Automated System for Drilling Operations Classification Using Statistical Features

Bilal Esmael
University of Leoben
Leoben 8700, Austria
Bilal@stud.unileoben.ac.at

Mohammad Arghad Arnaout
TDE GmbH
Leoben 8700, Austria
Arghad.Arnaout@tde.at

Rudolf K. Fruhwirth
TDE GmbH
Leoben 8700, Austria
Rudolf.Fruhwirth@tde.at

Gerhard Thonhauser
University of Leoben
Leoben 8700, Austria
Gerhard.Thonhauser@unileoben.ac.at

***Abstract*- operations classification is one of the most needed tasks in the oil & gas industry. It provides the engineers with detailed information about what is happening on the rig site.**

In this paper we propose an approach to classify drilling operations automatically using machine learning techniques. This approach takes as input the sensors data in a specific time range, and predicts the drilling operation.

Our approach is simple but effective, where for each sensor data (channel) a list of statistical features will be extracted, then features selection algorithms will be used to select the most informative features, and finally, a classifier will be trained based on these features.

In this paper many feature weighting and selection algorithms were tested to find which statistical measures clearly distinguish between many different rig operations. In addition, many classification techniques were employed to find the best one in terms of accuracy and speed.

Experimental evaluation with real data, from five different drilling scenarios, shows that our approach has ability to extract and select the best features and build accurate classifiers. The performance of the classifiers was evaluated by using the cross-validation method.

Index Terms – Operations classification, Statistical features, Features selection.

I. Introduction

Nowadays, it's very easy to monitor the basic drilling actions such as moving the drill string, rotating the drill string and circulating the drilling mud.

Many mechanical parameters, such as hook load and block position, are continuously measured during drilling oil wells. These parameters are measured by a group of sensors located around the drilling rig and wired to a measurement system called a mud-logging system. In addition, data transferring systems and data storing systems can be employed to transfer and store the sensors data anywhere in the world.

Although these systems are being developed rapidly, the techniques of data interpretation and analysis have not developed at the same speed, and there is a lack of systems able to make efficient use of all the data available to increase the efficiency of the drilling process.

Improving the drilling process relies on performance analysis that is primarily based on daily activity breakdowns [1].

Operations recognition systems break the total drilling time down into list of well-defined operations like drilling, rotating, make connection, etc.

Most operations recognition systems take as input the sensors data itself, and recognize the drilling operations. Paper [2] presents a drilling operations classification system using Support Vector Machine (SVM). The input of this system is five sensors values with a specific timestamp, and the output is one of six predefined operations.

Our approach is based on creating a compacted representation of the sensors data in a given time range, using a group of statistical features.

Many statistical features can be included in the representation. The authors of paper [3] used mean, variance, skewness, kurtosis and entropy as statistical features to classify audio signals.

The final goal of our research was not only to create a classifier for drilling operations, but also to find which statistical measures clearly distinguish between many different rig operations.

This research is a part of comprehensive research aiming at, not only classifying normal drilling operation, but also detecting drilling problems by adding a new group of features (textual features) extracted from daily morning reports.

The remainder of the paper is organized as follows: Section II presents the general framework of the proposed approach. Section III shows the details of statistical features extraction phase. Sections IV, V and VI introduce the details of features ranking and feature selection phases. Section VII shows the details of classification task, and the last section VIII shows the experimental results of the proposed approach.

II. The General Framework

In this section we will introduce the proposed approach that aims to recognize drilling operations using statistical features. Our operations recognition approach uses the classical steps of feature extraction, feature selection and classifier training, which are sketched in figure1 and further described below.

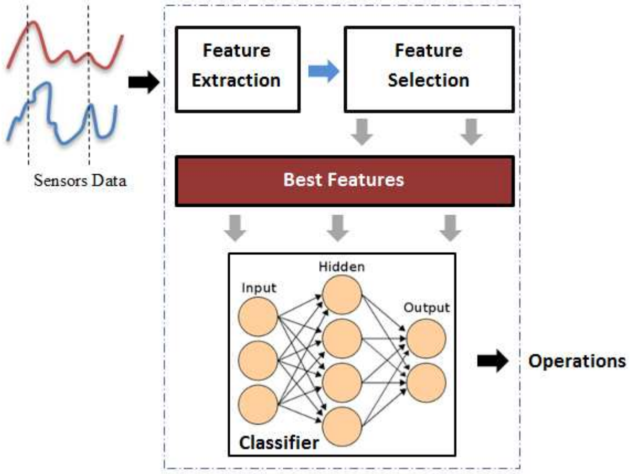

Figure 1. The general framework

III. Statistical Features Extraction

The first step of the approach is feature extraction, which is the transformation of patterns into features that are regarded as a compacted representation.
Many statistical measures were extracted to measure different properties of each channel as described in Figure2.

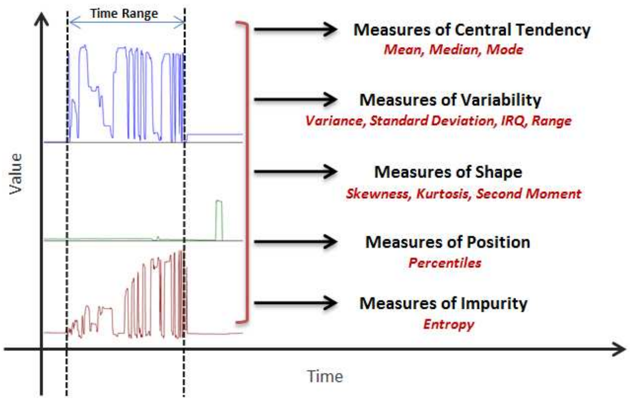

Figure 2. Statictical Features Extraction

In addition to the measures described in Figure 2, basic statistical functions were calculated like sum, min, max etc.
Skewness and kurtosis were used to measure the “asymmetry” and "peakedness" respectively, where second moment was used to measure the "width".
Percentiles were used to measure the “position”, where the p^{th} percentile is a value, Y_p, such that at most (100.p)% of the measurements are less than this value and at most 100(1- p)% are greater. Five percentiles values were included (p10, p25, p50, p75 and p90). Finally, entropy was used to measure the impurity.
Overall 22 statistical features were calculated for each channel namely: mean, median, mode, variance, standard deviation, interquartile range IRQ, range, skewness, kurtosis, second moment, p10, p25, p50, p75, p90, count, min, max, sum, first, last and entropy.

Data Channels
The aforementioned features were calculated for the 10 commonly used data channels described in Table1.
Some of these channels were normalized by dividing by the total depth of the selected well. In addition, one extra channel was generated which is the difference between mdbit and mdhole divided by the total depth.

Channels	Description
flowinav	Average mud flow-rate
hkldav	Average hook load
mdbit	Measured depth of the bit
mdhole	Measured depth of the hole
posblock	Block position
prespumpav	Average pump pressure
ropav	average rate of penetration
rpmav	Average drill string revolutions
tqav	Average torque
wobav	Average weight on bit

Table 1. Standard Data Channels

The total number of calculated features equals: Number of channels x Number of features = 11 x 22 = 242 features. All these features were calculated using simple software written in Matlab. This software takes as input a list of channels and a time range (start timestamp and end timestamp) and returns the mentioned statistical measures.

IV. Feature Ranking & Selection

High dimensional data, like our dataset, which has hundreds of features, can contain high degree of irrelevant and redundant information which may greatly degrade the performance of learning algorithms [4]. Therefore, feature selection becomes very necessary in our approach.

In the feature selection step, we seek to select a subset of relevant features with high predictive value. Feature selection was implemented to improve the performance of our learning models by increasing the accuracy of the classifiers and speeding up learning and classification processes. In addition, feature selection improved model interpretability because it is much easier to tell an engineer that from hundreds of features these 10 are important to the classification task than to explain the influence of the 242^{th} features.

In many applications, the best features can be selected using brute-force search, also known as exhaustive search. For a dataset with n features, exhaustive search needs (2^n-1) possibilities. In our case we have 242 features yielding 2^{242}-1 = 7.06×1072 possibilities to combine all the features. That means using exhaustive search is not feasible in finite time, and others selection algorithms should be considered.

Although many feature selections algorithms will remove the correlated features automatically, we preferred to start by removing these features in a separate initial step to drop the dimensionality of the data and increase the computational efficiency.
A correlation matrix (242×242) was calculated to check the correlation strength between features, then we searched for highly correlated ones and removed one of them.

V. Feature Ranking

The fastest way for feature selection, is ranking the features with some statistical test and selecting the k features with the highest score or those with a score greater than some threshold t. Such univariate filters do not take into account feature interaction, but they allow a first inspection of the data and most probably provide reasonable results [5].

We tested 10 different feature ranking algorithms (described in table2) and measured the performance of them.

Algorithm	Description
SAM	Calculates a weight according to "Significance Analysis for Microarrays"
PCA	Uses the factors of one of the principal components analysis as feature weights
SVM	Uses the coefficients of the normal vector of a linear support vector machine as feature weights
Chi Squared	Calculates the relevance of a feature by computing for each attribute the value of the chi-squared statistic with respect to the class attribute
Relief	Measures the relevance of features by sampling examples and comparing the value of the current feature for the nearest example of the same and of a different class
Gini Index	Calculates the relevance of the attributes based on the Gini impurity index
Information Gain	Calculates the relevance of the attributes based on the information gain
Correlation	Calculates the correlation of each attribute with the label attribute and returns the absolute or squared value as its weight.
Maximum Relevance	Selects Pearson correlation, mutual information or F-test depending on feature and label type (numerical/nominal).
Uncertainty	Calculates the relevance of an attribute by measuring the symmetrical uncertainty with respect to the class

Table 2. Feature Ranking Algorithms

Although the aforementioned algorithms did not produce identical results, there was about 70% of similarity between these results. For example most algorithms put flowin-p90, wobav-skewness, rpm-variance and prespumpav-range features in the top of the ranking list.

Feature number optimization

The resulting question now is: How many features should be used to get the best model in terms of accuracy? To answer this question, many tests were performed. We generated many models with different number of features and calculated the accuracy for each one. We started with the top 150 features and then reduced this number to 100, 50 and 25. Table 3 shows the results. For most algorithms, models trained with 50 features have the best accuracy.

Algorithm	Accuracy [%]			
	150 F	**100 F**	**50 F**	**25 F**
SAM	80.29	81.19	75.12	66.06
PCA	83.29	81.38	85.72	80.74
SVM	80.59	81.09	76.33	66.06
Chi Squared	82.31	82.41	83.19	79.68
Relief	81.29	82.2	83.19	78.57
Gini Index	80.89	80.69	81.59	80.08
Information Gain	81.6	81.19	81.88	80.39
Correlation	80.89	84.22	83.3	80.21
Maximum Relevance	80.69	82	79.31	79.58
Uncertainty	80.89	82.61	85.51	82.91

Table 3. Feature Ranking Comparison (150, 100, 50 and 25 Features)

To select the best number of features accurately we started with the top feature, and each time we added the next top feature until we finished all features. Figure 3 shows the accuracy curve as a function of the features number. It's clear that with 38 features we will get the most accurate classifier, but also with only 5 features we will get an acceptable result.

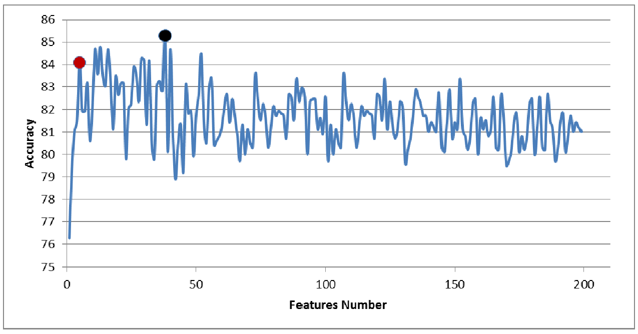

Figure 3. Accuracy curve as a function of the features numbers

VI. Forward Selection Methods

Forward selection method was used to bridge the gap between fast, but univariate filters, on the one hand, and slow, but multivariate exhaustive search, on the other hand.
Forward regression starts with creating models using exactly one feature. So we trained in the first step several networks using only the first feature as input, then the same procedure using the second feature as input and continued until the last feature was used as single model input [6]. The feature which yields the lowest error (ropav-p90) will be considered as the feature that has the most impact to the model.
In the second step we made new training runs with ropav-p90 as fixed input and adding exactly one of the remaining features as second input. We performed that procedure until all features were used as model input.
Many networks were trained to obtain as result the ranking of the input with respect to the model error. In Figure 4 the results are sketched, ropav-p90 has the leading impact followed by wobav-skewness, mdhole-p75, etc.
The first error values in Figure 4 give us the model errors using only ropav-p90 as input, the second values the errors using ropav-p90 & wobav-skewness as input, the third

values the errors using ropav-p90 & wobav-skewness & mdhole-p75, etc.

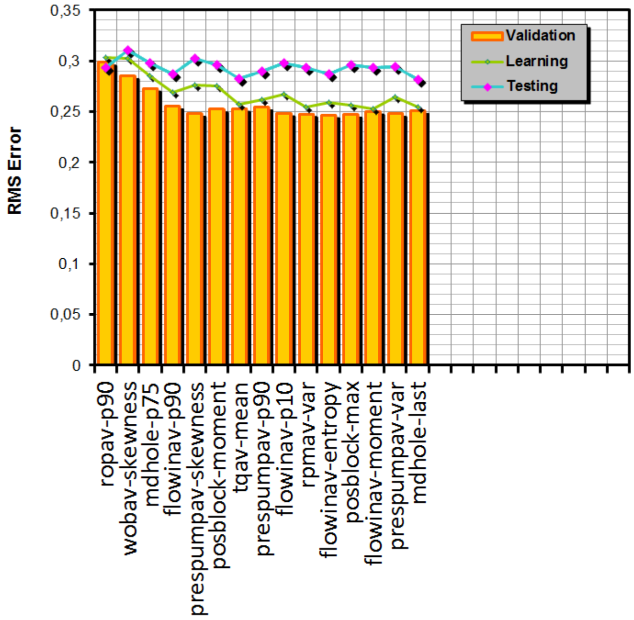

Figure 4. Forward Selection

VII. Classifiers Training

After extracting the features and selecting the most informative ones, we are ready to start classification process. Four classification techniques were constructed and employed in this study. These techniques are: Artificial Neural Network (ANN), Rule Induction (RI), Decision Tree (DT) and Naïve Bayes (NB).

Each one of these classifiers contains some parameters that can be tuned to improve the output of these classifiers. Many values and options of these parameters were tested to get the best results.

The performance of the classifiers was evaluated by using the cross-validation method. We found that the worst classifier –in most cases – is Naïve Bayes, and the best on is Rule Induction.

VIII. Experimental Results

To evaluate our approach, we collected data from four different drilling scenarios described in table4. We applied our approach to construct the features space, then we used RapidMiner[1] (an open-source system for data mining) to train the classifiers.

Scenario	Instances	Duration [day]	Depth [m]	Classes
#1	991	95	7825	5
#2	1250	190	4402	9
#3	770	87	4863	7
#4	470	41	4004	4

Table 4. Four drilling scenarios

[1] http://rapid-i.com/

At the beginning we discovered and pre-processed the data, then the classifiers were trained using the whole features set, Table 4 shows the results.

Scenarios	Accuracy [%]			
	ANN	**RI**	**NB**	**DT**
#1	78.2	79.08	65.02	72.55
#2	72.05	68.37	60.33	55.09
#3	78.12	78.90	63.95	75.26
#4	76.75	78.56	64.39	75.05

Table 4. Classification Results (all features)

After feature selection, we retrained the classifiers with only 38 features. Table 5 shows the results. The accuracy improvement rate is about 10%, and the classification and training process become much faster.

Scenarios	Accuracy [%]			
	ANN	**RI**	**NB**	**DT**
#1	82.51	85.45	67.78	76.75
#2	70.84	70.51	63.33	54.12
#3	80.52	86.41	66.74	79.34
#4	81.90	85.96	67.45	78.59

Table 5. Classification Results (38 features)

IX. Further Work

Our aim for future work is to extend this approach by adding a new group of features extracted from daily morning reports, and then using the whole features space to build a comprehensive classification and learning system that can be used not only for normal drilling operation recognition, but also for drilling problems detection and extracting the steps that have been taken to solve these problems.

Acknowledgment

We thank TDE Thonhauser Data Engineering GmbH for the support of this work and the permission to publish this paper.

References

[1] Thonhauser, G., "Using Real-Time Data for Automated Drilling Performance Analysis", OIL GAS European Magazine, Edition 4, 2004

[2] Adriane B. S. Serapiao, Rogerio M. Tavares, Jose Ricardo P. Mendes, Ivan R. Guilherme, "Classification of Petroleum Well Drilling Operations Using Support Vector Machine (SVM)", International Conference on Computational Inteligence for Modelling Control and Automation (CIMCA'06), 2006

[3] T. Lambrou, P. Kudumakis. "Classification of Audio Signals Using Statistical Features on Time and Wavelet Transform Domains", IEEE, 2002

[4] L. Yu, H. Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution", Twentieth international Conference on Machine Learning (ICML-2003), 2003

[5] Benjamin Schowe, "Feature Selection for high-dimensional data with RapidMiner", Technical University of Dortmund, 2010

[6] Rudolf K. Fruhwirth, G. Thonhauser, "Hybrid Simulation using Neural Networks to Predict Drilling Hydraulics in Real Time", SPE Annual Technical Conference and Exhibition held in San Antonio, Texas, U.S.A., 24–27 September 2006