Параллельный генетический алгоритм отбора значимых признаков в задаче формирования мероприятий для нефтяных скважин

В. В. Мокшин[1], И. М. Якимов[2]
Казанский национальный исследовательский технический университет им. А.Н. Туполева-КАИ
[1]vladimir.mokshin@mail.ru, [2]yakimovigormaks@mail.ru

П. И. Тутубалин[3]
Казанский национальный исследовательский технический университет им. А.Н. Туполева-КАИ
[3]ptyt@ya.ru

В. А. Суздальцев[4], И. А. Зарайский[5], Э. Г. Тахавова[6]
Казанский национальный исследовательский технический университет им. А.Н. Туполева-КАИ
[4]svlant@mail.ru, [5]ringo123@rambler.ru, [6]elzzy@yandex.ru

***Аннотация*. В работе предложен метод формирования мероприятий для нефтяных скважин. Предлагаемый метод основан на применении параллельного генетического алгоритма. Значимые признаки отбираются с использованием параллельного генетического алгоритма. Эффективность предлагаемого подхода демонстрируется на основе анализа данных функционирования нефтяных скважин.**

Ключевые слова: сложные системы; параллельный генетический алгоритм; отбор информационно значимых факторов

I. Введение

На сегодняшний день, в любой организации генерируется и накапливается большое количество информации. В данной работе рассматривается нефтегазодобывающая отрасль, в которой имеется существенный парк нефтяных скважин. Возникает задача планирования обслуживания и эксплуатации существующих объектов. При грамотном формировании мероприятий по скважинам позволит увеличить эффективность работы предприятия в целом. Для формирования того или иного решения по скважине используется телеметрическая информация. Поскольку скважин много, то формируется большой объем телеметрической информации. Для облегчения работы специалистов предлагается на основе собираемой информации по скважинам разработать модель, которая позволит формировать мероприятия по скважине на основе накопленной информации. Именно для организации более точной и безопасной работы оборудования, а следовательно, повышения эффективности процесса добычи полезных ископаемых широко используются автоматизированные информационные системы.

Организация сложных промышленных процессов базируется на принципе «черного ящика», когда поведение системы характеризуется набором признаков (входов) и соответствующих им ответных реакций (выходов). Функциональная зависимость состояний выходов от состояний входов лежит в основе задача моделировании подобных систем [1, 2] и представляется в виде целевой функции. Этот же принцип используется для организации сложных вычислительных процессов, требующих обработки большого объема данных и выделения среди полученной информации определенной структуры [3].

Признаки (feature), используемые в моделировании, оказывают большое влияние на качество результатов. Неинформативные или слабо информативные признаки могут существенно понизить эффективность модели. Для решения данной проблемы и для повышения результативности анализа данных большой эффект дает отбор значимых для модели входных признаков. Проблемы отбора признаков рассмотрены в работах [4]. Метод прямого включения признаков, метод обратного исключения, корреляционный метод, генетический алгоритм отбора значимых признаков [5–7] отбирают разное количество признаков.

Отбор признаков – это процесс выбора признаков, имеющих наиболее тесные взаимосвязи с целевой переменной. Одним из важных целей при создании алгоритмов отбора значимых признаков – это создание достаточно высокой точности работы, т.к. это напрямую влияет на возникновение ошибок работы системы, что может привести к потерям или иным серьезным проблемам, что недопустимо.

Исправить сложившееся положение можно с помощью автоматизированной информационной системы (АИС). Информационная система – это взаимосвязанная совокупность средств, методов и персонала, используемых

для хранения, обработки и выдачи информации для достижения цели управления. АИС представляет собой совокупность компонентов, объединённых регулируемыми взаимоотношениями для формирования организации как единого целого и обеспечения её целенаправленной деятельности. И как следствие этого определения, эффективность информационной системы может быть оценена только оценкой её вкладов в достижение организацией её стратегических целей, определить эффективность работы скважины и выдаче мероприятий для дальнейшей эксплуатации скважины. Для анализа проводятся статистические исследования и выявляются зависимости результативных показателей эффективности от производственно-технических факторов, что позволяет наилучшим образом оптимизировать работу предприятия.

Решение этих задач возможно на базе статистических методов исследования, которые являются достаточно развитой областью математики, и в то же время их применение для конкретной предметной области требует решения достаточно сложных вопросов, как в математическом, так и в практическом плане [8–10].

II. Планирование мероприятий

Оператор добычи нефти и оператор технической группы, на основании инструкций, рекомендаций, приказа и стандартов нефтедобывающего предприятия снимают производственно-технические показатели буровых скважин для мониторинга скважин. Результативность мониторинга скважин проверяется с помощью анализа состояния буровых скважин, по средствам которого составляется временной прогноз, отчет мероприятий по эксплуатации скважин и оценка степени влияния факторов на показатели и формируются мероприятия по скважинам. В работе предлагается проведение отбора значимых признаков с использованием параллельного генетического алгоритма. А далее на основе отобранных факторов происходит формирование модели для формирования мероприятий по скважинам. Формирования модели проводится на основе регрессионного анализа либо нейронной сети. В работе приводится сравнительный анализ рассматриваемых моделей [11].

III. Параллельный генетический алгоритм

Математически исследуемая система представляется в следующем виде:

$$\{\mathbf{Y}\} = \Phi\left[\{\mathbf{X}\},\{\mathbf{Z}\}\right] \quad (1)$$

где Φ – оператор системы, определяющий связь между указанными величинами;

$\{\mathbf{Z}\} = (z_1, z_2, ..., z_L)$ – множество векторов независимых признаков, которые влияют на систему, но не зависимые от самой системы;

$\{\mathbf{Y}\} = (y_1, y_2, ..., y_K)$ – множество векторов выходных показателей системы;

$\{\mathbf{X}\} = (x_1, x_2, ..., x_M)$ – множество векторов входных признаков (факторов), управляемые системой.

Суть алгоритма заключается в том, что для отбора значимых факторов запускаются короткие параллельные генетические алгоритмы [11]. Их количество $b = \overline{1,B}$. Число поколений в каждом коротком генетическом алгоритме N и размер популяции m. В каждом эволюционном пути $b = \overline{1,B}$ вычисляется частота появления входного признака x_i, $i = \overline{1,M}$. В параллельном генетическом алгоритме отбора значимых признаков длина эволюционного пути определяется энтропией популяции. Как только энтропия принимает устойчивую величину, отбор значимых признаков прекращается. Эксперименты показали, что для этого достаточно около 8-ми поколений.

В результате опишем параллельный алгоритм отбора значимых признаков с числом параллельных эволюционных путей B.

Входные данные параллельного генетического алгоритма:

B – число параллельных эволюционных путей;

n – размер временных отрезков наблюдений признаков;

m – размер популяции;

N – число поколений каждого эволюционного пути;

v_t – вероятность мутации особи поколения t (v_t = 1/M);

X – матрица входных признаков x_i, $i = \overline{1,M}$ размером $n \times M$;

Y – матрица размером $n \times 1$, временные наблюдения одного из $j = \overline{1,K}$ результативных показателей.

Шаг 1. Задание количества параллельных эволюционных путей B.

Шаг 2. Выполняются шаги 3, 4 для каждого эволюционного пути $b = \overline{1,B}$ в параллельном генетическом алгоритме.

Шаг 3. Примем $P(\mathrm{b}, \mathrm{N})$ как b-е поколение популяции на N-м эволюционном пути. Это результат генетического алгоритма с заданными параметрами генетического алгоритма отбора факторов $(\mathrm{Y}, \mathrm{X}, \mathrm{m}, \mathrm{N}, \mathrm{v_t})$.

Шаг 4. Вычисляется вес $r(\mathrm{i}, \mathrm{b})$ каждого входного признака x_i, $i = \overline{1,M}$, на b-м эволюционном пути. Параметр $r(\mathrm{i}, \mathrm{b})$ характеризует частоту появления входного признака,

$$r(j,b) = \frac{1}{m}\sum_{i=1}^{m} \omega_i(j), \quad (2)$$

где $\omega_1, \omega_2, ..., \omega_m \in P(b, N)$;

Шаг 5. Определим $\overline{r_i}$ – частоту появления входного признака i относительно всех эволюционных путей

$b=\overline{1,B}$ параллельного генетического алгоритма, определяемое формулой

$$\bar{r}_j = \frac{1}{B}\sum_{b=1}^{B} r(j,b). \quad (3)$$

Шаг 6. Полученные значения частоты появления признаков в матрице $\bar{r}_i$ сортируются в порядке убывания и определяется максимальное расстояния d между частотами $\bar{r}_i$ и $\bar{r}_{i+1}$, $i=\overline{1,M-1}$.

Шаг 7. Для кластеризации набора значимых признаков между упорядоченными в порядке убывания частотами появления признаков $\bar{r}_i$ и $\bar{r}_{i+1}$, $i=\overline{1,M-1}$ вычисляется максимальное расстояние $d\max$. Оно характеризует максимально допустимое расстояние для отбора признаков,

$$d_{\max} = \alpha\sqrt{\frac{1}{4B}}, \quad (4)$$

где $\alpha = 1,645$

Шаг 8. При выполнении условия $d \ge d\max$ выполняется шаг 9, иначе все входные признаки признаки будут являться значимыми.

Шаг 9. Среди отсортированных весов признаков матрицы $\bar{r}_i$ для дальнейшего исследования выбираются те признаки, веса которых находятся выше максимального расстояния d. Отобранные признаки включаются в массив $\overline{R(i)}$, где $i=\overline{1,M}$.

Выходными данными параллельного генетического алгоритма является массив отобранных признаков $\overline{R(i)}$ с числом параллельных эволюционных путей B:

Важной задачей является также определение наилучшего количества эволюционных путей в параллельном генетическом алгоритме (ПГА). Для решения этой задачи ПГА запускается U раз с увеличением $b=\overline{1,B}$. В итоге будет сформирован куб данных, где значением ячеек куба будет частота появления признаков x_i, $i=\overline{1,M}$ с учетом различного количества эволюционных путей $b=\overline{1,B}$ и с количеством повторных экспериментов U запуска ПГА. Далее для определения наилучшего количества эволюционных путей анализируется полученный куб данных. Определяется дисперсия весов входных признаков $\Delta r = r_{\max} - r_{\min}$ (рис. 1).

В случае, когда совокупность отобранных признаков становится стабильной, увеличение числа параллельных эволюционных путей прекращается (рис. 1).

Параметры ПГА:

Размер популяции, m – 50;

Количество поколений, N – 8;

Уровень мутации поколения t, v – 0.05;

Количество параллельных генетических алгоритмов, B –50.

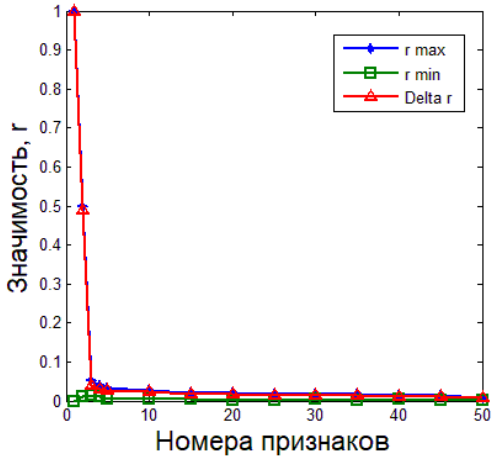

Рис. 1. Изменение значимости признаков с учетом повторных запусков ПГА для входных признаков x_i, $i=\overline{1,M}$

Результаты работы ПГА приведены на рис. 2.

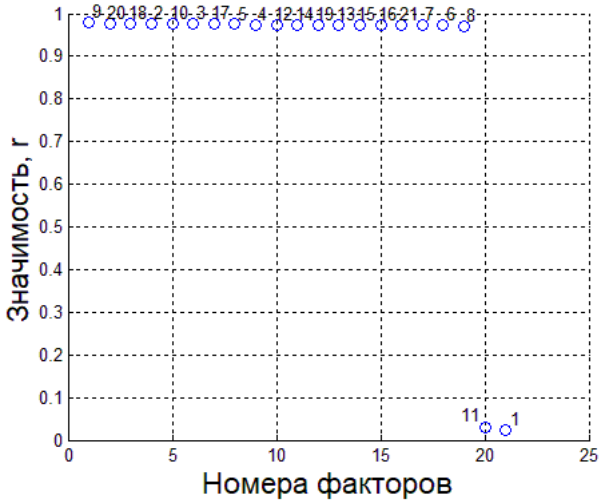

Рис. 2. Отсортированные входные переменные в соответствие с уровнем значимости для входных признаков x_i, $i=\overline{1,M}$

Из рис. 2 следует, что значимыми признаками были отобраны все, кроме 11-го и 1-го. То есть на основе отобранных признаком строится регрессионная модель [11] и нейронная сеть для поддержки принятия решений.

IV. Результаты

На основе отобранных признаков на примере y_1 построены регрессионные модели и сформирована нейронная сеть. Результаты генерации мероприятий для нефтяных скважин и оценка моделей по нейронной сети приведены в табл. 1.

ТАБЛИЦА I РЕЗУЛЬТАТ ГЕНЕРАЦИИ МЕРОПРИЯТИЙ

№	Y1	Y1 прогноз	Декодирование Y1 прогноз	Декодирование Y1 исходной	Ошибка
1	-0,36	-0,40	6	6	0,048
2	-1,15	-1,1	1	1	-0,048
3	-1,15	-1,1	1	1	-0,049
4	-0,36	-0,39	6	6	0,039
5	-0,51	-0,56	5	5	0,043
6	-0,36	-0,31	6	6	-0,044
7	-1,15	-1,15	1	1	0,001
8	1,87	1,89	20	20	-0,028
9	0,6	0,59	12	12	0,012
10	-0,51	-0,52	5	5	0,004

В данной таблице представлены исходные данные y_1, прогноз y_1, декодирование y_1 прогноз и декодирование y_1.

Нейронная сеть была обучена на 200 выборках и на 100 тестирующих выборках. Из них было выявлено, что 19 из 200 ошибочные, отсюда следует вероятность ошибки 0,095. А на тестирующей выборке ошибочные – 0 из 100.

ТАБЛИЦА II РЕЗУЛЬТАТ ГЕНЕРАЦИИ МЕРОПРИЯТИЙ ПО РЕГРЕССИОННОМУ АНАЛИЗУ ДЛЯ 10-ТИ СКВАЖИН

№	Y1	Y1 прогноз	Декодирование Y1 прогноз	Декодирование Y1 исходной
1	-1,160	0,755	1	1
2	-1,160	0,647	1	1
3	-1,160	0,794	2	1
4	-0,280	0,728	7	6
5	-0,456	0,863	3	5
6	-0,280	0,578	6	6
7	-1,160	0,767	1	1
8	-0,280	0,818	8	6
9	-0,280	0,401	8	6
10	-0,456	0,875	5	5

В табл. 2 представлены исходные данные y_1, результаты y_1 по регрессионному анализу и декодирование данных y_1. По этим данным можно сделать вывод, что результат генерации мероприятия по регрессионному анализу первой скважины равен 0.755, после декодирования y_1 по результату регрессионного анализа по (5) получилось число 1.

$$x_{si} = \frac{x_i - \overline{x}_i}{\sigma_i^2} \quad (5)$$

где x_{si} – стандартизованное значение i-го фактора, x_i – фактическое значение i-го фактора, $\overline{x}_i$ – среднее значение i-го фактора, σ_i^2 – среднеквадратическое отклонение i-го фактора.

Это соответствует мероприятию 1 по таблице закодированных ВНР (водонефтяной раздел) – в скважину заливают реагент 0,5 л. Одной этой капли достаточно, чтобы отделить нефть от воды. (1мг на 1 т. нефти).

По результатам регрессионного анализа следует, что не все данные совпали и это привело к выделенным ошибочным данным. Нейронная сеть, построенная с учетом отобранных входных признаков с помощью параллельного генетического алгоритма, дала более качественные результаты по сравнению с регрессионной моделью, и предлагаемый подход облегчит работу лица принимающего решения.

СПИСОК ЛИТЕРАТУРЫ

[1] I. Yakimov, Alexander Kirpichnikov, Vladimir Mokshin, Zuhra Yakhina. Rustem Gainullin. The comparison of structured modeling and simulation modeling of queueing systems. Communications in Computer and Information Science (CCIS), 800, 256-267 (2017).

[2] Pavel Innokentievich Tutubalin, Vladimir Vasilevich Mokshin. The Evaluation of the cryptographic strength of asymmetric encryption algorithms. 2017 Second Russia and Pacific Conference on Computer Technology and Applications (RPC) (Vladivostok, Russia, September 25-29, 2017). IEEE. Vladivostok, 2017. P. 180-183.

[3] Suzdaltsev V.A., Suzdaltsev I.V., Bogula N.Yu. Fuzzy rules formation for the construction of the predictive diagnostics expert system. Proceedings of 2017 20th IEEE International Conference on Soft Computing and Measurements, SCM 2017. pp. 481-482.

[4] Матвеев Ю.Н. Основы теории систем и системного анализа. Тверь: Твер.гос.техн. ун-т, 2007. 100 с.

[5] Mu Zhu, Hugh A. Chipman. Darwinian evolution in parallel universes: a parallel genetic algorithm for variable selection. Technometrix. 2006. Vol 48. N. 4. P. 491-502.

[6] Cantu-Paz E. Efficient and Accurate Parallel Genetic Algorithms. Massachusetts: Kluwer Academic Publishers. 2000. 162 p.

[7] Курейчик В.М., Курейчик В.В., Гладков Л.А. Теория и практика эволюционного моделирования. М: ФИЗМАТЛИТ. 2003. 432 с.

[8] Елисеева И.И., Юзбашев М.М. Общая теория статистики. М.: Финансы и статистика. 1995. 368 с.

[9] Царев Р.Ю. Модификация метода упорядоченного предпочтения через сходство с идеальным решением для задач многоцелевого принятия решения // Информационные технологии. 2007. № 7. С. 19-23.

[10] Madala H.R. and Ivakhnenko A.G. Inductive Learning Algorithms for Complex System Modeling. Florida. CRC Press. 1994. 368 p.

[11] Мокшин В.В. Параллельный генетический алгоритм отбора значимых факторов, влияющих на эволюцию сложной системы // Вестник КГТУ. 2009. № 3. С. 89-93.