УДК 004.438 DOI: 10.21440/0536-1028-2018-1-42-47

ИССЛЕДОВАНИЕ МАШИННОГО ОБУЧЕНИЯ ПРОГНОЗУ ПАРАМЕТРОВ ОБОГАТИТЕЛЬНОГО ПРОЦЕССА

ЛЕОНОВ Р. Е.

Статья посвящена актуальной проблеме обучения компьютера прогнозированию некоторых параметров обогатительных процессов. Рассмотрен прогноз содержания мелких классов в руде, поступающей на обогатительную фабрику, по данным минералогического состава руды и прогноз содержания общего железа после заключительной стадии мокрой магнитной сепарации по технологическим данным предварительных стадий обогащения. С учетом небольшого объема статистических данных о процессах использована линейная регрессия совместно с процессом кросс-валидации на этапе обучения. Результаты прогноза проверены на независимых данных. Рассмотрена зависимость эффективности прогноза от количества проходов кросс-валидации. Обучение компьютера выполнено в приложении Anaconda3 языка Python 3.6.0, отладка программ обучения и последующего контроля произведена в Spyder и IPython.

Ключевые слова: машинный прогноз; обогатительные процессы; кросс-валидация.

Прогноз является одной из актуальных задач во многих отраслях производственной и научной деятельности. Горное производство, в частности обогащение, не является исключением. В 1970–80-е гг. широкое распространение в этой области получили методы распознавания образов [1–3]. С появлением в настоящее время новых технологий обработки информации, основанных на компьютерной базе, вновь возник интерес к вопросам прогноза с использованием обучения компьютеров и обработки больших массивов информации.

Большинство работ в этой области основано на теории распознавания образов. Однако применение компьютеров с их огромными возможностями позволило получить и новые качественные результаты. При этом следует учесть два обстоятельства.

Во-первых, эффективное обучение компьютера прогнозированию возможно только на представительном фактическом материале. За основу обучения процессу распознавания и в дальнейшем прогнозирования берут реальные данные – признаки и соответствующую каждому набору признаков выходную величину, представляющую основной интерес для прогнозирования, – отклик.

Чем больше таких наборов *признаки–отклик*, тем эффективнее процесс обучения машины и тем в последующем эффективнее прогноз. При этом предполагается, что условия, при которых получены данные для обучения, сохранятся в дальнейшем.

Во-вторых, в условиях рынка предприятия как правило не публикуют первичную информацию о технологических параметрах. Поэтому исходные наборы данных скупы и для обучения компьютера приходится использовать небольшие выборки. Положение осложняется тем, что из этой небольшой выборки необхо-

Леонов Рафаил Ефимович – кандидат технических наук, профессор кафедры автоматики и компьютерных технологий. 620144, г. Екатеринбург, ул. Куйбышева, 30, Уральский государственный горный университет. E-mail: lnprep2011@yandex.ru

димо еще удалить часть наборов *признаки–отклик* для создания контрольной выборки, с помощью которой в дальнейшем оценивают эффективность прогноза, которую осуществляет обученный компьютер.

Как неоднократно указывалось [4], оценка эффективности прогноза на нескольких экземплярах той выборки, по которой производилось обучение, дает излишне оптимистические результаты. В дальнейшем, при предъявлении компьютеру нового набора признаков, который отсутствовал в обучающей выборке, прогнозируемый отклик может сильно отличаться от того, который будет наблюдаться в действительности.

Исследована эффективность обучения и прогноза двух параметров технологического процесса обогащения: содержания класса C_x менее 0,15 мм в руде, поступающей на обогатительную фабрику (y_1), и содержания общего железа $Fe_{общ}$ после третьей стадии мокрой магнитной сепарации (y_2). Указанные параметры выбраны из тех соображений, что непосредственный оперативный контроль их затруднен.

В качестве признаков в первом случае рассматривались: содержание массивной руды (x_1), содержание брекчеевидной руды (x_2), вкрапления (x_3).

Во втором случае признаками были: содержание класса +0,53 мм в промпродукте первой стадии мокрой магнитной сепарации (ММС-1) – (x_1), плотность промпродукта ММС-1 – (x_2), содержание класса +0,53 мм в песках гидроциклона (x_3).

Для обучения и последующего прогноза использована линейная регрессия. В данном случае имеются некоторые особенности применения регрессии.

Известно, что регрессия не является способом аппроксимации, и предсказывать будущие события по ней рискованно. Это способ интерполяции. Получаемое уравнение действительно только на том наборе данных, по которым это уравнение получено. Несмотря на то что предполагается сохранение условий, по которым получено уравнение регрессии, проверка возможности предсказания результатов должна производиться на независимом наборе данных, который не был использован при выводе уравнения регрессии.

В связи с небольшим набором экспериментальных данных в обоих исследованиях для обеспечения независимости процесса обучения от процесса контроля использована кросс-валидация [4].

Обучение компьютера произведено по программам на языке Python: программы на Python широко доступны, бесплатны, имеется большое число специализированных программных пакетов, в том числе и для обучения компьютера. Использована версия Python 3.6.0, среда Anaconda3. Программы отлажены и проверены в приложении IPython – Spyder. Одной из целей проверки было сравнение эффективности прогноза на контрольных данных, предъявленных компьютеру, обученному на выборке без кросс-валидации и после кросс-валидации.

Целесообразно остановиться на способе кросс-валидации для правильной интерпретации полученных результатов.

Обучающая выборка делится на n частей (классов). Из обучающей выборки отбирается одна часть. На оставшихся после отбора данных производится обучение компьютера. Затем отобранная часть предъявляется компьютеру и на ней оценивается эффективность предсказания отклика. Считается, что такой контроль качества обучения достаточно эффективен, так как контрольная выборка не участвовала в процессе обучения. Затем отобранная выборка возвращается в обучающую совокупность и отбирается следующая из n частей. Обучение и контроль повторяются для всех n частей.

В итоге получаем n обученных компьютеров (моделей). В данном случае это n уравнений регрессии, из которых окончательно отбирается лучшее – то, которое

показывает минимальную остаточную дисперсию (минимальное остаточное среднеквадратичное отклонение) на заранее отобранной контрольной выборке. Эта лучшая модель используется в дальнейшем при прогнозе.

Далее приведены некоторые полученные результаты. За основу обучения взяты фрагменты программ, использованных для анализа продаж недвижимости, которые существенно переработаны, документированы, отлажены. Результаты прогноза по y_1 позволяют предсказать содержание мелких классов в руде в зависимости от ее минералогического состава. По второй решаемой задаче можно

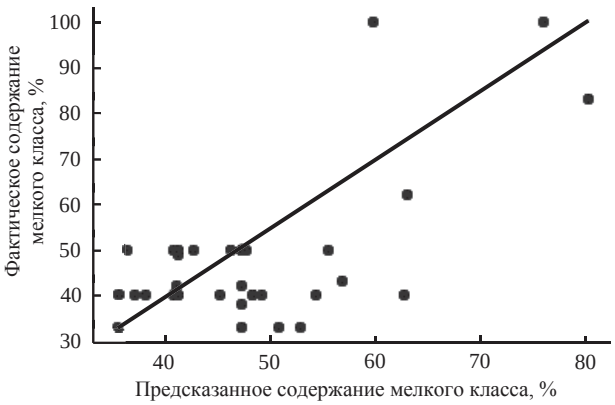

Рис. 1. Зависимость фактического содержания мелкого класса от предсказанного

прогнозировать содержание общего железа на выходе последней стадии магнитной сепарации y_2 в зависимости от измеряемых технологических показателей на предыдущих стадиях обогащения. Эти результаты имеют некоторое самостоятельное значение, хотя основной целью было исследование возможности применения прогноза с использованием кросс-валидации при малых выборках, полученных на действующих горных объектах.

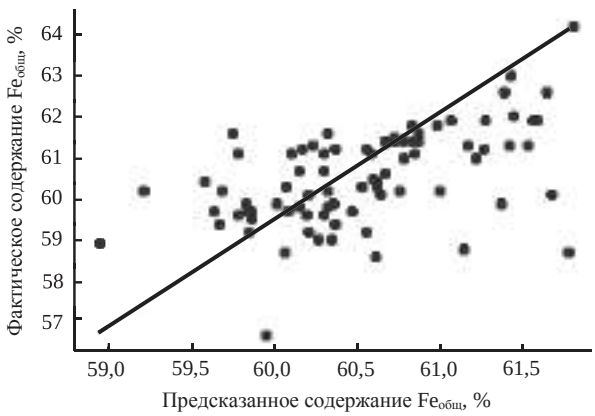

Рис. 2. Зависимость фактического содержания $Fe_{общ}$ от предсказанного

Обучающая выборка по прогнозу y_1 состояла из 40 значений, контрольная – из 8 значений, случайным образом отобранных из общей первоначальной совокупности данных (48 значений).

Обучающая выборка для прогноза y_2 состояла из 80 значений, контрольная – из 21 значения, отобранных случайным образом из первоначальной совокупности (101 значение). В обоих случаях контрольные экземпляры не участвовали в процессе обучения.

Поскольку отклик в обоих случаях зависит от трех факторов, зависимость отклика от факторов графически представить невозможно. Ввиду этого, на рис. 1, 2 приведены фактические значения отклика (точки) и рассчитанные по уравнению регрессии значения. В случае, если бы рассчитанные значения точно совпали с наблюдаемыми, все фактические значения легли бы на прямую линию. Как видно из рисунков, в действительности это не так, в обоих случаях имеется довольно большое расхождение. Это расхождение вызвано малой представительностью данных.

Таблица 1

Обобщенные результаты контроля качества прогноза

Показатель	y_1	y_2
Величина обучающей выборки	40	80
Величина контрольной выборки	8	21
Остаточное стандартное отклонение (без кросс-валидации)	10,197	1,470
Остаточное стандартное отклонение (кросс-валидация $n = 5$)	11,153	1,530
Коэффициент детерминации	0,419	0,278
Коэффициенты оптимальной модели	$y = a_0 + a_1x_1 + a_2x_2 + a_3x_3$	
a_0	52,0515	62,1020
a_1	–0,095	–0,080
a_2	–0,134	0,014
a_3	–0,0605	–0,0500
Оптимальное количество n	25	50
Остаточное стандартное отклонение при оптимальном n	10,040	1,468

Так коэффициент детерминации для y_1 составил $K = 0{,}419$, для y_2 $K = 0{,}278$. Учитывая, что максимальное значение этого коэффициента при идеальном обучении составляет единицу, следует признать данные малопредставительными. Однако и в этом случае, несмотря на небольшие объемы выборок, метод кросс-валидации позволил получить результаты, почти соизмеримые с обучением по всей выборке. При этом соблюдено основное правило прогноза: контрольная выборка должна быть отделена от обучающей.

В табл. 1 приведены обобщенные результаты контроля качества прогноза.

Следует остановиться на важном вопросе – определении количества групп n, на которые следует делить обучающую выборку при кросс-валидации. Практические рекомендации в этой области предлагают определять объем группы при кросс-валидации в 20 % от объема обучающей выборки.

В табл. 2 приведены остаточные стандартные отклонения прогноза в зависимости от величины n.

Из табл. 2 видно, что, применяя кросс-валидацию, следует проверить модели для различных значений n, из которых затем можно выбрать лучшую модель,

которая дала минимальное остаточное стандартное отклонение. Видно также, что лучшая модель при использовании кросс-валидации позволила получить результаты прогноза на независимых данных лучшие, чем модель без кросс-валидации. Действительно, если не принимать во внимание $n = 2$ для y_1, что означает использование для обучения половины данных (20 значений) и для прогноза при кросс-валидации оставшиеся 20 значений, то видно, что при большом числе n и для y_1, и для y_2 остаточное стандартное отклонение стабилизируется и в обоих случаях оно меньше, чем без кросс-валидации.

Таблица 2

Остаточные стандартные отклонения прогноза в зависимости от величины *n*

n	y_1	y_2
2	5,21	1,58
3	9,60	1,53
4	10,35	1,57
5	11,15	1,53
10	10,13	1,50
15	10,18	1,50
20	10,18	1,18
25	10,04	–
30	10,04	1,47
35	10,04	–
39	10,04	–
40	–	1,47
50	–	1,47
70	–	1,47
79	–	1,46

Из проведенных исследований можно сделать следующие выводы. Использование готовых пакетов языка Python для обучения компьютера позволяет при наличии обучающих данных получить модели, пригодные для прогноза технологических процессов горного производства.

Использование процесса кросс-валидации даже при наличии небольшого набора данных позволяет получить удовлетворительные результаты прогноза.

На примере таких разных параметров горного процесса, как крупность классов и содержание общего железа на выходе обогатительной фабрики, показано, что прогноз может быть выполнен по единообразной методике с использованием стандартных пакетов языка Python.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Вапник В. Н., Червоненкис А. Я. Теория распознавания образов. М.: Наука, 1974. 416 с.
2. Нильсон Н. Обучающиеся машины. М.: Мир, 1967. 180 с.
3. Браиловский В. Л., Лунц А. Л. Формулировка задачи распознавания объектов со многими параметрами и методы ее решения // Изв. АН СССР. Техническая кибернетика. 1969. № 1. С. 20–33.
4. Коэльо Л. П., Ричарт В. Построение систем машинного обучения на языке Python. М.: ДМК Пресс, 2016. 302 с.

Поступила в редакцию 26 октября 2017 года

INVESTIGATION OF MACHINE LEARNING TO FORECAST THE PARAMETERS OF A CONCENTRATING COMPLEX

Leonov R. E. – The Ural State Mining University, Ekaterinburg, the Russian Federation. E-mail: lnprep2011@yandex.ru

The article is dedicated to an urgent problem of machine learning to forecast some parameters of concentrating processes. The forecast of fine grains content in ore which comes to a concentrating mill is studied according to the data of ore mineral composition, together with the forecast of the content of total iron after the final phase of wet magnetic separation according to the technological data of the preliminary phases of separation. With the account of small amount of statistical data about the processes, the linear regression together with the process of cross-validation at the training phase is used. The results of the forecast have been verified with independent data. The forecast efficiency dependence is examined on the quantity of cross-validation procedures. Machine learning has been fulfilled with the application Anaconda3 of the Python 3.6.0 language, learning program debug and further control have been fulfilled with Spyder and IPython.

Key words: machine forecast; concentrating processes; cross-validation.

REFERENCES

1. Vapnik V. N., Chervonenkis A. Ia. *Teoriia raspoznavaniia obrazov* [Theory of pattern recognition]. Moscow, Nauka Publ., 1974. 416 p.
2. Nil'son N. *Obuchaiushchiesia mashiny* [Learning machines]. Moscow, Mir Publ., 1967. 180 p.
3. Brailovskii V. L., Lunts A. L. [Defining the problem of pattern recognition with a variety of parameters and the procedures for its solution]. *Izvestiia Akademii nauk SSSR. Tekhnicheskaia kibernetika – Bulletin of the Academy of Sciences of the USSR. Engineering Cybernetics*, 1969, no. 1, pp. 20–33. (In Russ.)
4. Koel'o L. P., Richart V. *Postroenie sistem mashinnogo obucheniia na iazyke Python* [Development of the systems of machine learning in the Python language]. Moscow, DMK Press Publ., 2016. 302 p.