УДК 004.89

К ВОПРОСУ О РАЗРАБОТКЕ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ ДЛЯ КЛАССИФИКАЦИИ СКВАЖИН-КАНДИДАТОВ НА ГЕОЛОГО-ТЕХНИЧЕСКИЕ МЕРОПРИЯТИЯ ПО ПРИЧИНАМ ОТКЛОНЕНИЯ

Хакимов Р.Ф.

ФГБОУ ВО «Уфимский государственный авиационный технический университет», Уфа, e-mail: wilmgc@gmail.com

Процесс согласования и отклонения скважин-кандидатов на геолого-технические мероприятия является трудоемким и подвержен человеческому фактору ввиду большого количества скважин-кандидатов и возможных причин отклонения, в связи с этим возникает задача разработки программного обеспечения для решения этой задачи. В данной статье рассматриваются результаты исследования возможности применения методов машинного обучения для классификации скважин-кандидатов на мероприятие «Вывод из бездействия» по причинам отклонения для повышения эффективности и надёжности процесса согласования и отклонения. В качестве классификатора предложен многослойный перцептрон. Для балансировки данных используются методы генерации синтетических экземпляров, в частности метод Syntenhic Minority Oversample Technique и метод Random Oversampling. Новизна представленного решения связана с использованием методов машинного обучения в процессе согласования и отклонения скважин после первичного подбора на геолого-техническое мероприятие. Использование нейросетевого классификатора для классификации скважин-кандидатов на мероприятие «Вывод из бездействия» по причинам отклонения позволяет с достаточной для повышения эффективности процесса точностью классифицировать заданную скважину. В результате оценки классификатора была получена средняя точность 96, 67, 72% по метрикам precision, accuracy, recall. Предложенное программное решение можно реализовать в виде API-службы, принимающей входные данные о скважине и возвращающей несколько наиболее вероятных причин отклонения. Такая служба сможет взаимодействовать с соответствующими прикладными приложениями для работы со скважинами для вывода на пользовательский интерфейс наиболее подходящих причин отклонения.

Ключевые слова: добыча нефти, геолого-технические мероприятия скважин, нормализация данных, балансировка данных, машинное обучение, методы классификации, верификация данных, нейронные сети

TO THE QUESTION OF SOFTWARE DEVELOPMENT FOR THE CLASSIFICATION OF WELL-CANDIDATES FOR GEOLOGICAL AND TECHNICAL ACTIONS BY REJECTION REASONS

Khakimov R.F.

Ufa State Aviation Technical University, Ufa, e-mail: wilmgc@gmail.com

Process of approval of wells-candidates for geological and technical actions is labour intensive and human-susceptible because of a large number of wells-candidates and possible rejection reasons. The purpose of the paper is to research possibility of using machine learning methods for classification of wells-candidates to geological and technical action called «Well recommissioning» by rejection reasons to improve efficiency and reliability of approval process. A multilayer perceptron is used as a classifier. Methods for generating synthetic instances are used to balance the data, in particular, the Synthetic Minority Oversample Technique method and the Random Oversampling method. Novelty of the presented solution is the use of machine learning methods in the process of approval and rejection wells-candidates after the initial selection for an action. The use of a neural network classifier for classification of wells-candidates for the action «Well recommissioning» by rejection reasons allows to classify a well with sufficient accuracy to increase efficiency of the process. As a result of the assessment of the classifier an average accuracy of 96%,67%,72% was obtained on precision, accuracy, recall metrics. Presented solution may be implemented in the form of an API service that accepts input data about a well and returns several of the most probable rejection reasons for a well. Such a service would be able to interact with appropriate applications to display the most appropriate rejection reasons on the user interface.

Keywords: oil production, geological and technical actions for wells, data normalization, data balancing, machine learning, classification methods, data verification, neural networks

При эксплуатации нефтяных скважин для увеличения производительности и повышения экономической эффективности скважины применяются геолого-технические мероприятия (ГТМ) – комплекс мер геологического, технологического, технического и экономического характера. Существует несколько видов ГТМ, таких как гидроразрыв пласта, обработка призабойной зоны, смена частоты ЭЦН, оптимизация, вывод из бездействия.

При подборе скважин-кандидатов на определенное мероприятие используются различные параметры скважины, такие как пластовое давление, забойное давление, текущие дебиты жидкости и нефти, вязкости и другие, а также статистические данные скважины.

Автоматизация первичного подбора скважин на геолого-технические мероприятия с помощью информационных систем позволяет значительно повысить эффективность и оперативность данного процесса [1].

После процесса первичного подбора скважин происходит процесс согласования ГТМ для каждой предложенной скважины – специалист либо согласовывает ГТМ либо отклоняет с указанием причины. Причины отклонения могут быть связаны с недостатком данных, аномальными данными либо организационными причинами. Список причин отклонения может отличаться для каждого вида ГТМ.

Процесс согласования и отклонения геолого-технических мероприятий для скважин является трудоемким ввиду большого количества скважин и причин отклонения. Автоматическая классификация скважин по причинам отклонения может существенно повысить оперативность работы специалистов.

В данной статье рассмотрен подход к разработке системы классификации скважин-кандидатов на геолого-технические мероприятия по причинам отклонения.

Постановка задачи

Имеющаяся в системе автоматизированного подбора скважин на геолого-технические мероприятия информация может прямо или косвенно указывать на невозможность проведения данного мероприятия. К примеру, показатель пластового давления может указывать на причины «Низкое пластовое давление» либо «Аномально высокое пластовое давление», информация о состоянии скважины может указывать на причину «Запущена в работу в текущем месяце», информация о целевых параметрах скважины может указывать на причину «Отсутствие эффекта от ГТМ».

Таким образом, наличие в системе данных о скважине даёт возможность составить предварительную классификацию либо ранжирование скважин по причинам отклонения. Наличие множества возможных причин отклонения для каждого отдельного вида геолого-технического мероприятия характеризует данную задачу как задачу многоклассовой классификации.

Необходимо разработать программное решение в виде API-службы, позволяющее обучить классификатор на заданной обучающей выборке, а также классифицировать заданную скважину-кандидата по причинам отклонения, соответствующим данному типу геолого-технического мероприятия.

Математическая постановка задачи классификации скважин-кандидатов на геолого-технические мероприятия может быть сформулирована следующим образом.

Дано: множество параметров скважин-кандидатов на геолого-технические мероприятие $X^{\text{ГТМ}} = \{x_1, x_2, \ldots, x_n\}$, множество причин отклонения для данного геолого-технического мероприятия $Y^{\text{ГТМ}} = \{y_1, y_2, \ldots, y_n\}$, где ГТМ $\in$ {ОПТ, ОПЗ, ГРП, УВЧ, ВБД, ВБД ПФ}, где ОПТ – оптимизация, ОПЗ – обработка призабойной зоны, ГРП – гидроразрыв пласта, УВЧ – изменение частоты ЭЦН, ВБД – вывод из бездействия, ВБД ПФ – вывод из бездействия прочего фонда – возможные геолого-технические мероприятия.

Разработать: алгоритмы $a^{\text{ГТМ}}: X^{\text{ГТМ}} \rightarrow Y^{\text{ГТМ}}$, способные классифицировать произвольную скважину $x^{\text{ГТМ}} \in X^{\text{ГТМ}}$.

Данная задача относится к классу задач многоклассовой классификации.

Подход к решению задачи многоклассовой классификации

Для решения задачи классификации скважин-кандидатов по причинам отклонения использован метод классификации с помощью нейронной сети. Обучающая выборка построена на основе данных из системы согласования скважин-кандидатов на геолого-технические мероприятия.

При построении обучающей выборки необходимо использовать только те скважины-кандидаты на геолого-технические мероприятия, признак согласования либо причина отклонения которых встречались в исходной выборке более 10 раз. Это ограничение связано с тем, что использование классов, для которых размер обучающей выборки меньше определённого числа не позволяет выполнить балансировку данных без потери обучающей способности нейронной сети.

При разработке классификатора были задействованы этапы, представленные на рис. 1.

Исходными параметрами алгоритма является вектор параметров для каждой i-й скважины $X^i = \{x_1^i, x_2^i, \ldots, x_n^i\}$, который включает в себя такие данные, как:

– Остановочные параметры скважины (дебит жидкости, нефти, забойное давление, пластовое давление, давление насыщения, линейное давление, обводненность).

– Потенциальные параметры скважины (дебит жидкости, нефти, обводненность, забойное давление).

– Параметры экономической эффективности скважины (Суммарные затраты на ГТМ, NPV, PI).

– Предыдущая причина отклонения (если возможно).

– Признак месторождения.

Рис. 1. Этапы разработки классификатора

На основе этих данных алгоритм классифицирует скважину в один из 11 классов Y_i (ГТМ принят, низкая экономическая эффективность, аварийный фонд, некорректные данные, и т.д.).

В связи с тем, что входные данные алгоритма имеют разные типы и области значений, данные в исходном виде могут оказывать разное влияние на обучение нейронной сети [2]. Для того чтобы избавиться от этого фактора, применяется нормализация данных. Среди входных параметров выделяются:

– Численные (Дебиты, давления, обводненность, экономические параметры).

– Категориальные (Признак месторождения, предыдущая причина отклонения).

Для нормализации численных параметров применяется метод минимаксной нормализации – $X' = \frac{X - X\min}{X\max - X\min}$ [3]. Для нормализации категориальных параметров применяется метод унитарного кода, при котором каждому из возможных значений признака сопоставляется отдельный бинарный признак [4].

В наборах данных часто оказывается так, что какие-либо классы присутствуют в большем количестве, чем другие. Такие наборы данных могут негативно влиять на обучение нейронной сети, так как более присутствующие классы будут оказывать большее влияние на обучение, чем менее присутствующие [5, 6]. Для решения проблемы несбалансированности данных в обучающей выборке используется алгоритм SMOTE (Synthetic Minority Oversample Technique), суть которого заключается в генерации искусственных экземпляров миноритарного класса [7]. Искусственные экземпляры генерируются в «соседних» областях с помощью алгоритма ближайшего соседа (KNN).

При разбиении исходных данных на обучающую и тестовую выборки в задаче многоклассовой классификации важно, чтобы каждый класс был представлен равно как в обучающей, так и в тестовой выборке, иначе классификатор может иметь недостаточно данных для обучения либо для проверки. Для равномерного распределения классов среди обучающей и тестовой выборок используется алгоритм стратифицированного разделения.

Таким образом, исходное множество X делится на N подмножеств $X_k \subseteq X, k \le N$, где N – количество классов, в соответствии с принадлежностью к классу k. После этого из каждого подмножества случайным образом выбирается $T_k = |X_k| * h$ элементов, где h – коэффициент разделения, для тестовой выборки – $X^{test} = \bigcup_i^N T_i$, а из оставшихся элементов формируется обучающая выборка – $X^{train} = X \setminus X^{test}$.

Для решения поставленной задачи используется архитектура многослойного перцептрона, со следующими слоями (рис. 2).

– Входной слой, 36 элементов.

– Скрытый слой, 128 элементов.

– Скрытый слой, 128 элементов.

– Выходной слой, 11 элементов.

В скрытых слоях используется функция активации линейного выпрямителя (ReLU) – $f(x) = \begin{cases} 0, x < 0 \\ x, x > 0 \end{cases}$. Для выходного слоя используется функция активации Softmax – $f_i(\vec{x}) = \frac{e^{x_i}}{\sum_{j=1}^{J} e^{x_j}}$. Областью определения функции Softmax является (0,1), и её результат можно интерпретировать как вероятность попадания в заданный класс.

Для того чтобы предотвратить переобучение нейронной сети, используется метод регуляризации Dropout, суть которого заключается в отключении определённого количества случайных нейронов слоя на каждом шаге обучения. В данном случае метод Dropout применяется к скрытым слоям нейронной сети.

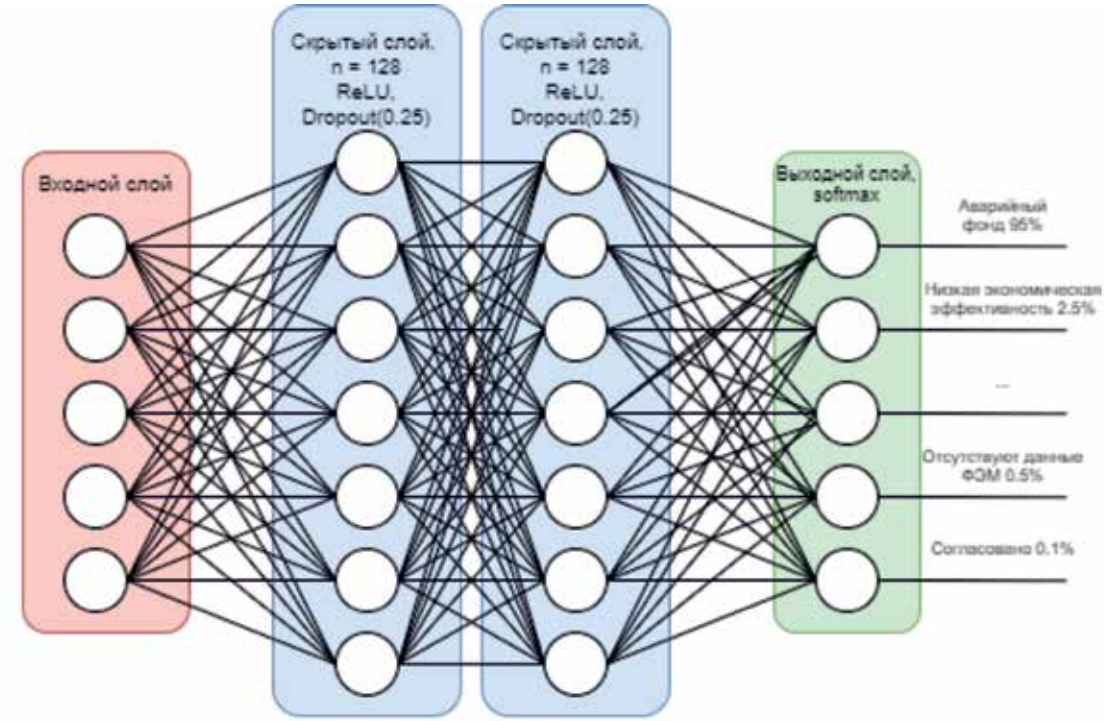

Рис. 2. Архитектура нейронной сети

Для реализации заданной модели использовался язык программирования Python, а также пакет keras для реализации нейросетевой модели и пакет sklearn для обработки исходных данных и валидации результатов. API-служба реализована на языке Python с помощью фреймворка Flask.

Оценка результатов

Для проведения эксперимента был выбран метод скользящего контроля со случайными разбиениями. В данном методе исходная выборка X^L делится N различными способами на две непересекающиеся выборки $X^L = X_n^m \bigcup X_n^k$, где X_n^m – обучающая выборка, а X_n^k – тестовая выборка. Алгоритм классификации обозначим $a_n = \mu(X_n^m)$, значение оценки качества $Q_n = Q(a_n; X_n^k)$. После вычисления среднего арифметического значения оценок по всем выборкам получим оценку скользящего контроля: $CV(\mu, X^L) = \frac{1}{N}\sum_{n=1}^{N} Q(\mu(X_n^m); X_n^k)$. В данном эксперименте исходная выборка делится на 10 различных случайных разбиений на обучающую и тестовую выборку.

При оценке точности классификатора в задачах многоклассовой классификации принято использовать метрики Accuracy, Precision и Recall. Для оценки точности классификатора можно представить многоклассовый классификатор как множество бинарных классификаторов для каждого класса. Таким образом, введем понятия для результатов классификации: истинно положительный результат, ложно положительный результат, истинно отрицательный результат, ложно отрицательный результат. В задачах оценки классификатора такие результаты называются True Positive (TP), False Positive (FP), True Negative (TN), False Negative (FN). Метрика $accuracy = \frac{TP + TN}{TP + TN + FP + FN}$ показывает долю верно предсказанных результатов, метрику $precision = \frac{TP}{TP + FP}$ можно интерпретировать как долю объектов, названных классификатором положительными и при этом действительно являющимися положительными, а метрика $recall = \frac{TP}{TP + FN}$ показывает, какую долю объектов положительного класса из всех экземпляров положительного класса определил классификатор. На следующей диаграмме показаны усреднённые по всем разбиениям оценки accuracy, precision, recall для геолого-технического мероприятия «Вывод из бездействия».

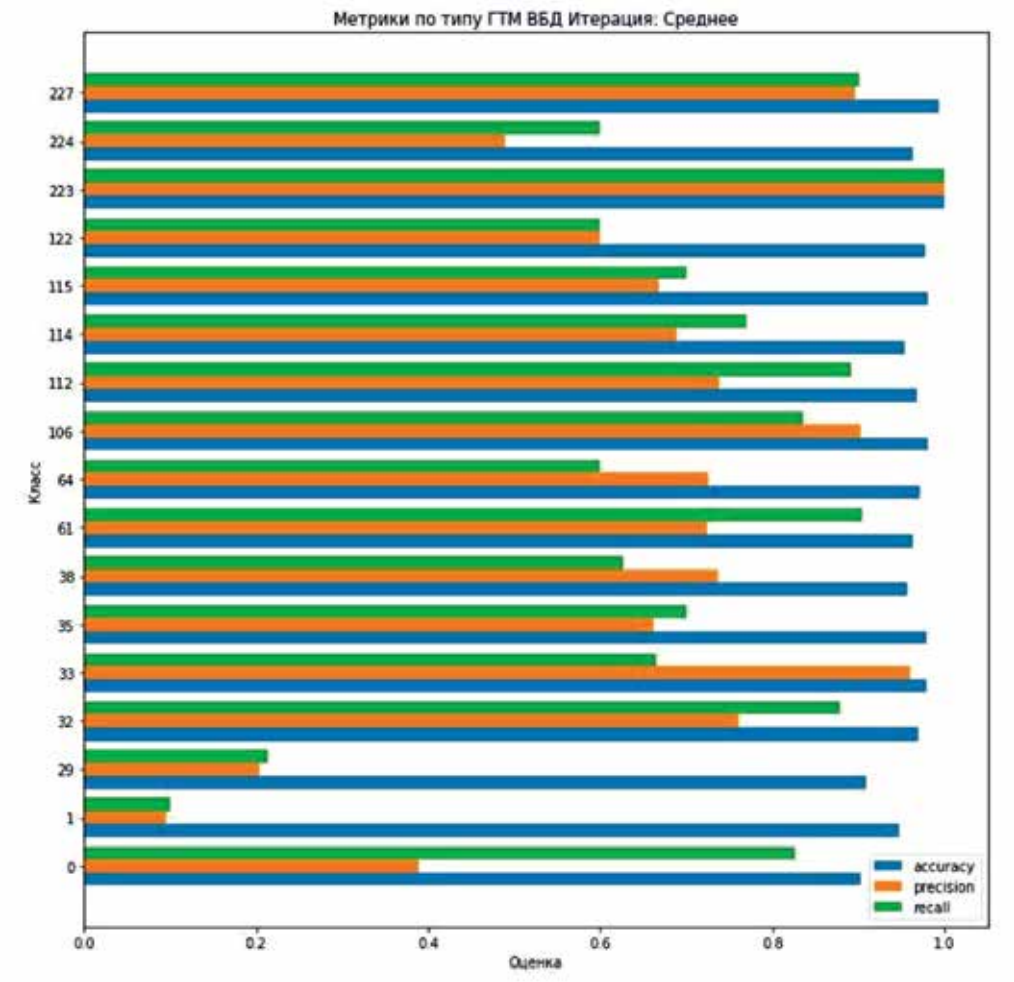

Рис. 3. Результаты оценки классификатора

Усреднив оценки по всем классам, получим итоговую оценку классификации по геолого-техническому мероприятию «Вывод из бездействия» accuracy = 96 %, precision = 67 %, recall = 72 %.

Для большинства результирующих признаков точность классификации является довольно высокой. Самая низкая точность классификации получилась для класса 1 (Стоит бригада), accuracy = 0.94, precision = 0.10, recall = 0.11. Для повышения точности классификации по этой причине отклонения необходима информация о бригадах, которой в данном исследовании не было в исходной выборке.

Заключение

Процесс согласования и отклонения скважин-кандидатов на геолого-технические мероприятия является трудоёмким и ресурсозатратным, автоматизация некоторых аспектов данного процесса может повысить его эффективность и надёжность.

В ходе исследования реализовано программное решение для классификации скважин-кандидатов на геолого-техническое мероприятие «Вывод из бездействия» по причинам отклонения с помощью нейросетевого классификатора. Результаты эксперимента показали, что нейронная сеть с заданными параметрами решает задачу классификации скважин с точностью, достаточной для отображения оператору в качестве предложения для принятия решения. Увеличение количества входных параметров о скважине, таких как информация о предыдущих мероприятиях, информация о мероприятиях на данном пласте, может существенно увеличить точность классификатора.

Список литературы

1. Ситников А.Н., Асмандияров Р.Н., Пустовских А.А., Шеремеев А.Ю., Зулькарниев Р.З., Колупаев Д.Ю., Чебыкин Н.В., Кириллов А.А. Формирование программ геолого-технических мероприятий с помощью цифровой ин-

формационной системы «Подбор ГТМ» // PROНЕФТЬ. Профессионально о нефти. 2017. № 2 (4). С. 39–46.

2. Sola J. & Sevilla Joaquin. Importance of input data normalization for the application of neural networks to complex industrial problems. Nuclear Science, IEEE Transactions. 1997. Vol. 44. No 18. P. 1464–1468.

3. Patro S., Sahu K.K. Normalization: A preprocessing stage // arXiv preprint. 2015. [Electronic resource]. URL: https://arxiv.org/ftp/arxiv/papers/1503/1503.06462.pdf (date of access: 16.05.2021).

4. Potdar, Kedar & Pardawala, Taher & Pai, Chinmay. A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers. International Journal of Computer Applications. 2017. Vol. 175. No. 4. P. 7–9.

5. Shuo Wang, Member, and Xin Yao. Multiclass Imbalance Problems: Analysis and Potential Solutions. IEEE Transactions On Systems, Man, And Cybernetics. Part B: Cybernetics. 2012. Vol. 42. No. 4. P. 1119–1130.

6. Никулин В.Н., Канищев И.С., Багаев И.В. Методы балансировки и нормализации данных для улучшения качества классификации // Компьютерные инструменты в образовании. 2016. № 3. С. 16–24.

7. Chawla N.V., Bowyer K.W., Hall L.O., Kegelmeyer W.P. SMOTE: Synthetic Minority Over-Sampling Technique. Journal of Artificial Intelligence Research. 2002. Vol. 16. P. 321–357.