

УДК 681.785.57

© И. В. Скутин, К. А. Жаринов

ПРИМЕНЕНИЕ МЕТОДОВ МАТЕМАТИЧЕСКОГО ПРОГРАММИРОВАНИЯ ДЛЯ ГРАДУИРОВКИ БИК-АНАЛИЗАТОРОВ

В статье рассматривается возможность повышения точности анализа различных образцов, проводимого с помощью анализатора, работающего в ближней инфракрасной области. Для решения задачи предлагается создание градуировки с помощью методов математического программирования.

Кл. сл.: градуировка, БИК-анализатор, математическое программирование

ВВЕДЕНИЕ

На сегодняшний день во многих отраслях промышленности, в том числе и в сельском хозяйстве, одной из главных задач является исследование состава продукции. Подобные данные просто необходимы для оптимизации технологического процесса (сокращение расхода сырья, повышение качества продукции), определения стоимости продукции, идентификации и обнаружения фальсификации образцов. Для решения данной задачи в настоящее время все чаще используют методы, основанные на спектральном анализе в ближней инфракрасной (БИК) области [1].

Преимущества метода БИК-анализа в том, что он занимает значительно меньше времени и не требует применения дополнительных реактивов, высокой квалификации персонала и хорошо оборудованных лабораторий, как в случае с методами химического анализа. Анализ проводится с помощью специальных приборов, удобных в эксплуатации и позволяющих в короткие сроки получить необходимые результаты с достаточно высокой точностью.

Однако, прежде чем использовать анализатор, его необходимо предварительно отградуировать, чтобы найти взаимосвязь между содержанием анализируемых компонентов и спектральными характеристиками. Для этого отбираются градуировочный и валидационный (проверочный) наборы образцов с известными референтными значениями, а также регистрируются их спектральные данные. Референтные значения градуировочного и валидационного наборов — это анализируемые параметры образцов, которые определяются при помощи независимых референтных методов, имеющих заданную точность и воспроизводи-

мость. Величины, измеряемые непосредственно анализатором, называются спектральными данными. Диапазон изменения референтных значений образцов должен полностью охватывать диапазон возможных изменений этих значений при последующем анализе неизвестных образцов. Кроме того, для создания наиболее устойчивой модели образцы в градуировочном наборе должны быть однородно распределены в диапазоне изменения анализируемых свойств, а также количество валидационных образцов должно быть не меньше 50 [2]. Таким образом, формируются матрица X спектральных данных градуировочного набора размерности $n \times f$ (n — количество образцов в градуировочном наборе, а f — число точек спектра) и матрица Y референтных значений градуировочного набора размерности $n \times k$ (k — количество исследуемых компонентов). Матрицы X' и Y' валидационного набора отличаются от градуировочного меньшим количеством образцов.

Известно, что в БИК-спектроскопии используемые количественные связи зачастую слишком сложны, т. к. спектры в БИК-области содержат множество широких перекрывающихся полос поглощения (отражения, рассеивания), которые накладываются на полосу интересующего компонента. Это не позволяет построить простую градуировочную зависимость между интенсивностями характерной полосы с содержанием интересующего вещества. В таком случае формируют многомерную градуировочную модель (MVC — multivariate calibration), которая связывает весь массив спектральной информации с содержанием интересующего компонента.

Наиболее часто для создания MVC используют метод проекции на латентные структуры (PLS —

partial least squares) и метод регрессии на главных компонентах (PCR — principal component regression) [3]. PCR — довольно простой метод, требующий меньше вычислительной мощности, чем PLS, его используют для хорошо изученных наборов, где построение MVC с удовлетворяющими характеристиками не вызывает особых трудностей. PLS создан на основе PCR и существенно расширяет его возможности. Метод хорошо работает, когда в спектральных данных существует сильная корреляция или имеется набор спектральных данных, число столбцов в котором сильно превышает число строк [4]. Однако при наличии сложных связей в спектральных данных или отсутствии ярко выраженных особенностей для создания лучшей MVC возможности обоих методов весьма ограничены за счет удаления латентных переменных высших порядков.

Нейронные сети все больше набирают популярность в задачах создания MVC. Тем не менее наряду с преимуществами они имеют два основных недостатка. Их архитектура должна быть определена заранее, т. к. изменение ее в процессе обучения довольно сложная задача. Вторым недостатком в том, что нейронные сети могут застревать в локальных минимумах во время обучения, а во многих случаях это нежелательно [5].

Таким образом, поиск более выгодной альтернативы в создании MVC для решения конкретных задач, где общие методы слабы, является актуальной задачей. В качестве такой альтернативы можно предложить технику математического программирования (МП), которая успешно используется в подобных задачах построения MVC [6].

ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

Математическое программирование в задачах многомерной градуировки

Математическое программирование — это область математики, разрабатывающая теорию, численные методы решения многомерных задач с ограничениями. В отличие от классической математики, МП занимается математическими методами решения задач нахождения наилучших вариантов из всех возможных.

Хотя градуировочные характеристики могут описываться самыми разными функциями, считается вполне достаточным ограничиться случаем линейной зависимости. Это обусловлено тем, что разработчики методик чаще всего сводят градуировку к нахождению линейной зависимости. Другими словами, мы будем полагать, что величины $\mathbf{X}$ и $\mathbf{Y}$ связаны линейной зависимостью, а задачу градуировки анализатора можно записать в виде

$$\mathbf{Y} = \mathbf{X}\mathbf{b} + \mathbf{e}, \quad (1)$$

где $\mathbf{b}$ — вектор регрессионных коэффициентов, $\mathbf{e}$ — вектор, содержащий значения ошибки. Следовательно, градуировка спектрометра заключается в подборе таких регрессионных коэффициентов $\mathbf{b}$, чтобы квадрат ошибки f был минимален. Таким образом, целевую функцию для задачи оптимизации можно представить следующим образом:

$$\mathbf{f} = (\mathbf{Y} - \mathbf{X}\mathbf{b})^T (\mathbf{Y} - \mathbf{X}\mathbf{b}), \quad (2)$$

где $\mathbf{f}$ — вектор $(f_1, f_2, \dots, f_n)$, содержащий квадраты указанной выше ошибки $\mathbf{e}$. Если ввести ограничения на область поиска экстремальных значений целевой функции, коэффициенты регрессии $\mathbf{b}$ должны находиться внутри пространства гиперкуба, где гиперкуб — это фигура в евклидовом пространстве E_n (n — размерность пространства). Например, если $n = 3$, то гиперкуб — это куб в трехмерном евклидовом пространстве, имеющий длину ребра, равную d . Целевую функцию (2) можно представить в виде задачи квадратичного программирования:

$$\mathbf{f} = (\mathbf{Y} - \mathbf{X}\mathbf{b})^T (\mathbf{Y} - \mathbf{X}\mathbf{b}) \xrightarrow{0 \leq x \leq d} \min. \quad (3)$$

Для удобства записи введем следующие переменные:

$$\begin{aligned} \mathbf{C} &= 2(\mathbf{X}^T \cdot \mathbf{X}); \\ \mathbf{D} &= -2(\mathbf{X}^T \cdot \mathbf{Y}). \end{aligned} \quad (4)$$

Тогда задача (3) может быть представлена в виде классической задачи квадратичного программирования:

$$\mathbf{f}(x) = \mathbf{D}^T \mathbf{b} + \mathbf{b}^T \mathbf{C} \mathbf{b} \xrightarrow{0 \leq b \leq d} \min, \quad (5)$$

где $\mathbf{b}$ — искомый вектор $(b_1, b_2, \dots, b_n)$, вектор $\mathbf{d}$ $(d_1, d_2, \dots, d_n)$, представляет указанное выше пространство ограничений. Оптимальное значение параметра $\mathbf{d}$ определяется с помощью процедуры валидации, что подразумевает конечную проверку созданной модели и изменение размерности гиперкуба до тех пор, пока критерии оценки качества градуировочной модели не будут удовлетворять необходимым условиям.

Данную задачу можно решить с помощью искусственного базиса и симплекс-метода. Однако у такого подхода есть ряд недостатков, например, если $\mathbf{C}$ будет положительно полуопределенной матрицей, то данный метод будет расходиться. Существует более простой и эффективный метод решения подобной задачи, называемый методом решения задачи о дополнительнойности. Для того чтобы представить задачу квадратичного программирования в виде задачи о дополнительнойности

сти, существует специальная методика, подразумевающая выполнения следующих операций [7].

Для начала записывается функция Лагранжа для данной задачи квадратичного программирования (5):

$$L(\mathbf{u}, \mathbf{v}) = \mathbf{f}(\mathbf{x}) - \mathbf{u}^T \mathbf{b} - \mathbf{v}^T (\mathbf{d} - \mathbf{b}). \quad (6)$$

На основании (6) можно записать условия Куна—Таккера для данной задачи квадратичного программирования (5):

$$\begin{aligned} \nabla \mathbf{f} - \mathbf{u} - \mathbf{v} &= 0; \\ \mathbf{u}^T \mathbf{b} &= 0; \quad \mathbf{v}^T (\mathbf{d} - \mathbf{b}) = 0; \\ \mathbf{d} - \mathbf{b} &\geq 0; \\ \mathbf{u}, \mathbf{v} &\geq 0. \end{aligned} \quad (7)$$

Затем, для того чтобы заменить ограничение-неравенство ограничением-равенством, предполагается введение дополнительного вектора $\mathbf{S}$. Причем $\mathbf{d} - \mathbf{b} - \mathbf{S} = 0$, и благодаря этой подстановке условия Куна—Таккера (7) представляются следующим образом:

$$\begin{aligned} \nabla \mathbf{f} - \mathbf{u} - \mathbf{v} &= 0; \\ \mathbf{u}^T \mathbf{b} &= 0; \quad \mathbf{v}^T \mathbf{S} = 0; \\ \mathbf{S} &\geq 0; \quad \mathbf{u}, \mathbf{v} \geq 0. \end{aligned} \quad (8)$$

На последнем шаге для системы, записанной в (8), делаются следующие замены:

$$\begin{aligned} \mathbf{M} &= \begin{pmatrix} 2Q & -1 \\ 0 & 1 \end{pmatrix}; \quad \mathbf{q} = \begin{pmatrix} \mathbf{D}^T \\ -\mathbf{d} \end{pmatrix}; \quad \mathbf{w} = \begin{pmatrix} \mathbf{u} \\ \mathbf{S} \end{pmatrix}; \\ \mathbf{z} &= \begin{pmatrix} \mathbf{b} \\ \mathbf{y} \end{pmatrix}. \end{aligned} \quad (9)$$

Задача (5) принимает вид эквивалентной задачи о дополнительности:

$$\begin{aligned} \mathbf{w} &= \mathbf{Mz} + \mathbf{q}; \\ \mathbf{w} &\geq 0; \quad \mathbf{z} \geq 0; \quad \mathbf{w}^T \mathbf{z} = 0. \end{aligned} \quad (10)$$

Решив данную систему, получим вектор коэффициентов $\mathbf{z}$, который и будет оптимальным вектором регрессионных коэффициентов $\mathbf{b}$ для целевой функции задачи построения MVC-анализатора, записанной в уравнении (2).

Таким образом, при исследовании образца с неизвестной концентрацией интересующего компонента y производится регистрация спектра X_0 на анализаторе, а затем с помощью выражения

$$\hat{y} = \mathbf{X}_0^T \cdot \mathbf{b} \quad (11)$$

можно оценить значение y .

Критерии для оценки качества градуировочной модели

Для оценки качества градуировочной зависимости была использована статистическая характеристика, широко применяющаяся в зарубежной практике, — SEV (Standard Error of Validation), стандартная ошибка валидации [8].

SEV характеризует ошибку отклонения между референтными и предсказанными по градуировочному уравнению значениями для образцов дополнительного набора (т. е. серии образцов, не входивших в градуировочный набор и используемых для проверки градуировки):

$$SEV = \sqrt{\sum_{i=1}^{d_v} (\hat{y}_i - y_i)^2 / d_v}, \quad (12)$$

d_v — число образцов дополнительного набора; y_i — референтное значение анализируемого компонента для i -го образца дополнительного набора; $\hat{y}_i$ — предсказанное по градуировочному уравнению значение анализируемого компонента для i -го образца дополнительного набора.

ЭКСПЕРИМЕНТАЛЬНАЯ ЧАСТЬ

В наших экспериментах мы сравнивали метод создания MVC с помощью техники МП с известными классическими методами — PLS и PCR.

Наборы данных и оборудование для создания многомерной градуировочной модели

Предлагаемый способ создания MVC для градуировки БИК-анализаторов оценивался на трех различных наборах данных. Для измерения спектральных данных был выбран БИК-анализатор "ИНФРАЛЮМ ФТ-10" — современный анализатор, выпускаемый компанией "Люмэкс". Спектры образцов измерялись в диапазоне волновых чисел 8 000–14 000 см^{-1} с разрешением 16 см^{-1} .

Первый набор — образцы зерен кукурузы с известными референтными данными процентного содержания жира. Градуировочный набор содержит 67 образцов, валидационный набор — 10 образцов.

Второй набор — образцы шрота подсолнечника с известными референтными данными процентного содержания протеина. Градуировочный набор содержит 74 образца, валидационный набор — 10 образцов.

Третий набор — образцы комбикормов с известными референтными данными процентного содержания кальция. Градуировочный набор содержит 53 образца, валидационный набор — 10 образцов.

Результаты приведены в табл. 1–3.

Табл. 1. Результаты предсказания массовой доли жира в зернах кукурузы для образцов валидационного набора по градуировочным моделям, построенным различными способами

Образец	Реф. значение	Метод построения многомерной градуировочной модели		
		PCR	PLS	МП
kuk-0350	5.04	4.59	4.42	4.85
kuk-0327	7.99	6.60	7.05	6.75
kuk-0259	6.79	5.92	6.16	6.59
kuk-0288	7.70	7.32	7.25	6.88
kuk-0322	3.41	4.74	4.59	4.07
kuk-0232	6.29	6.30	6.63	6.18
kuk-0248	4.50	5.21	4.79	4.42
kuk-0375	4.28	5.31	4.84	4.30
kuk-0307	3.27	5.42	5.07	3.32
kuk-0246	5.70	5.20	5.05	5.32
Значение SEV		1.19	0.99	0.62

Табл. 2. Результаты предсказания массовой доли протеина в шроте подсолнечника для образцов валидационного набора по градуировочным моделям, построенным различными способами

Образец	Реф. значение	Метод построения многомерной градуировочной модели		
		PCR	PLS	МП
chm-1141	43.99	40.51	43.02	43.54
chm-1152	44.04	42.45	43.04	43.43
chm-1145	42.58	41.03	42.82	42.77
chm-1266	38.90	38.75	38.78	38.44
chm-1133	40.26	39.02	40.22	40.44
chm-1280	41.80	38.92	40.22	41.01
chm-1118	36.79	37.18	37.80	37.57
chm-1119	37.09	38.41	39.28	38.46
chm-1219	35.57	37.69	36.20	35.11
chm-1126	39.97	39.19	40.19	40.38
Значение SEV		1.99	1.25	0.91

Табл. 3. Результаты предсказания массовой доли кальция в комбикормах для образцов валидационного набора по градуировочным моделям, построенным различными способами

Образец	Реф. значение	Метод построения многомерной градуировочной модели			Образец	Реф. значение	Метод построения многомерной градуировочной модели		
		PCR	PLS	МП			PCR	PLS	МП
pk-0050	2.59	2.33	2.64	2.75	pk-0073	2.11	1.95	2.59	2.73
pk-0808	3.10	2.27	1.95	2.84	pk-0833	1.34	2.32	0.57	0.78
pk-0001	3.55	2.09	2.44	2.98	pk-0026	2.86	2.17	2.58	2.98
pk-0819	1.04	2.51	2.44	1.35	pk-0069	2.36	2.58	2.83	2.33
pk-0034	1.95	2.08	2.41	2.02	Значение SEV		0.86	0.87	0.51
pk-0043	3.61	2.96	2.82	3.22					

ВЫВОДЫ

Полученные экспериментальные данные показывают, что при градуировке БИК-анализаторов методами МП для определения содержания жира

в зернах кукурузы, протеина в шроте подсолнечника, кальция в комбикормах значения стандартной ошибки валидации (SEV) меньше аналогичных значений, полученных при градуировке альтернативными методами.

Это позволяет сделать вывод о том, что техника математического программирования может применяться в качестве метода для поиска регрессионных коэффициентов в задачах построения МВС, и в частности для градуировки БИК-анализаторов. Методы математического программирования в определенных случаях позволяют создавать градуировки, обеспечивающие более высокую точность, чем градуировки, полученные с помощью классических методов.

СПИСОК ЛИТЕРАТУРЫ

1. Крищенко В.П. Ближняя инфракрасная спектроскопия. М.: Интерагротех, 1997. 638 с.
2. Wang Y., Veltkamp D., Kowalski B.R. Multivariate instrument standardization // *Analytical chemistry*. 1991. Vol. 63, nu. 23. P. 2750–2756.
3. Esbensen K.H. Multivariate analysis in practice. Oslo: Camo, 2000. 597 p.
4. Estienne F, Massart D.L. Multivariate calibration with Raman data using fast principal components and partial least square method // *Analytical Chemistry Acta*. 2001. Vol. 450, nu. 1-2. P. 123–129.

5. Bishop C.M. Neural networks for pattern recognition. Oxford: Clarendon Press, 1995. 482 p.
6. Skutin I., Zharinov K., Sushkova M. Mathematical programming method as a means of calibration of NIR analyzers // *Journal of Chemistry & Chemical Engineering*. 2012. Vol. 6, nu. 8. P. 692–697.
7. Sinha S.M. Mathematical programming theory and methods. Elsevier Science, 2006. 628 p.
8. ASTM standard E 1655-00. Practices for infrared multivariate quantitative analysis. 2000. 28 p.

Санкт-Петербургский государственный технологический институт (Технический университет)

Контакты: Скутин Илья Владимирович,
Skytin.Ilya@gmail.com

Материал поступил в редакцию 30.09.2013

THE APPLICATION OF MATHEMATICAL PROGRAMMING METHODS FOR CALIBRATION OF NIR ANALYZERS

I. V. Skutin, K. A. Zharinov

Saint-Petersburg State Technological Institute (Technical University), RF

The possibility of improving the accuracy of the analysis of various samples carried out by means of an analyzer operating in the near infrared region is considered in this article. To solve the problem we suggest to create calibration using the mathematical programming methods.

Keywords: calibration, NIR analyzer, mathematical programming

REFERENCES

1. Wang Y., Veltkamp D., Kowalski B.R. Multivariate instrument standardization. *Analytical chemistry*, 1991, vol. 63, nu 23, pp. 2750–2756.
2. Esbensen K.H. Multivariate analysis in practice. 2000, Oslo, Camo, 597 p.
3. Estienne F, Massart D.L. Multivariate calibration with Raman data using fast principal components and partial least square method. *Analytical Chemistry Acta*. 2001, vol. 450, nu. 1-2, pp. 123–129.
4. Bishop C.M. Neural networks for pattern recognition. 1995, Oxford, Clarendon Press, 482 p.
5. Skutin I., Zharinov K., Sushkova M. Mathematical programming method as a means of calibration of NIR analyzers. *Journal of Chemistry & Chemical Engineering*.

ing. 2012, vol. 6, nu. 8, pp. 692–697.

6. Sinha S.M. Mathematical programming theory and methods. 2006, Elsevier Science, 628 p.
7. ASTM standard E 1655-00. Practices for infrared multivariate quantitative analysis. 2000, 28 p.

Contacts: Skutin Ilya Vladimirovich,
Skytin.Ilya@gmail.com

Article arrived in edition: 30.09.2013