УДК 543.422.8, 681.3.06

АНАЛИЗ ПРОБ С НЕИЗВЕСТНОЙ МАТРИЦЕЙ С ИСПОЛЬЗОВАНИЕМ АЛГОРИТМОВ DATA MINING

© Е. И. Молчанова[1], Е. Н. Коржова[2], Т. В. Степанова[2], В. В. Кузьмин[1]

Статья поступила 16 февраля 2016 г.

Предложено при определении ограниченного числа аналитов в пробах сложного химического состава с неизвестной матрицей объединить алгоритмы Data Mining (задачи кластеризации и регрессии). Это позволяет компенсировать влияние компонентов вмещающей среды на интенсивность аналитической линии определяемого элемента. Разработанная технология опробована при рентгенофлуоресцентном определении S, Fe, Cu, Zn, As в пробах флотоконцентрата при переработке полиметаллических руд и V, Fe в синтетических пленочных образцах, адекватных по физико-химическим свойствам пробам сварочных аэрозолей, собранных на фильтр. Погрешность результатов анализа уменьшилась в 1,5 – 5 раз по сравнению с использованием классического уравнения регрессии Лукас-Туса. Применение разработанной технологии существенно повышает экспрессность анализа при использовании рентгеновских спектрометров последовательного действия.

Ключевые слова: алгоритмы Data Mining; кластер; гетерогенные материалы; рентгенофлуоресцентный анализ; модели градуировочных функций; уравнения регрессии; метод наименьших квадратов; градуировочные образцы; погрешность адекватности.

Постановка задачи

При анализе гетерогенных материалов модели градуировочных функций нередко задают уравнениями регрессии, переменными в которых служат интенсивности I_j аналитических линий элементов пробы j [1]. Особенностью этих уравнений является то, что коэффициенты, описывающие межэлементные взаимодействия, определяют методом наименьших квадратов (МНК) по группе градуировочных образцов (ГО). Часто на практике возникают аналитические задачи, когда регистрируют только параметры для аналитов, и неизвестны характеристики вмещающей среды. В таких условиях погрешность адекватности обычных уравнений регрессии недопустимо велика [2].

Выбор методов исследования

Data Mining — исследование и обнаружение «машиной» (алгоритмами, средствами искусственного интеллекта) в «сырых» данных скрытых знаний, которые ранее не были известны, нетривиальны, практически полезны, доступны для интерпретации человеком [3].

Методы Data Mining помогают решать следующие основные задачи анализа данных:

классификация (определение класса объекта по его характеристикам);

регрессия (определение значения некоторого параметра объекта по известным его характеристикам);

поиск ассоциативных правил (нахождение частных зависимостей между объектами или событиями);

кластеризация (поиск независимых групп и их характеристик во всем множестве анализируемых данных).

В Data Mining задачу классификации рассматривают как задачу определения значения одного из параметров анализируемого объекта на основании значений других параметров. Определяемый параметр часто называют зависимой переменной, а параметры, участвующие в его определении, — независимыми переменными. Если значениями независимых и зависимых переменных являются действительные числа, то задача называется задачей регрессии.

Задачу классификации и регрессии решают в два этапа. На первом выделяют обучающую выборку. В нее входят объекты, для которых известны значения как независимых, так и зависимых переменных. На втором этапе построенную модель применяют к анализируемым объектам (к объектам с неопределенным значением зависимой переменной).

В практике физических и химических методов анализа задачи построения модели градуировочной функции методики (первый этап) и анализа неизвестных проб (второй этап) принято рассматривать как задачи регрессии. Можно ожидать, что объединение алгоритмов задач кластеризации и регрессии позволит получить новые знания об объекте исследования (принадлежность к кластеру) и повысить точность результатов анализа.

[1] Иркутский государственный университет путей сообщения, г. Иркутск, Россия; e-mail: moleli59@gmail.com
[2] Иркутский государственный университет, г. Иркутск, Россия; e-mail: rfa@chem.isu.ru

Разработка технологии Data Mining

Предлагаемая технология Data Mining объединяет методы регрессии (этап построения градуировочной функции по обучающей выборке градуировочных образцов), классификации и кластеризации (этап анализа на основе определения принадлежности состава пробы (**C**) к соответствующему кластеру ГО). При использовании рентгенофлуоресцентного метода на этапе анализа от пробы измеряют только интенсивности аналитических линий k определяемых элементов, которые сопоставляют с аналогичными аналитическими характеристиками одного градуировочного образца-соседа с целью компенсации неучтенного влияния матрицы. Состав образца-соседа $(\mathbf{C})_c$ выбирают наиболее близким к составу пробы (**C**). Так как число градуировочных образцов ограничено, можно допустить некоторую разницу в содержаниях элементов, интенсивности линий которых регистрируют. Влияние этой разницы на результат анализа учитывают с помощью выражения:

$$C_i = \frac{C_i^c I_i}{I_i^c} + a_1\left(1-\frac{I_i}{I_i^c}\right) + \sum_j^k \alpha_{ij}(I_j - I_j^c)I_i. \qquad (1)$$

Здесь C_i — содержание определяемого элемента i; I_i и I_j — интенсивности аналитических линий соответственно определяемого i и мешающего j элементов из числа k; с — индекс, обозначающий образец-сосед; коэффициенты a_1, α_{ij} оценивают предварительно на этапе градуирования.

Это выражение получено на основе классического уравнения Лукас-Туса [4]:

$$C_i = a_0 + I_i\left(a_1 + \sum_j \alpha_{ij} I_j\right). \qquad (1а)$$

Для выбора состава образца-соседа $(\mathbf{C})^c$ при анализе каждой пробы выполняют кластерный анализ выборки из n градуировочных образцов. Каждый ГО представляет собой отдельный кластер.

Тогда матрица отличий **D** векторов состава пробы (**C**) и соседа $(\mathbf{C})^c$ будет иметь вид вектора-строки:

$$\mathbf{D} = d[\{(\mathbf{C}), (\mathbf{C})^{c1}\}]d[(\mathbf{C}), (\mathbf{C})^{c2}) \ldots d[(\mathbf{C}), (\mathbf{C})^{cn}]. \qquad (2)$$

Здесь d — мера расстояния между составом пробы и образца-соседа.

С учетом того, что количественно влияние элемента j на интенсивность I_i зависит не только от содержания C_j, но и от величины эффектов поглощения и подвозбуждения, при оценке d учитывали вес каждого значения C_j. В качестве веса C_j использовали регрессионные коэффициенты α_{ij}, рассчитанные для выражения (1).

В выражение (1) входят величины I_j, а не C_j, поэтому для оценки меры расстояния d использовали метрику[3] δ в форме:

$$\delta = a_1\left|1-\frac{I_i}{I_i^c}\right| + \sum_j^k \alpha_{ij}|I_j - I_j^c|I_i. \qquad (3)$$

Исходя из того, что адекватность модели (1) будет тем выше, чем меньше поправочный член (3), критерием принадлежности пробы к кластеру конкретного ГО служит величина $\min\delta$.

В алгоритме выбора образца-соседа для каждой пробы дополнительно накладывали ограничение $I_i/I_i^c \to 1$ (или по массе $M/M^c \to 1$ для ненасыщенных слоев).

После выбора образца-соседа C_i для пробы рассчитывали по выражению (1).

Предложенную технологию испытали при решении двух задач.

Задача 1. Определение (%) Fe (2,0 – 36,8), Zn (0,1 – 1,2), Cu (0,05 – 6,4), S (3,7 – 44,5), As (0,02 – 3,3) в пробах флотоконцентрата при переработке полиметаллических руд.

Аппаратура: вакуумный рентгеновский спектрометр последовательного действия «Спектроскан» МАКС-GV (НПО «Спектрон», Санкт-Петербург), рентгеновская трубка с Pd-анодом, напряжение — 40 кВ, сила тока — 2 – 4 мА.

Характеристика анализируемого материала. При проведении экспериментальных исследований градуировочные образцы были представлены выборкой из 11 проб анализируемого продукта, в которых содержания определяемых элементов установлены химическим, атомно-абсорбционным и рентгенофлуоресцентным методами. Состав нерудного компонента, представляющего матрицу проб, оставался неизвестен. Для каждой пробы было приготовлено по три излучателя в виде двухслойных таблеток на основе борной кислоты, представляющих независимые измельчения материала пробы. Таким образом, в эксперименте использовали 33 ГО, для которых зарегистрировали интенсивности характеристического рентгеновского излучения только пяти аналитов (S, Fe, Cu, Zn, As). Проводили коррекцию зарегистрированных интенсивностей на фон.

Применение технологии кластеризованной регрессии. Для примера на рис. 1 приведены зависимости $C_i = f(I_i)$ для Fe и S в ГО. На обоих графиках точки для ГО разделяются на два класса: I — 18 ГО с большим (14,6 – 36,8 %) содержанием Fe; II — 15 ГО с малым (2,0 – 5,8 %) содержанием Fe. Это объясняется тем, что в ГО II класса существенно выше доля нерудного компонента, влияние которого не учитывают.

[3] Метрика — правило вычисления расстояний между любой парой объектов исследуемого множества [3].

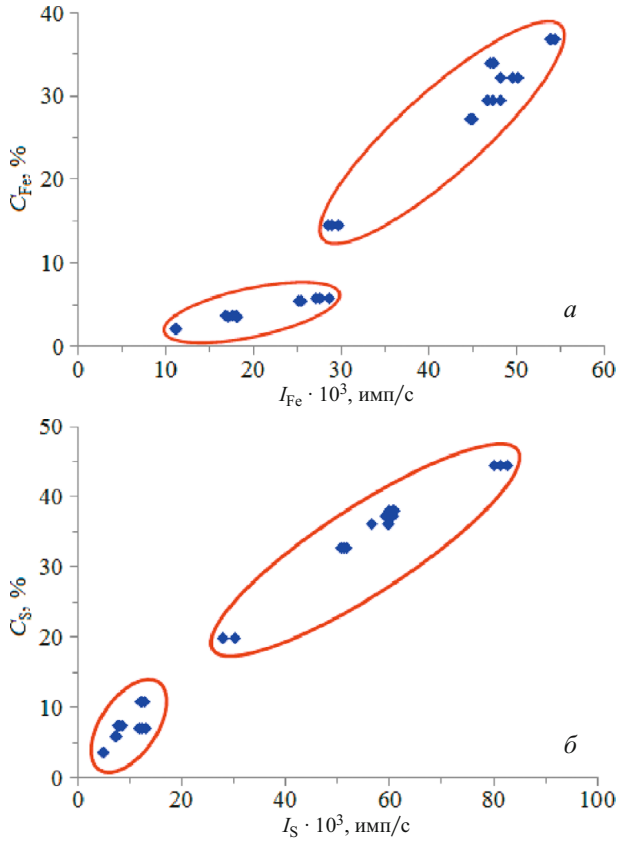

Рис. 1. Зависимости $C_i = f(I_i)$ при определении Fe (*а*) и S (*б*) в 33 ГО флотоконцентрата

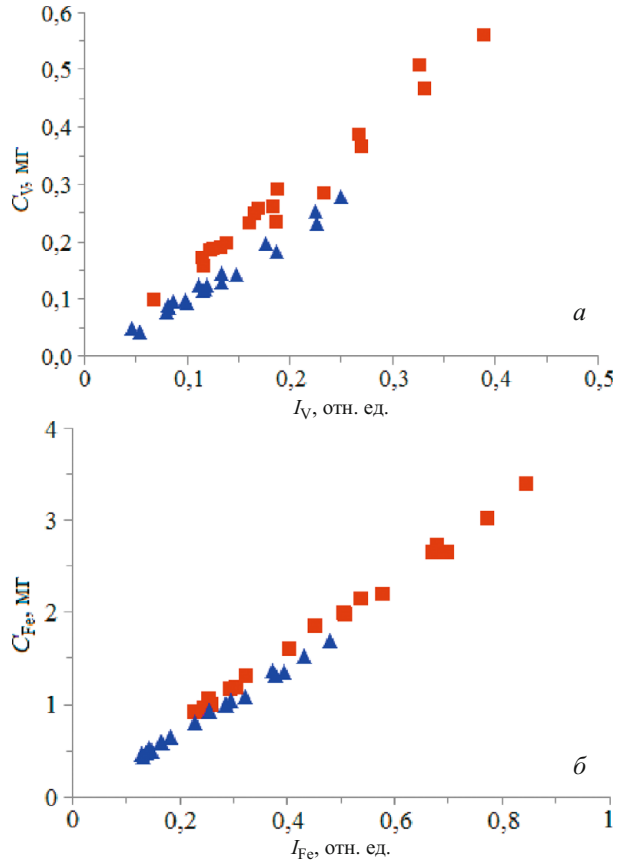

Рис. 2. Зависимости $C_i = f(I_i)$ для определения Fe (*а*) и V (*б*) в синтетических пленках: ▲ — $M < 80$ мг, ■ — $M > 80$ мг

Сначала с помощью полученных характеристик всех 33 ГО рассчитали коэффициенты уравнения (1а), используя взвешенный метод наименьших квадратов (статистический вес $1/\sqrt{C_i}$) для всего диапазона C_i [3]. По полученным коэффициентам рассчитали содержания S, Fe, Cu, Zn и As в этих ГО. В таблице приведена остаточная погрешность (коэффициент вариации V_0) определения аналитов в ГО. Невысокая точность определения химического состава ГО и отсутствие данных о характеристиках нерудного компонента приводят к высокой остаточной погрешности адекватности уравнения (1а).

На следующем этапе влияние нерудного компонента компенсировали сопоставлением аналитических характеристик пробы с таковыми для ГО, принадлежащего к тому же кластеру (группе) по выражению (1).

Как видно из таблицы, разработанная технология позволяет в 2 – 5 раз повысить точность определения аналитов в пробах с неизвестной матрицей без измерения дополнительных спектральных характеристик, что существенно повышает экспрессность анализа с применением рентгеновских спектрометров последовательного действия.

Остаточная погрешность определения аналитов в ГО флотоконцентрата

Уравнение	Коэффициент вариации V_0 (%) для аналита				
	Fe	Zn	Cu	S	As
(1а)	7,0	8,4	11	9,2	25
(1)	3,0	3,2	5,0	3,8	5,2

Задача 2. Определение (мг) V (0,05 – 0,6), Cr (0,04 – 0,7), Mn (0,04 – 1,2), Fe (0,4 – 3,4) и Ni (0,04 – 1,2) в синтетических градуировочных образцах, адекватных по физико-химическим характеристикам сварочным аэрозолям, собранным на фильтр.

Аппаратура: рентгеновский спектрометр последовательного действия VRA-30 (Carl Zeiss, Германия), рентгеновская трубка с Rh-анодом, напряжение — 40 кВ, сила тока — 40 мА. Для учета аппаратурного дрейфа интенсивности аналитических линий регистрировали в относительных единицах, используя в качестве образца-репера таблетку, спрессованную из оксидов определяемых элементов и борной кислоты.

Характеристика анализируемого материала. Образцы представляют собой органические пленки, содержащие тонкоизмельченный порошок (носитель аэрозолей) известного химического состава. Их получали по технологии [4, 5], порошковый носитель аналитов готовили в соответствии с рекомендациями работы [6]. Массовая доля порошка в пленке изменяется от 4 до 10 %. Аналиты — оксиды Fe_2O_3, Mn_2O_3, V_2O_5, Cr_2O_3 и NiO, в качестве наполнителя использовали соединения CaF_2, SiO_2 и NaF. Было приготовлено 38 пле-

ночных образцов, их масса (M) изменяется от 40 до 100 мг.

Применение технологии кластеризованной регрессии. Для примера на рис. 2 приведены зависимости $C_i = f(I_i)$ для Fe и V. Видно, что точки для ГО разделяются на два класса: I — 19 ГО, имеющих массу $M < 80$ мг, и II — 19 ГО с $M > 80$ мг.

Остаточная погрешность определения Fe и V во всех ГО по уравнению (1а) с введением поправок только на определяемые элементы (V, Cr, Mn, Fe и Ni) характеризуется коэффициентом вариации V_0, равным 4,2 и 10,4 % соответственно. При использовании технологии кластеризованной регрессии после выбора образца-соседа и расчета C_i для пробы по выражению (1) коэффициент вариации V_0 для Fe и V составил 3,3 и 6,9 % соответственно, т.е. погрешность результатов анализа по сравнению с использованием уравнения Лукас-Туса уменьшилась примерно в 1,5 раза.

Таким образом, предложена технология анализа проб с неизвестной матрицей с использованием алгоритмов Data Mining. Применение алгоритма кластеризации в задаче регрессии позволяет не вводить поправки на влияние компонентов вмещающей среды. Следует отметить, что поле состава ГО должно покрывать поле состава анализируемых материалов, т.е. выборка ГО должна быть репрезентативной.

ЛИТЕРАТУРА

1. **Молчанова Е. И., Смагунова А. Н., Козлов А. В., Азьмуко Н. Н.** Уравнения связи в рентгенофлуоресцентном анализе (обзор) / Заводская лаборатория. 1994. Т. 60. № 2. С. 12 – 21.
2. **Базыкина (Коржова) Е. Н., Смагунова А. Н., Слободняк Т. Г., Кубарев С. В.** Рентгеноспектральный анализ технологических растворов / Заводская лаборатория. 1981. Т. 47. № 9. С. 56 – 59.
3. **Барсегян А. А., Куприянов М. С., Степаненко В. В., Холод И. И.** Методы и модели анализа данных: OLAP и Data Mining. — СПб.: БХВ-Петербург, 2004. — 336 с.
4. **Lucas-Tooth H. J., Price B. J.** A mathematical method for the investigation of inter-element effects in X-ray fluorescent analysis / Metallurgia. 1961. Vol. 64. N 383. P. 149 – 152.
5. **Молчанова Е. И., Смагунова А. Н., Смагунов А. В.** Способы повышения точности построения градуировочной характеристики с помощью уравнений связи в рентгенофлуоресцентном анализе / Заводская лаборатория. 2000. Т. 66. № 4. С. 16 – 20.
6. Пат. РФ № 2239170, МПК[7] G 01 N1/28. Способ изготовления стандартных образцов атмосферных аэрозолей, нагруженных на фильтр / Коржова Е. Н., Смагунова А. Н., Кузнецова О. В., Козлов В. А.; заявл. 30.08.02; опубл. 13.04.04, бюл. № 17.
7. Пат. РФ № 2324915, МПК G 01 N1/28. Способ изготовления стандартных образцов атмосферных аэрозолей, нагруженных на фильтр / Коржова Е. Н., Смагунова А. Н., Карпукова О. М., Козлов В. А.; заявл. 20.03.06; опубл. 20.05.08.
8. **Степанова Т. В., Смагунова А. Н., Коржова Е. Н.** Выбор порошка-носителя аналитов для приготовления градуировочных образцов при рентгенофлуоресцентном анализе сварочных аэрозолей / Аналитика и контроль. 2015. Т. 19. № 2. С. 1 – 7.

REFERENCES

1. **Molchanova E. I., Smagunova A. N., Kozlov A. V., Az'muko N. N.** Uravneniya svyazi v rentgenofluorestsentnom analize (obzor) [The coupling equations in X-ray fluorescence analysis (review)] / Zavod. Lab. 1994. Vol. 60. N 2. P. 12 – 21 [in Russian].
2. **Bazykina (Korzhova) E. N., Smagunova A. N., Slobodnyak T. G., Kubarev S. V.** Rentgenospektral'nyi analiz tekhnologicheskikh rastvorov [X-ray spectral analysis of technological solutions] / Zavod. Lab. 1981. Vol. 47. N 9. P. 56 – 59 [in Russian].
3. **Barsegyan A. A., Kupriyanov M. S., Stepanenko V. V., Kholod I. I.** Metody i modeli analiza dannykh: OLAP i Data Mining [Methods and models of data analysis: OLAP and Data]. — St. Petersburg: BKhV-Peterburg, 2004. — 336 p. [in Russian].
4. **Lucas-Tooth H. J., Price B. J.** A mathematical method for the investigation of inter-element effects in X-ray fluorescent analysis / Metallurgia. 1961. Vol. 64. N 383. P. 149 – 152.
5. **Molchanova E. I., Smagunova A. N., Smagunov A. V.** Sposoby povysheniya tochnosti postroeniya graduirovochnoi kharakteristiki s pomoshch'yu uravnenii svyazi v rentgenofluorestsentnom analize [Ways to improve the accuracy of construction of calibration characteristics using coupling equations in x-ray fluorescence analysis] / Zavod. Lab. 2000. Vol. 66. N 4. P. 16 – 20. [in Russian].
6. RF Pat. No. 2239170, **Korzhova E. N., Smagunova A. N., Kuznetsova O. V., Kozlov V. A.** Sposob izgotovleniya standartnykh obraztsov atmosfernykh aérozolei, nagruzhennykh na fil'tr [A procedure for preparing standard reference samples of atmospheric aerosols deposited on the filter], 2004 [in Russian].
7. RF Pat. No. 2324915, **Korzhova E. N., Smagunova A. N., Karpukova O. M., Kozlov V. A.** Sposob izgotovleniya standartnykh obraztsov atmosfernykh aérozolei, nagruzhennykh na fil'tr [A procedure for preparing standard reference samples of atmospheric aerosols deposited on the filter], 2008 [in Russian].
8. **Stepanova T. V., Smagunova A. N., Korzhova E. N.** Vybor poroshka-nositelya analitov dlya prigotovleniya graduirovochnykh obraztsov pri rentgenofluorestsentnom analize svarochnykh aérozolei [The choice of analytes' powder-carrier for preparing calibration samples when analyzing welding fumes using X-ray fluorescence] / Analit. Kontrol'. 2015. Vol. 19. N 2. P. 1 – 7 [in Russian].