

УДК 519.67

Мухамедиев Р.И.,*Университет имени Сулеймана Демиреля***Кучин Я.И., Якунин К., Саннова С.***Институт информационных и вычислительных технологий,
Алматы, Казахстан*

ОБРАБОТКА ДАННЫХ ГЕОФИЗИЧЕСКОГО ИССЛЕДОВАНИЯ СКВАЖИН С ПОМОЩЬЮ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ

Аннотация: В работе рассматриваются несколько алгоритмов машинного обучения, апробированных на задаче классификации данных электрического каротажа урановых скважин. Описываются физические аспекты процесса добычи урана на пластово-инфильтрационных месторождениях, каротаж скважин и результаты, полученные при применении некоторых классификационных методов.

По мнению авторов, геофизические исследования скважин широко используются на различных стадиях поисков, разведки и отработки месторождений полезных ископаемых, и в целом, могут быть охарактеризованы как технологический процесс получения количественной информации о геологических объектах, реализуемый путем совместного использования технических средств измерений, а также методического, алгоритмического, петрофизического и метрологического обеспечения. В результате исследований, авторы пришли к выводу, что наилучшим методом распознавания является использование нейронных сетей.

Ключевые слова: машинное обучение, каротаж, скважина, классификация, геофизическое исследование скважин.

1. Введение

На современном этапе развития геологоразведочных работ геофизические исследования скважин (ГИС) стали одним из основных источников информации о составе и свойствах геологических объектов, условиях их залегания[1].

Они широко используются на различных стадиях поисков, разведки и отработки месторождений полезных ископаемых, и в целом, могут быть охарактеризованы как технологический процесс получения количественной информации о геологических объектах, реализуемый путем совместного использования технических средств измерений, а также методического, алгоритмического, петрофизического и метрологического обеспечения.

При отработке пластово-инфильтрационных месторождений урана геофизическим методом исследования скважин отводится особое место и роль. Связано это, прежде всего, с тем, что подавляющее большинство скважин в этом случае проходятся без отбора керна. Поэтому ГИС является практически единственной информационной поддержкой всех разноплановых видов работ, выполняемых при отработке этих месторождений.

Подземное скважинное выщелачивание (ПСВ) - это сложный физико-химический процесс, протекающий в горных породах. Динамика этого процесса, его особенности определяются природными факторами: фациально-литологическими особенностями рудовмещающих горизонтов, физическими свойствами рудовмещающей толщи и всего разреза в целом. Контроль за ходом выщелачивания и управление процессом ПСВ предполагает тщательное изучение и учет всех этих природных факторов, что в свою

очередь предъявляет повышенные требования к максимальной полноте всей информации о среде.

Объекты исследований при ГИС – геологическая среда и элементы конструкции скважины. Они характеризуются определенными физическими свойствами, вещественным составом, определенными геометрическими размерами и могут быть описаны в виде конкретной геологической и петрофизической модели.

Большинство геологических, технических и геотехнологических задач, решение которых является конечной целью ГИС, может быть получено лишь в результате комплексной интерпретации данных различных методов ГИС, основанных на разных физических явлениях.

При отработке месторождений урана способом ПСВ ГИС являются основным, а зачастую и единственным методом получения наиболее полной информации об особенностях геологического разреза и характеристики уранового оруденения по каждой конкретной скважине. Они основаны на изучении естественных и искусственных физических полей во внутрискважинном, околоскважинном и межскважинном пространстве и проводятся с целью:

- изучения геологического разреза по всему стволу скважины в целом;
- детального изучения фациально-литологического строения рудовмещающих горизонтов;
- выявления рудных интервалов и параметров уранового оруденения (мощность, средние содержания, стволовые запасы);

Комплексная интерпретация с учетом всех данных ГИС требует больших затрат времени и профессионализма интерпретатора, в то время как технология производства требует быстрого принятия решений. Исчерпывающей теории для автоматического литологического расчленения с учетом всех видов каротажа сейчас нет. В этой связи была предпринята попытка использовать машинное обучение (МО) для проведения комплексной интерпретации для литологического расчленения пород вдоль ствола скважины. Для обучения использовались данные электрического и индукционного каротажа.

2. Электрический каротаж

Электрический каротаж – это метод исследования горных пород, основанный на регистрации параметров естественного или искусственного электрических полей.

Электрический каротаж, основанный на регистрации параметров естественного электрического поля, представляет собой каротаж потенциалов самопроизвольной поляризации (ПС). Измеряемой величиной является разность электрического потенциала ПС ($\Delta U_{ПС}$). Единица измерения – милливольт (мВ).

Измеряемой величиной является кажущееся удельное электрическое сопротивление (ρ_K) среды. Единица измерения – ом-метр (Ом·м). При отработке месторождений урана выполняется стандартный электрокаротаж подошвенными градиент-зондами, размеры которых выбраны постоянными для данного района (месторождения) работ.

В скважинах, заполненных промывочной жидкостью на непроводящей основе, а также обсаженных полиэтиленовыми (непроводящими) трубами, электрокаротаж с целью литолого-стратиграфического расчленения разреза скважин не выполняется.

Данные стандартного электрокаротажа являются одними из основных для получения информации о литолого-стратиграфическом и фациально-литологическом строении разреза скважин. Кроме того, используются для оценки фильтрационных свойств пород, слагающих рудовмещающий горизонт.

Минимальные требования к методическому обеспечению заключаются в наличии корреляционных зависимостей, связывающих геоэлектрические, гранулометрические параметры с фильтрационными свойствами пород.

Индукционный каротаж основан на измерении кажущейся удельной электрической проводимости δ_k пород в переменном электромагнитном поле в частотном диапазоне от десятков до сотен килогерц. В методе реализованы варианты измерения как активной компоненты кажущейся удельной электрической проводимости δ_a^a , которая пропорциональна ЭДС, так и реактивной компоненты δ_a^p , пропорциональной ЭДС, сдвинутой по фазе относительно тока генераторной цепи зонда на величину $\pi/4$. Единица измерения – сименс на метр (См/м), дробная – миллисименс на метр (мСм/м).

Является основным методом при определении мест перетоков технологических растворов из продуктивных в вышележащие горизонты и оценке их растекания в процессе ПСВ.

3. Алгоритмы машинного обучения

Машинное обучение – сфера исследований, которая направлена на анализ и разработку обучаемых систем. Задачи обучения, как правило, делятся на задачи обучения с учителем (supervised learning) и без учителя (unsupervised learning). При обучении без учителя решается задача кластеризации, когда система самостоятельно, используя некоторые заранее заданные критерии, формирует классы, куда относит рассматриваемые объекты. При обучении с учителем решается задача классификации объектов, после того как система была обучена на каком то наборе примеров. В задаче интерпретации данных каротажа такой набор примеров формируется экспертами. В настоящей работе исследованы следующие классификационные методы:

- Linear Discriminant Analysis Classifier (LDAC) [2]
- Support Vector Classification (SVM) (Linear SVM и Non-linear SVM) [3]
- Diagonal Linear Discriminant Analysis (DLDA)
- k-Nearest-Neighbor (k-NN) [4]

Кроме указанных алгоритмов использовались и многослойные нейронные сети с прямым распространением сигнала, описание которых и ход экспериментов приведены в других работах авторов [6,7,8].

3.1. Алгоритм k-Nearest-Neighbor (k-NN)

Алгоритм основан на подсчёте количества объектов каждого класса в сфере (гиперсфере) с центром в распознаваемом объекте. Объект относится к тому классу, объектов которого больше всего в этой сфере. В данном методе предполагается, что веса объектов выбраны единичными для всех объектов. Если веса не одинаковы, то вместо подсчёта количества объектов можно суммировать их веса. Таким образом если в сфере вокруг распознаваемого объекта 10 эталонных объектов класса А и 15 ошибочных/пограничных объектов класса Б, то точка будет отнесена к классу А.

Также веса объектов в сфере можно описать как обратно-пропорционально зависимые от расстояния до распознаваемого объекта. Таким образом, чем ближе объект, тем более значимым он является для данного распознаваемого объекта.

В итоге, метрический классификатор можно описать так:

$$a(u, X) = \arg \max_{y \in Y} \sum_{i=1}^n [y_i^* = y] w(i, u)$$

где $w(i, u)$ – вес i -того соседа распознаваемого объекта u ,
 $a(u; Xl)$ – класс объекта u , распознанный по выборке Xl .

Радиус гиперсферы может быть как фиксированным, так и динамичным. Причём в случае с динамическим радиусом, радиус для каждой точки подбирается так, чтобы количество объектов в каждой сфере было фиксированным. Таким образом, при распознавании в областях с разной плотностью выборки, количество «соседних» объектов (по которым и происходит распознавание) будет одинаково. То есть, в областях с низкой плотностью не будет ситуации, когда данных для распознавания недостаточно.

3.2. Support Vector Classification (Linear SVM и Non-linear SVM)

Этот алгоритм относится к группе граничных методов: он определяет классы при помощи границ областей. В основе метода лежит понятие плоскостей решений. Плоскость решения разделяет объекты с разной классовой принадлежностью. В пространствах высоких размерностей вместо прямых необходимо рассматривать гиперплоскости — пространства, размерность которых на единицу меньше, чем размерность исходного пространства. В R^3 , например, гиперплоскость — это двумерная плоскость.

Метод опорных векторов отыскивает образцы, находящиеся на границах классов (не меньше двух), т.е. опорные вектора и решает задачу нахождения разделения множества объектов на классы с помощью линейной решающей функции. Метод опорных векторов строит классифицирующую функцию f в виде:

$$f(x) = \text{sign}(\langle w, x \rangle + b),$$

где $\langle, \rangle$ – скалярное произведение,

w – нормальный (перпендикулярный) вектор к разделяющей гиперплоскости,

b – вспомогательный параметр, который равен по модулю расстоянию от гиперплоскости до начала координат. Если параметр b равен нулю, гиперплоскость проходит через начало координат.

Объекты, для которых $f(x)=1$ попадают в один класс, а объекты с $f(x)=-1$ — в другой.

Linear Discriminant Analysis Classifier (LDAC)

В математической статистике классификацию часто называют дискриминацией. Дискриминантный анализ - раздел многомерного статистического анализа, который включает статистические методы классификации многомерных наблюдений.

Линейный дискриминантный анализ — метод статистики и машинного обучения для поиска линейной комбинации переменных, наилучшим образом разделяющих два или более класса объектов или событий. ЛДА пытается выразить зависимую переменную (метку класса) через линейную комбинацию других признаков или измерений. Полученная комбинация может быть использована в качестве линейного классификатора. Признаки, используемые для отличия одного класса (подмножества) от другого, называются дискриминантными переменными.

Пусть обучающая выборка записана в матрицы $X1$ и $X2$, в которых имеется по $I1$ и $I2$ строк (объектов). Число столбцов (переменных) одинаково. Исходные предположения состоят в следующем:

каждый класс ($k=1$ или 2) представляется нормальным распределением;
 ковариационные матрицы этих двух классов одинаковые $\Sigma1 = \Sigma2 = \Sigma$.

Классификационное правило в LDA следующее: новый образец x относится к тому классу, к которому он ближе в метрике Махаланобиса:

$$d_k^2 = (x - \mu_k) \sum^{-1} (x - \mu_k)^T, k=1,2$$

На практике неизвестные математические ожидания и ковариационная матрица заменяются их оценкам

$$m_k = \frac{1}{I_k} \sum_{i=1}^{I_k} x_i,$$

$$S = \frac{1}{I_1 + I_2 - 2} (\tilde{X}_1^T \tilde{X}_1 + \tilde{X}_2^T \tilde{X}_2),$$

где $\tilde{X}_k$ – центрированная матрица X_k .

4. Сравнительный анализ методов машинного обучения

В качестве исходного набора данных, на которых были проведены вычислительные эксперименты, использовались результаты трех видов электрического каротажа (ИК – индукционный каротаж, KS – каротаж методом кажущихся сопротивлений, PS – Каротаж методом естественной поляризации) проведенные на 18 скважинах месторождения «Буденовское» и 12 скважин месторождения «Инкай».

Данные каротажа каждой скважины месторождения «Буденовское» были проинтерпретированы экспертами, которые выделили 10 основных литотипов (крупные и мелкозернистые пески, глина, смеси песка и глины, галечник и т.п.). Указанные данные, содержащие в общей сложности около 8000, значений использовались как для обучения нейронных сетей (NN) так и для оценки качества ее интерпретации.

Данные месторождения «Инкай» обрабатывались сходным образом. Отличие заключалось в том, что эти данные вместо экспертных оценок имели оценки по грансоставу, полученные методом ядерного опробования и не имели показателей индукционного каротажа. Общее количество литологических типов по данному месторождению составило около 120 пород, которые были объединены экспертом с получением 20 значимых для целей исследования литотипов. В процессе обучения ИНС множество данных разбивается на три группы: собственно обучающий набор данных, тестовое и поверочное множество. В задачах, где необходимо обучить сеть распознаванию рядов, используется скользящее окно данных [5]. В нашем случае, данные на вход сети поступали в виде скользящего окна данных с параметрами 5+1+5 (пять точек выше + текущая + пять точек ниже текущей). Разбиение такого множества на обучающий, тестовый и поверочный наборы затруднительно (каждая группа из 11 значений связана с другой), поэтому в работах [6,7] предложено наряду со стандартной методикой оценки (тестовое и поверочное множество данных) использовать для оценки качества распознавания данные скважин, которые никак не участвовали в процессе обучения (такие скважины условно названы «сырыми»). Такой подход позволяет полностью исключить возможность запоминания сетью или иным алгоритмом классификации данных распознаваемого набора.

4.1. Результаты, полученные по данным месторождения «Буденовское»

Проведенные вычислительные эксперименты (язык Python, пакет FANN) позволили получить результаты, показанные в Таблицах 3, 4, 5, 6, 7, 8, 9.

Таблица 3 - Результаты применения нейронной сети (NN) по всем комбинациям входных данных: KS, PS, IK, KS PS, KS IK, PS IK, KS PS IK

N	IK	KS	PS	KS_IK	KS_PS	PS_IK	KS_PS_IK
1	30	33	29	32	32	32	30
2	54	73	41	76	74	51	74
3	50	58	37	59	62	46	60
4	35	51	43	50	48	33	48
5	50	67	41	69	69	47	67
6	47	71	32	71	69	42	69
7	43	47	38	52	50	39	47
8	47	66	33	70	69	42	66
9	32	52	27	57	56	33	56
10	51	67	34	65	65	49	64
11	64	63	45	63	60	48	62
12	32	55	47	54	58	40	54
13	32	49	47	51	54	38	52
14	40	68	29	74	73	36	73
15	12	18	30	18	17	18	18
16	62	72	45	74	71	51	70
17	44	55	33	53	55	39	52
18	47	64	44	68	70	48	70
Среднее	43	57	37	59	58	41	57

Примечание. Колонки 2-8 содержат среднее после 10 итераций значение результата распознавания в %. Строки 1-18 – значения качества распознавания для каждой из 18 скважин при подаче на вход сети результатов разного вида каротажа и их сочетаний.

Последняя колонка – все доступные виды каротажа. Последняя строка – среднее значение процента распознавания по всем 18 скважинам.

Отметим, что полученные результаты показывают существенное влияние на качество распознавания данных каротажа кажущихся сопротивлений.

Классификационные алгоритмы LDAC, SVM, DLDA, k-NN оценивались в разных сочетаниях и также с применением плавающего окна данных (Таблицы 4, 5, 6, 7, 8).

Таблица 4 - Результаты применения LDAC

Комбинация входа	Без окна	Окно 3	Окно 5	Окно 7
KS_PS_IK	0,53	0,58	0,58	0,57
KS	0,50	0,54	0,54	0,54
PS	0,28	0,28	0,27	0,28
IK	0,33	0,34	0,34	0,34
KS_PS	0,52	0,57	0,57	0,57
KS_IK	0,52	0,57	0,57	0,57
PS_IK	0,33	0,34	0,34	0,34

Таблица 5 - Результаты применения Linear SVM

Комбинация входа	Без окна	Окно 3	Окно 5	Окно 7
KS_PS_IK	0,46	0,50	0,51	0,51
KS	0,41	0,44	0,44	0,44
PS	0,28	0,28	0,28	0,27
IK	0,35	0,35	0,35	0,35
KS_PS	0,43	0,46	0,46	0,47
KS_IK	0,46	0,50	0,50	0,50
PS_IK	0,35	0,35	0,35	0,35

Таблица 6 - Результаты применения Non-linear SVM

Комбинация входа	Без окна	Окно 3	Окно 5	Окно 7
KS_PS_IK	0,52	0,31	0,28	0,28
KS	0,55	0,58	0,50	0,41
PS	0,35	0,33	0,32	0,30
IK	0,43	0,45	0,43	0,40
KS_PS	0,54	0,40	0,31	0,28
KS_IK	0,53	0,43	0,34	0,30
PS_IK	0,43	0,31	0,28	0,27

Таблица 7 - Результаты применения DLDA

Комбинация входа	Без окна	Окно 3	Окно 5	Окно 7
KS_PS_IK	0,44	0,54	0,52	0,50
KS	0,28	0,28	0,28	0,28
PS	0,33	0,36	0,36	0,37
IK	0,44	0,54	0,52	0,50
KS_PS	0,45	0,55	0,53	0,51
KS_IK	0,33	0,34	0,33	0,33
PS_IK	0,45	0,54	0,52	0,51

Таблица 8 - Результаты применения алгоритма k-NN (50 соседних точек)

Комбинация входа	Без окна	Окно 3	Окно 5	Окно 7
KS PS IK	0,52	0,54	0,53	0,52
KS	0,54	0,59	0,59	0,58
PS	0,33	0,33	0,34	0,35
IK	0,41	0,41	0,41	0,41
KS PS	0,52	0,56	0,54	0,54
KS IK	0,52	0,56	0,56	0,55
PS IK	0,39	0,40	0,40	0,40

Таблица 9 - Сравнительные результаты применения алгоритмов к данным месторождения «Буденовское»

Комбинация входа	NN	LDAC	Linear SVM	Non-linear SVM	DLDA	k-NN
KS PS IK	57	58	50	31	54	54
KS	57	54	44	58	28	59
PS	37	28	28	33	36	33
IK	43	34	35	45	54	41
KS PS	58	57	46	40	55	56
KS IK	59	57	50	43	34	56
PS IK	41	34	35	31	54	40
Среднее	50.28	46	41.14	40.14	45	48.42

Примечание. В качестве результатов алгоритмов LDAC, SVM, DLDA, k-NN взяты лучшие показатели, полученные при применении плавающего окна размером 3.

4.2. Кривые обучения

При распознавании важную роль играет способность алгоритма обучаться, переобучаться (*overfitting*) или не дообучаться (*underfit*). Правильный баланс между *underfit* и *overfit* означает поиск такого алгоритма и его параметров, который был бы способен показать приемлимые результаты на тестовом множестве (или множестве *crossvalidation*). Недообученный алгоритм будет показывать одинаково плохие результаты и на тестовом и на тренировочном множествах, в то время как переобученный будет демонстрировать высокий результат на тренировочном и низкий на тестовом. Для выявления данных свойств алгоритмов были проведены дополнительные эксперименты на тестовом множестве данных для алгоритмов на базе NN и k-NN. Эксперименты показали, что с ростом числа скважин используемых для обучения показатели ошибок распознавания для NN уменьшаются, а для k-NN изменяются слабо. При этом, хотя оба алгоритма показывают близкие результаты по показателю точности (*accuracy*), показатели полноты (*recall*) и «аккуратности» (*precision*) для NN значительно выше (Рис.1 и 2).

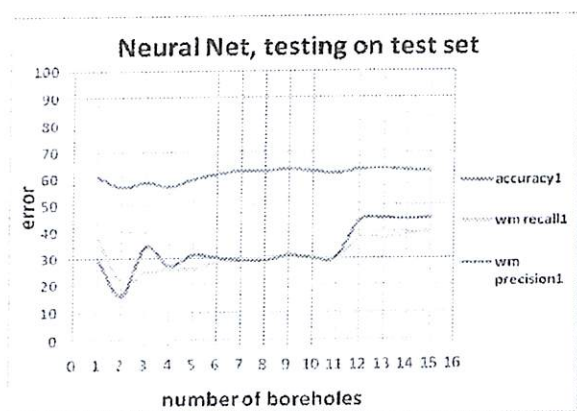


Рис.1. Кривые обучения NN

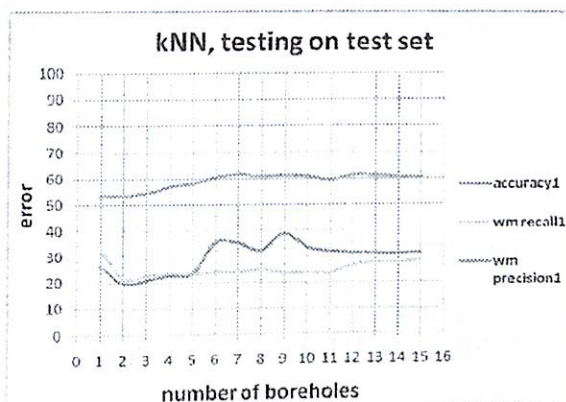


Рис. 2. Кривые обучения k-NN

5. Заключение

Результаты экспериментов показывают, что алгоритм k-NN показывает в среднем наилучший результат (исключая NN). При этом для каротажа кажущихся сопротивлений (KS), данный алгоритм абсолютный рекордсмен. Этот алгоритм показывает также наиболее близкие по сравнению с NN результаты в случае комбинаций входов. Очевидно, что LDAC и k-NN могут использоваться для оценки нижней границы распознавания при использовании NN. Наилучшим методом распознавания в нашем случае является использование нейронных сетей.

Список использованной литературы

- 1 ЗАО НАК «Казатомпром». ТОО ИВТ. Методические рекомендации по комплексу геофизических методов исследования скважин при подземном выщелачивании урана. Алматы, 2003 г.
- 2 Linear discriminant analysis. – 2012. [http://en.wikipedia.org/wiki/Linear_discriminant analysis](http://en.wikipedia.org/wiki/Linear_discriminant_analysis).
- 3 Support vector machine. – 2012. http://en.wikipedia.org/wiki/Support_vector_machine.
- 4 k-nearest neighbor algorithm. – 2012. http://en.wikipedia.org/wiki/K-nearest_neighbor_algorithm.
- 5 Kostikov D.V. Tools of interpretation of geophysical researches of boreholes on the basis of the transformed logging diagrams by means of a multilayered neural network. – M.: Russian state library, 2007. – P. 189
- 6 Ravil I. Muhamediyev, Yan Kuchin, Elena Muhamediyeva (2012). Recognition of Geological Rocks At the Bedded-infiltration Uranium Fields by Using Neural Networks // IEEE Conference on Open Systems. – Kuala Lumpur: IEEE Xplore database, 2012. – 6 p.
- 7 Ravil I. Muhamediyev, Yan Kuchin, Elena Muhamediyeva (2012). Geophysical Research of Boreholes: Artificial Neural Networks Data Analysis // The 6th International Conference on Soft Computing and Intelligent Systems and the 13th International Symposium on Advanced Intelligent Systems. – Kobe: scis2012.j-soft.org, 2012. – P. 825-829.

8 R. Muhamediyev, E. Amirgaliev, S. Iskakov, Y. Kuchin, E. Muhamedyeva. Integration of Results of Recognition Algorithms at the Uranium Deposits. JACIII, 2014, Vol.18 No.3. pp. 347-352

Мухамедиев Р.И.,

Сулейман Демирель атындағы университет, Қаскелең, Қазақстан

Кучин Я.И., Якунин К., Санинова С.

Ақпараттық және есептеуіш технологиялар институты, Алматы, Қазақстан

МАШИНАЛЫҚ ОҚУ АЛГОРИТМДЕРІН ПАЙДАЛАНА ОТЫРЫП ГЕОФИЗИКАЛЫҚ БАРЛАУ ҰНҒЫМЛАРЫНЫҢ ДЕРЕКТЕРІН ӨНДЕУ

Аңдатпа. Мақалада уран кен орындарында электр каротаждар деректерін классификациялауда қолданыс тапқан бірнеше машиналық оқыту алгоритмдері қарастырылады. Қабатты инфильтрационды кен орындарында уран алу үдерісінің физикалық сипаты, ұңғылар каротажы және кейбір классификациялау алгоритмдерінің қолданыс нәтижелері қарастырылады.

Кілт сөздер: машиналық оқыту, машиналық оқыту, ұңғы, классификациялау, ұңғыларды геофизикалық зерттеу

Muchamediyev R.I.,

Suleyman Demirel University, Kaskelen, Kazakhstan

Kuchin, Y.I., Yakunin K., Sainova S.,

Institute of Information and Computing Technology, Almaty, Kazakhstan

DATA PROCESSING OF GEOPHYSICAL EXPLORATION WELLS THROUGH THE USE OF MACHINE LERANING

Abstract. The paper deals with several machine learning algorithms, tested on the problem of classifying the data electric logging of uranium wells. It describes the physical aspects of uranium mining process at stratum-infiltration deposits, well logging, and the results obtained by applying some methods of classification.

According to the authors, geophysical surveys wells are widely used in various stages of exploration, prospecting and mining of mineral deposits, and, in general, can be described as a process of obtaining quantitative information about the geological objects, implemented by sharing technical means of measurement, as well as methodological, algorithmic, petrophysical and metrology software. As a result of studies, the authors concluded that the best use of neural networks is the recognition method.

Key words: machine learning, logging, well, classification, , geophysical exploration wells.