УДК 518.852

***С.И. Носков*[1]**

[1]*Иркутский государственный университет путей сообщения, г. Иркутск, Российская Федерация*

О КЛАСТЕРИЗАЦИИ ДАННЫХ НА ОСНОВЕ СВОЙСТВ МЕТОДОВ ИДЕНТИФИКАЦИИ ПАРАМЕТРОВ ЛИНЕЙНОЙ РЕГРЕССИИ

Аннотация. В работе предложен способ кластеризации выборки данных при построении линейной регрессионной модели, основанный на свойствах методов наименьших модулей и антиробастного оценивания параметров. Первый из них обеспечивает равенство нулю числа ошибок аппроксимации, не меньшего числа параметров модели, а для второго число максимальных по модулю ошибок не меньше, чем число параметров плюс единица.

Ключевые слова: кластеризация выборки, регрессионная модель, методы идентификации параметров, методы наименьших модулей и антиробастного оценивания.

***S.I. Noskov*[1]**

[1]*Irkutsk state University of railway engineering, Russian Federation.*

ON DATA CLUSTERING BASED ON PROPERTIES OF METHODS FOR IDENTIFICATION OF LINEAR REGRESSION PARAMETERS

Annotation. The paper proposes a method for clustering a data sample when constructing a linear regression model, based on the properties of the methods of least absolute deviation and anti-robust parameter estimation. The first of them ensures that the number of approximation errors equals zero, which is not less than the number of model parameters, and for the second, the number of maximum errors in absolute value is not less than the number of parameters plus one.

Keywords: sample clustering, regression model, parameter identification methods, least modulus and anti-robust estimation methods.

Рассмотрим линейное регрессионное уравнение (модель)

$$y_k = \sum_{i=1}^{m} \alpha_i x_{ki} + \varepsilon_k,\ k = \overline{1,n}, \quad (1)$$

где y − зависимая, а x_i − i-ая независимая переменные, α_i − i-ый подлежащий оцениванию параметр, ε_k − ошибки аппроксимации, k − номер наблюдения, n − длина выборки данных. Будем считать все переменные и ошибки уравнения (1) детерминированными.

Представим уравнение (1) в векторной форме:

$$y = X\alpha + \varepsilon,$$

где $y = (y_1, \dots, y_n)^T$, $\alpha = (\alpha_1, \dots, \alpha_m)^T$, $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$, X− $(n \times m)-$ матрица с компонентами x_{ki}. При вхождении в уравнение (1) свободного члена первый столбец матрицы X будет состоять из единиц.

Одной из традиционных задач регрессионного анализа является кластеризация данных, состоящая в разбиении исходной выборки данных (X,y) на непересекающиеся подвыборки, или, что то же, раздиление множества номеров наблюдений P={1, 2, …, n} на p подмножеств $\mathrm{P}_i, i = \overline{1,\mathrm{p}}$:

$$\mathrm{P} = \bigcup_{i=1}^{\mathrm{p}} P_i,\ P_i \cap P_j = \emptyset,\ i \neq j.$$

Так, в работе [1] кластеризацию предлагается производить на основе метода k-means. В [2] представлен краткий обзор публикаций по алгоритмам кластеризации. В частности, рассмотрены алгоритмы k-means, dbscan и иерархические агломеративные алгоритмы, в которых для вычисления межкластерного расстояния используются методы ближайшего соседа, полной и средней связи и метод Уорда. В [3] описан алгоритм динамической кластеризации, позволяющий решать задачи классификации объектов на массивах больших данных при значительной размерности пространства классифицируемых признаков. Алгоритм позволяет в режиме реального времени отслеживать эффекты, производимые

изменениями обучающей выборки, и может быть использован для обработки и классификации больших данных в условиях потока. В [4] рассмотрен модифицированный алгоритм Хамелеон, предназначенный для работы с линейно неразделимыми зашумленными данными различных объемов.

В работе [5] кластеризация выборки производится с использованием метода наименьших модулей (МНМ), а в [6] – антиробастного оценивания (МАО) параметров модели (1).

Разбиение выборки на три подвыборки (т.е. для p=3) может быть осуществлено на основе свойств МНМ и МАО. Так, в [7] доказано, что если rank $X = m$, то при использовании МНМ для оценивания параметров модели (1) выполняется неравенство:

$$|P_1| \geq m, \tag{2}$$

где

$$P_1 = \{s \in P | \ \varepsilon_s = 0\},$$

а символом $|P_1|$ обозначено число элементов в множестве P_1.
При применении же МАО справедливо неравенство [8]:

$$|P_2| \geq m + 1, \tag{3}$$

где

$$P_2 = \{s \in P | \ |\varepsilon_s| = \max_{k=\overline{1,n}} |\varepsilon_k|\}.$$

Таким образом, множества P_1 и P_2 содержат номера наблюдений, характеризующихся противоположным по смыслу содержанием – в первое входят номера с нулевыми ошибками аппроксимации, во второе – с максимальными по модулю ошибками.

Сформируем третье множество P_3 номеров наблюдений, не вошедших в P_1 и P_2:

$$P_3 = P \setminus (P_1 \cup P_2).$$

Для реальных выборок вследствие того, что множества P_1 и P_2 сформированы с использованием разных методов идентификации параметров, может оказаться, что

$$P_1 \cap P_2 \neq \emptyset.$$

В этом случае номера наблюдений, вошедшие в P_1 и P_2 одновременно, следует «переместить» в множество P_3.

Разумеется, для использования описанного выше способа кластеризации выборки на три подвыборки требуется, чтобы ее длина была приемлемой, а именно, должно выполняться необходимое (но не достаточное) условие:

$$n > 2\,m + 1.$$

Оно не является обременительным, поскольку для корректного применения методов регрессионного анализа должно выполняться условие:

$$n >> m.$$

В классической монографии [9] оно конкретизируется следующим образом:

$$n \geq 4m.$$

Предложенный способ кластеризации выборки данных может быть эффективно применен для решения задач, связанных с уточнением оценок параметров уравнения (1).

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Кротов Л.Н., Кротова Е.Л., Рукшин С.А. Математическое моделирование для кластеризации данных на основе обучающей выборки // Международный научно-исследовательский журнал. -2016. - № 5-3 (47). - С. 116-118.

2. Беликова М.Ю., Каранина С.Ю., Глебова А.В. Экспериментальное сравнение алгоритмов кластеризации в задаче группировки данных о грозовых разрядах // Кибернетика и программирование. - 2018. - № 1. - С. 15-2-6.

3. Печеный Е.А., Нуриев Н.К., Старыгина С.Д. Динамическая кластеризация потока больших данных // Математические методы в технике и технологиях - ММТТ. - 2019. - Т. 3. - С. 19-21.

4. Ляховец А.В. Исследование динамической кластеризации линейно неразделимых зашумленных данных различных объемом с помощью модифицированного алгоритма Хамелеон // Международный научно-исследовательский журнал. - 2013. - № 10-2 (17). - С. 55-61.

5. Носков С. И., Ильюшонок Д. М. Подход к кластеризации выборки данных на основе метода наименьших модулей // Южно-Сибирский научный вестник. - 2020. - № 6. - С. 255-259.

6. Носков С.И. Применение метода антиробастного оценивания параметров для кластеризации выборки данных // Вестник кибернетики. - 2021. - № 3 (43). - С. 46-50.

7. Лакеев А.В., Носков С.И. Метод наименьших модулей для линейной регрессии: число нулевых ошибок аппроксимации // Современные технологии. Системный анализ. Моделирование. - 2012. - № 2. - С. 48-50.

8. Носков С.И. Метод антиробастного оценивания параметров линейной регрессии: число максимальных по модулю ошибок аппроксимации // Южно-Сибирский научный вестник. - 2020. - № 1 (29). - С. 51-54.

9. Дрейпер Н, Смит Г. Прикладной регрессионный анализ, 3-е изд. – Вильямс. - 2016. - 912 с.

Информация об авторе

Сергей Иванович Носков – д.т.н., профессор, профессор кафедры «Информационные системы и защита информации», Иркутский государственный университет путей сообщения, г. Иркутск, e-mail: sergey.noskov.57@mail.ru

Author

Sergey Ivanovich Noskov – Doctor of Technical Science, Professor, the Subdepartment Information systems and information security, Irkutsk State Transport University, Irkutsk, e-mail: sergey.noskov.57@mail.ru

Для цитирования

Носков С.И. О кластеризации данных на основе свойств методов идентификации параметров линейной регрессии // «Информационные технологии и математическое моделирование в управлении сложными системами»: электрон. науч. журн. – 2022. – №4(16). – С. 82-85 – DOI: 10.26731/2658-3704.2022.4(16).82-85 – Режим доступа: http://ismm-irgups.ru/toma/416-2022, свободный. – Загл. с экрана. – Яз. рус., англ. (дата обращения: 23.12.2022).

For citation

Noskov S.I. O klasterizacii dannykh na osnove svojstv metodov identifikacii parametrov linejnoj regressii [On data clustering based on properties of methods for identification of linear regression parameters] // Informacionnye tehnologii i matematicheskoe modelirovanie v upravlenii slozhnymi sistemami: ehlektronnyj nauchnyj zhurnal [Information technology and mathematical

85 modeling in the management of complex systems: electronic scientific journal], 2022. No. 4(16). P. 82-85. DOI: 10.26731/2658-3704.2022.4(16).82-85 [Accessed 23/12/22].