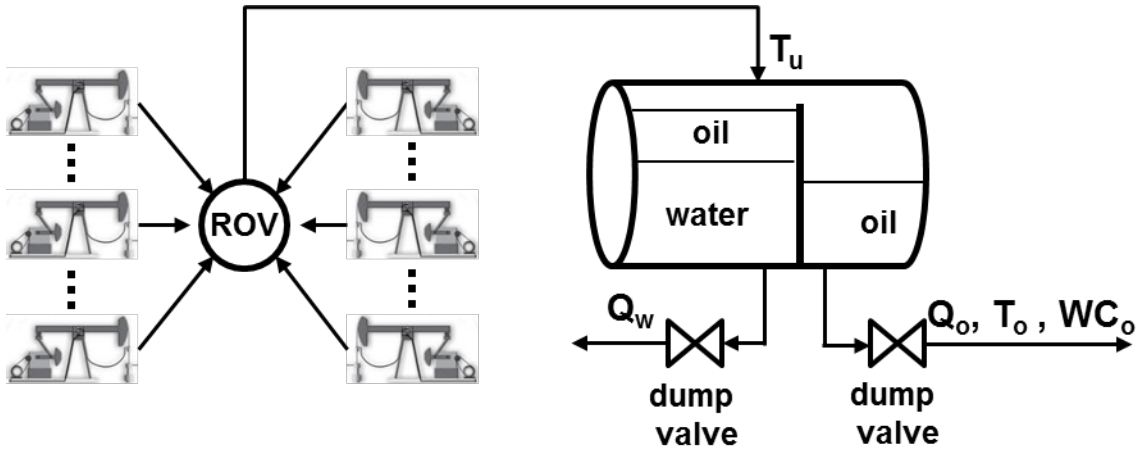


Figure 3. Physical setup and instrumentation of test separator. Each well of a given pad is connected to the well test separator through the remotely operated valve (ROV); only one well is connected at a time. Tu and To are inlet and oil outlet temperatures respectively. Qo and Qw are oil and water rates, respectively. WCo is water cut in separated oil.

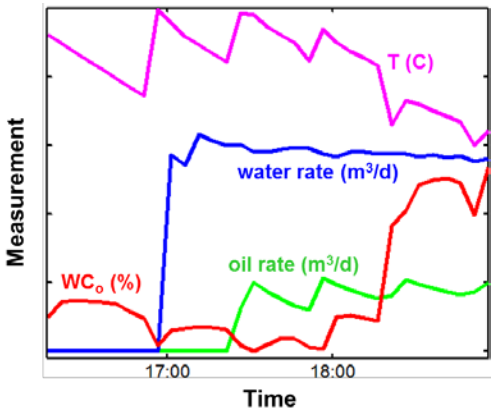
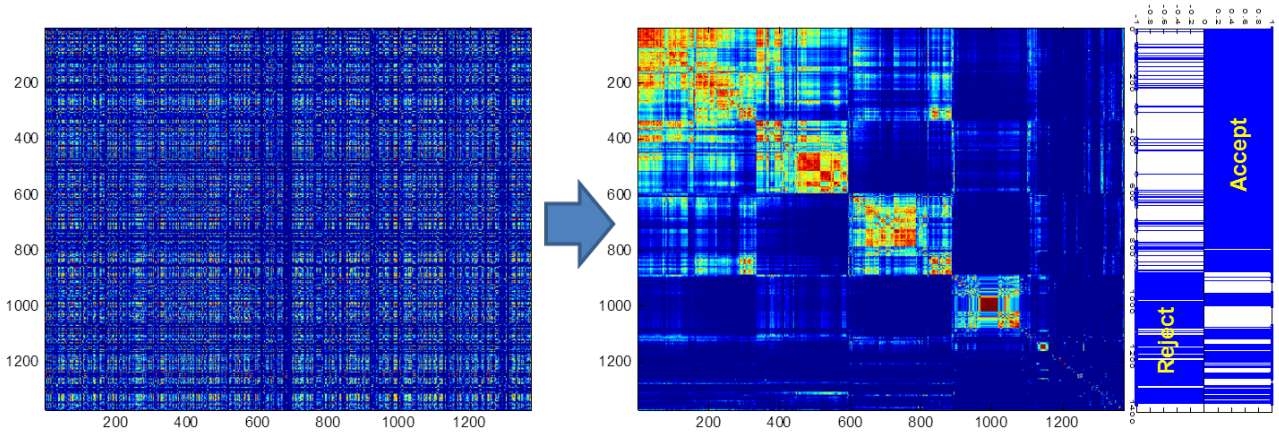


Figure 4. Representative data for a typical rate measurement test.

Technical Approach and Results

In our work, we first use a matrix rearrangement technique to visualize the structure in the data and to obtain an initial set of points for the expert to label. This is done in an interactive fashion where the data structure is displayed to the user in a format that makes it easy for the user to select the right points. Once the initial labeled set is obtained, semi-supervised learning is performed on the data to propagate the label information and obtain class probabilities for the unlabeled samples. These probabilities are then used to calculate the uncertainty of each unlabeled point, and the top few are selected for the expert to label in the next iteration. We compare the improvement in accuracy using this approach to the baseline, which is to select examples at random and label them. The techniques used in each step of the iterative process are detailed below.

The first step in our approach is to visualize structure in the data and select a good initial set for labeling. This is analogous to the simple example with clear clusters where labeling one point from each cluster would be sufficient. Unfortunately most real datasets do not display this clear structure, and a method that handles more complex data sets is required. Let N be the total number of data points available and M be the dimension of the data. In our example, we extract 12 features from each test ($M = 12$) and use historical data from a pad for a period of six months, which gives us 1370 data points ($N = 1370$). A good "similarity" matrix for the data set, $\mathbf{K}$, is computed using Gaussian radial basis functions as $\mathbf{K}(i,j) = \mathbf{K}(j,i) = \exp(-\alpha d(\boldsymbol{X}_i,\boldsymbol{X}_j))$, where α is a scaling parameter and $d(X_i,X_j)$ is the Euclidean distance between the i^{th} data point, $\boldsymbol{X}_i$, and the j^{th} data point, $\boldsymbol{X}_j$. $\mathbf{K}(i,j)$ has a maximum value of one when the distance between two points is zero and then drops off at a rate controlled by the scaling parameter α. In this paper, we use a hierarchical clustering method with optimal leaf ordering to rearrange the similarity matrix so that cluster-like structure becomes apparent. For example, if there are two clear clusters in the data, a well arranged similarity matrix would show two distinct diagonal blocks and selecting one point from each block would be easy. An extension of this concept is to create a tree structure whose nodes correspond to individual data points and then agglomerate the points which are most similar together to get a hierarchical cluster structure. Once this is done, the leaves of the tree may still be rearranged while honoring the tree structure to get an ordering of the leaves so that the sum of distances between the leaf nodes is minimized via dynamic programming (Bar-Joseph et al., 2001). The original similarity matrix (points ordered in chronological order) and the rearranged matrix are shown in Figure 5a and 5b, respectively. It is easy to see the structure in the data and the groups of similar data points in 5b, whereas 5a does not provide any such information. In order to get a good initial set for expert labeling, points are now chosen from the rearranged similarity matrix by making sure that all the blocks are sampled.



5a. Similarity matrix (data in chronological order) **5b. Rearranged similarity matrix (with expert labels)**

Figure 5. Comparison of similarity matrices before and after rearrangement to visual data structure. The similarity matrix is shown as a heat map (red indicates a high value of similarity and blue indicates a low value). The stem plot to the right of Figure 5b shows the known labels for each test (row).

As an aside, since we do actually have expert labels for all tests in this case, we go ahead and plot the class indicators (test accepted (+1) or rejected (-1)) alongside the rearranged similarity matrix in 5b, and it can be seen that just arranging the data points according to their similarity in an unsupervised fashion seems to separate the accepted tests from the rejected tests. We

also see that accepted tests tend to be more similar (although there are some distinct blocks), whereas rejected tests can be quite different from each other as there may be a number of reasons for their rejection. Although this application is attempting to classify the data into two classes (accept and reject), zooming in to these regions and using their structure may help generate labels for a more detailed diagnostics module.

The next step is to use information in the limited set of labels and the structure contained in all the unlabeled data together to improve the performance of the classifier. We use the method proposed by Zhu and Ghahramani (2002) that makes use of the similarity matrix to propagate labels using simple matrix operations. The basic idea is to convert the similarity matrix into a transition probability matrix **P**, such that $\mathbf{P}(i,j) = \mathbf{K}(i,j)/\sum_{l=1}^{N}\mathbf{K}(l,j)$, where **P**(*i*,*j*) gives the probability of going from point *i* to point *j*. The label information is represented as a matrix **L**, whose elements **L** (*i*,*c*) give the probability that the ith data point belongs to class *c*. For the data points for which labels are available, **L** (*i*,*c*) = 1 if the label for data point *i* is class c and 0 for other classes. The unlabeled data points are initialized with random probability assignments. The available labels are then propagated through this transition probability matrix by multiplying the initial label matrix repeatedly with the transition probability matrix as $\mathbf{L}_{new}=\mathbf{P}\,\mathbf{L}_{old}$. If two points are very similar, then there is a higher probability that they will share the same labels, otherwise label information from one point is not propagated easily to the other. Technical details about fixing the labeled data points, renormalizing each row of the label matrix before each iteration to ensure convergence of the algorithm, and other details may be found in Zhu and Ghahramani (2002). After convergence, the limited amount of label information is propagated to all unlabeled data points, and we have a probability distribution at each data point. The class with the highest probability is declared as the class to which the data point belongs.

Finally, class probability distribution for each data point may then be used to compute the entropy of the predicted distribution ($\mathbf{E}(i) = -\sum_c \mathbf{L}(i,c)log(\mathbf{L}(i,c))$). This entropy is a measure of the uncertainty in the prediction and the simple active learning algorithm we use selects the data points with the highest entropy for consideration by the expert for labeling.

To summarize, our workflow involves the following steps to learn a classifier with limited expert input.

1) Automatically rearrange the similarity matrix to emphasize cluster information and allow the expert to interactively select the first few data points for labeling
2) Use the available labels and unlabeled data to learn the labels for the entire data set using semi-supervised learning
3) Use the predicted probability distributions for class membership to identify the data points with highest uncertainty and present them to the user for additional label input
4) Iterate between steps 2 and 3 until the change in predicted labels is below a small threshold

The increase in accuracy of the learned classifier as more and more data points are added is shown in Figure 6. This is compared to a baseline approach where the points are selected randomly and a decision tree is used to learn from the labeled data only. The plot also shows the results of using semi-supervised learning only with randomly selected points and active learning/sampling with classical supervised learning. The results show that using either active learning to select points for labeling with classical supervised learning or using random sampling of labels with semi-supervised learning improves the performance of the learned classifier over the baseline over all levels of availability of labeled data. However, the most dramatic improvement is obtained when using both active learning for selecting the points to label and semi-supervised learning to propagate the learned labels to unlabeled data points.

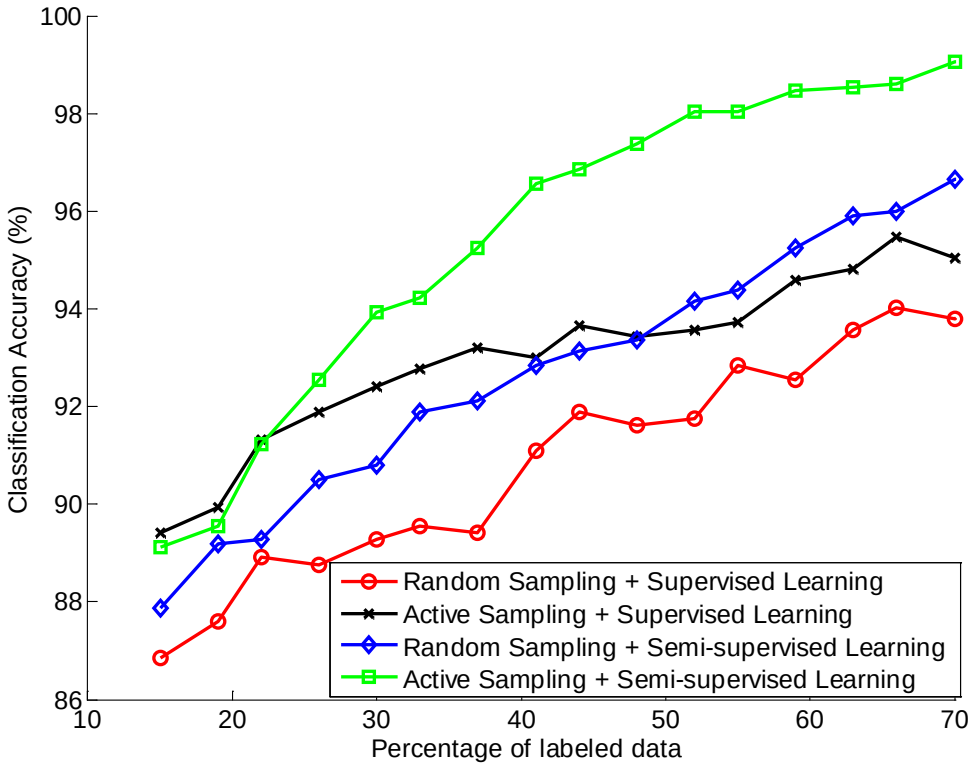


Figure 6. Comparison of classification accuracy (measured on the entire data set) with increasing labeled data using various approaches.

Conclusions

The rapid proliferation of sensing and data acquisition systems combined with cheap data storage is resulting in the age of "Big Data" and the most scarce resource going forward is likely to be expert time to go through, clean and label the data. This has led to the rising importance of adopting new strategies and workflows for the development of advanced analytics models. This trend is already visible in other industries and is likely to be a big challenge for the oil and gas industry as well. In this work, we explored the use of active learning to intelligently identify data points with the highest value of information and semi-supervised learning to combine the information from labeled and unlabeled sources in an optimal fashion. It was demonstrated that these advanced machine learning methods show considerable promise and can be integral tools in making the entire analysis procedure more interactive and productive. Effective adoption of these advanced machine learning techniques for interactive learning will be dependent on their integration with excellent data visualization and effective user interface design.

References

Z. Bar-Joseph et al., "Fast optimal leaf ordering for hierarchical clustering", In Proceedings of the Ninth International Conference on Intelligent Systems for Molecular Biology, 2001.

A. S. El-Bakry , SPE, et al., "Decision Support and Workflow Automation for the Development and Management of Hydrocarbon Assets Using Multi-Agent Systems", SPE 150285, presented at the SPE Intelligent Energy International Conference and Exhibition, Utrecht, The Netherlands, 27-29 March, 2012.

K.E. Killian et al., “Leveraging Technology and Support Infrastructure and Experience in Upstream Digital Technology Applications”, SPE 150220, presented at the SPE Intelligent Energy International Conference and Exhibition, Utrecht, The Netherlands, 27-29 March, 2012.

B. Settles, “Active Learning Literature Survey”, Computer Sciences Technical Report 1648, University of Wisconsin–Madison. 2009.

J. J. Shyeh et al., “Examples of Right-Time Decisions from High Frequency Data”, SPE 112150, presented at the SPE Intelligent Energy International Conference and Exhibition, Amsterdam, The Netherlands, 25-27 February, 2008.

X. Zhu and Z. Ghahramani, “Learning from labeled and unlabeled data with label propagation”, Technical Report CMU-CALD-02-107, Carnegie Mellon University, 2002.

X. Zhu and A. B. Goldberg, “Introduction to Semi-Supervised Learning”, Synthesis Lectures on Artificial Intelligence and Machine Learning, Morgan & Claypool, 2009.