BUILDING REUSABLE PREDICTION MODELS FOR TECHNOLOGICAL DATA USING BUSINESS INTELLIGENCE TOOLS

A.V. Vaytulevich, F.V. Stankevich
Scientific Supervisor: Docent, Dr. A.V. Kudinov
Tomsk Polytechnic University, Russia, Tomsk, Lenin str., 2, 634050
E-mail: vaanval@gmail.com

СОЗДАНИЕ ПОВТОРНО ИСПОЛЬЗУЕМЫХ МОДЕЛЕЙ ДЛЯ ПРОГНОЗИРОВАНИЯ ТЕХНОЛОГИЧЕСКИХ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ СРЕДСТВ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА

А.В. Вайтулевич, Ф.В. Станкевич
Научный руководитель: доцент, к.т.н. А.В. Кудинов
Томский политехнический университет, Россия, г. Томск, пр. Ленина, 2, 634050
E-mail: vaanval@gmail.com

Аннотация. *Большинство предприятий, занятых в сфере добычи нефти и газа, используют автоматизированные системы для мониторинга состояния оборудования. Процесс мониторинга может создавать большие объёмы данных. Поскольку эти данные потенциально могут содержать ценные с точки зрения улучшения технологического процесса знания, требуется их анализ. Ввиду сложности создания и поддержания моделей прогнозирования на основе таких данные, требуется использование методов для создания моделей, которые могут быть использованы повторно. В данной работе предлагается технология, представляющая собой последовательность шагов по созданию переиспользуемых моделей интеллектуального анализа данных нефтегазового оборудования.*

Introduction. Full-scale use of information systems in industry over the last 10 years yielded vast amounts of stored technological data. This allowed researchers to reconsider approaches to optimize key business processes (production, sales and resource management) based on retrospective analysis of said data. Due to its high volumes specialized tools should be used for such analysis, including various Business Intelligence (BI) approaches [1]. Introducing BI support into business processes can be impeded by need to use expert knowledge to build relevant models, as well as to verify analysis results. Presented research focuses on the methods and instruments aimed to apply data mining to analyze the data generated by oil wells. We will analyze such data and use it to predict various parameters of oil wells and to prevent upcoming equipment failures. In order to implement data mining solutions CRISP-DM will be used, since it incorporates iterative approach needed for countering possible design flaws [2].

Research area analysis. Many oil production companies equip most of their oil with automatic technological parameters registering systems which monitor oil production and related processes. Internal regulations of a company usually dictate a certain procedure used to collect data (i.e. how frequently measurements are made, how they are filtered and stored), which may generate dozens of values per second for a single object leading to a big data archive being stored.

The state of an oil well can be described by a set of parameters associated with equipment sensors and sensors measuring physical parameters of underground oil layers. Analyzing an oil well state allows estimating its future production, as well as predicting possible equipment failures. Predictions of similar parameters may vary depending on oil well type (production wells, exploration wells or injection wells) or on geological structures involved. An average oil well penetrates several oil layers with varying physical properties like pressure or temperature. However, these properties can be the same for different wells, given the same layer is penetrated. Therefore a single model based on similar physical properties can hypothetically be built which can then be reused for multiple oil wells. This will allow reducing the amount of models needed (compared to having a unique model for each oil well), cutting costs for their creation, storing and managing. In addition, some wells (which are mostly newly drilled) may have no historical data stored. By using a model associated with a similar well one can estimate possible short-term output using only current data from a new well.

Implementation. Oil wells with most data stored were determined by means of plain statistical analysis prioritizing the most recent data (less than 5 years old). A list of 45 oil wells of each of possible 3 types (exploration, injection and producing wells) was formed. Visual analysis of plotted data confirmed that some technological parameters of an oil well may have high correlation. Hence, the main goal for the next steps is to detect groups of such highly-correlating technological parameters to build models upon. A model would then consist of a set of parameters that correlate well for most oil wells.

A model using Microsoft Decision Tree Algorithm was trained for each well using the values of all parameters of the well. Since values for each parameter were not evenly distributed in time, the training set was split into equal time periods. The values of each parameter were then interpolated to match the starts of these periods, thus forming a uniform grid having timestamps as its lines and technological parameters as its columns. The modeling resulted in both correlation diagrams (Fig. 1) and its numeric representations [3]. The resulting correlations contained information on correlations within separate oil wells only. To build models applicable for multiple wells data had to be further analyzed using a different algorithm. To do this the data was first aggregated into a combined adjacency matrix for all correlation graphs acquired in the previous step with intersections representing correlation value aggregated from all models where a specific pair of parameters was present. Preliminary manual analysis of the resulting matrix, however, showed that some parameter pairs had high correlation values while being present only in minority of oil wells which had to be filtered out. To do this, a model allowing us to classify each parameter pair as being either “reliable” or “unreliable” had to be built (where “unreliable” would mean aforementioned rare occurrences). A total of 68 % of remaining data was removed using this. The filtered training set containing only reliable correlations could finally be used to form groups of parameters which were strongly linked in most training sets. To do this, filtered correlation matrix was used as input to another clustering model which split training data into said groups. As a result clusters of parameters were obtained and could be used to train corresponding universal prediction models (Fig. 2).

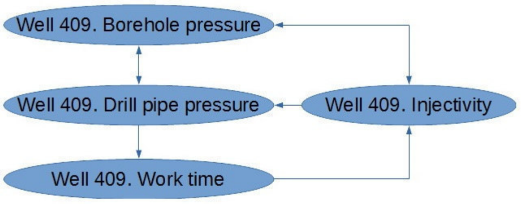

Fig 1. Example of parameters correlation graph

These models used Microsoft Time Series Algorithm for short-term prediction. Unaltered data from parameter groups was used to train these models, with a single model assigned to each group. Before deployment each prediction model’s accuracy was estimated using data from other wells.

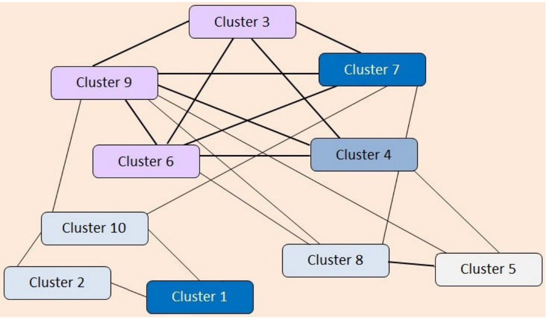

Fig. 2. Universal technological parameter clusters

Results. We have discussed and developed a technology that is able to identify groups of strongly correlating technological parameters which can be used to create prediction models suitable for most oil wells of a given oil field. This technology was used to determine parameter groups and build corresponding prediction models for a single oil field of an oil enterprise in Western Siberia. The quality of resulting models was evaluated against testing data in five cases: (1) single prediction, by using a model built for a specific oil well, (2) single prediction for one parameter type, by using a model built for a specific parameter type, (3) group prediction, by using a reusable group model, (4) group prediction without data, by using aforementioned models to predict data for newly drilled wells. Several existing approaches estimating mean absolute error (MAE) and mean absolute percentage errors (MAPE) to estimate model quality of were used [4]. As a result of the model testing, single prediction proved to be most precise method with 3.15 % error for 969 test cases for one-day prediction. This, however, cannot be used in real-world scenario due to high amount of technological parameters each oil well has (350 on average). Suggested approach of using universal reusable parameter group models showed 8.71 % error which was considerably lower than when using linear approximation (23.1 % error). Additionally, it could be used to predict technological parameter values even when no data was present (in case of new oil wells) with error slightly lower (22.47%) than using linear approximation for oil wells with historical data (23.1 %).

Conclusion and future research. The result of this work is an approach for building reusable prediction models of technological data of oil wells composed of five steps: (1) determining of oil wells that are used as training data sources; (2) determining dependencies in technological data within separate oil wells; (3) classification of these dependencies into “reliable” and “unreliable”; (4) clustering of dependencies to determine the groups of dependent technological data; (5) training reusable models based on these groups and estimating their quality.

The prototype of intelligence data analysis system of technological data of oil production was created as the result of this work. The prediction accuracy for reusable group models was 91.29 % using mean absolute percentage error for estimation (error being 8.71 %) for 969 data samples for one day prediction compared to later obtained data.

REFERENCES

1. Rud O. Business Intelligence Success Factors: Tools for Aligning Your Business in the Global Economy. – Hoboken: John Wiley & Sons, 2009. – 283 p.
2. Azevedo A., de Amorim J.L. KDD, SEMMA and CRISP-DM: A parallel overview. – Hoboken: John Wiley & Sons, 2004.
3. Resources and Tools for IT Professionals. Data Mining Algorithms [Electronic resource]. – Retrieved 01.04.2014 from http://technet.microsoft.com/en-us/library/ms175595.aspx.
4. Chuchueva I. The main estimates for the accuracy of time series prediction [Electronic resource] // Mathematical Bureau. – Retrieved 01.04.2014 from http://www.mbureau.ru/blog/osnovnye-ocenki-tochnosti-prognozirovaniya-vremennyh-ryadov.