УДК 518.852

С.И. Носков[1], С.В. Беляев[1]

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

СПОСОБ КЛАСТЕРИЗАЦИИ ВЫБОРКИ ДАННЫХ НА ОСНОВЕ ПРИМЕНЕНИЯ КРИТЕРИЯ СОГЛАСОВАННОСТИ ПОВЕДЕНИЯ

Аннотация. В работе дан краткий обзор публикаций по кластеризации данных с помощью методов регрессионного анализа. Приведено краткое описание известного способа разбиения выборки данных на подвыборки на основе разделяющих регрессий, сводящегося к задаче минимизации сумм ошибок аппроксимации на всех этих подвыборках. Кроме того, рассмотрен способ решения задачи кластеризации с помощью обобщенного критерия согласованности поведения и его непрерывной формы. Решен численный иллюстративный пример.

Ключевые слова: выборка данных, разделяющая регрессионная модель, параметры, метод наименьших модулей, задача нелинейного программирования, критерий согласованности поведения, булевы переменные.

S.I. Noskov[1], S.V. Belyaev[1]

[1] *Irkutsk State Transport University, Irkutsk, Russian Federation*

METHOD OF CLUSTERING A DATA SAMPLE BASED ON THE APPLICATION OF THE BEHAVIORAL CONSISTENCY CRITERION

Abstract. The paper provides a brief overview of publications on data clustering using regression analysis methods. A brief description of a known method for dividing a data sample into subsamples based on separating regressions is given, which is reduced to the problem of minimizing the sums of approximation errors on all these subsamples. In addition, a method for solving a clustering problem using a generalized behavior consistency criterion and its continuous form is considered. A numerical illustrative example is solved.

Keywords: data sample, separating regression model, parameters, least absolute values method, nonlinear programming problem, behavior consistency criterion, Boolean variables.

Введение. При обработке различных типов данных (пространственных данных (cross section data), временных рядов (time series), панельных данных (panel data)) средствами регрессионного анализа часто возникает необходимость в их кластеризации, состоящей в разбиении исходной выборки данных на непересекающиеся подвыборки, называемые кластерами.

Так в работе [1] оцениваются важность учета уровня кластеризованности данных. При этом подчеркивается, что при построении линейной модели с применением метода наименьших квадратов отклонение от действительного значения зависимой переменной меняется при увеличении этого уровня. В [2] рассматриваются три подхода к построению моделей линейной регрессии с кластеризованными данными. Первые два подхода основаны на применении многоуровневых моделей со “случайным интерцептом”, вторая модель отличается от первой учетом средних значений по каждому кластеру. Третья модель основана на оценивание попарных разниц между двумя случайными элементами в кластере.

Применение различных подходов к кластеризации данных при построении линейной регрессии используются при создании алгоритмов машинного обучения. Так в статье [3] представлен алгоритм построения кластерной линейной регрессии (КЛР), который включает в себя 3 основных этапа: а) определение числа кластеров; б) определение границ кластеров; в) применение линейной регрессии ко всем кластерам. CLR показал в среднем более высокую точность по сравнению с другими алгоритмами. В работе [4] рассматривается метод оценки электропотребления в зависимости от различных

параметров с целью его снижения. Исходная информация для анализа содержалась в профилях нагрузки клиентов. Профили нагрузки были разбиты на кластеры при помощи алгоритма k-средних, затем полученные кластеры применялись при построении регрессионных моделей. Были построены три регрессионные модели: а) модель температурной чувствительности профиля нагрузки; б) кластерная регрессионная модель; б) кластерно-температурная регрессионная модель. В [5] рассматривается подход к кластеризации потоковых данных. Линейная регрессия применяется при определении возможности объединения двух кластеров. В исследовании [6] приведено сравнение логистической регрессии и логистической регрессии с случайным интерцептом (константой) для кластеризованных данных. Последняя показала большую определенность, если кластеризация применялась для определения риска.

В работе [7] предлагается расширение модели кластерной линейной регрессии. Рассматривается алгоритм объедения нескольких функциональных предикторов на основе метода главных компонент. Для определения функциональных коэффициентов разработан эффективный алгоритм с использованием методов k-средних и М-оценок. Различные методы построения кластерной линейной регрессии описываются в [8]. Представлены следующие виды моделей и алгоритмов: смешанное целочисленное нелинейное программирование и смешанное логическое квадратичное программирование; модели, основанные на комбинировании кластеризации и смешанного-целочисленного линейного программирования; модели нелинейного программирования и модели, основанные на нечеткой логике; модель негладкой оптимизации; модели многокритериальной оптимизации; смесевые модели; кластерно-взвешанные модели. Также описываются алгоритмы, основанные на приведенных моделях, и результаты их сравнения.

В [9, 10] подробно описан кластерный анализ – определены его процедуры, алгоритмы кластеризации и их общая классификация. Приведено сравнение алгоритмов кластеризации на тестовых данных. Также в [10] приведен пример использования методов кластеризации при анализе производственной деятельности и классификации предприятий.

В [11] приведен подход к построению линейной регрессии, основанный на порождении бинарных признаков и кластеризации. Переход к бинарному признаку осуществляется с использованием нечеткой классификации. Показано, что совместное использование исходных действительных и порожденных бинарных признаков позволило существенно повысить качество прогноза по линейной регрессии.

Кластерный анализ применяется в различных практических сферах. Так, в [12] приведено применение кластерного подхода для решения ряда экономических задач. В работе [13] кластерный анализ применяется при изучении деятельности судостроительных и судоремонтных предприятий. Произведена кластеризация по показателям пропускной способности, а также по экономическим факторам. На основе одного из кластеров построена двухфакторная линейная модель. В [14] проведено исследование уровня информатизации общества в регионах. Определены два кластера: регионы, лидирующие по вопросам информатизации и регионы, условно отстающие в информатизации экономической и социальной сфер. Для каждого кластера построены линейные регрессионные модели, произведена их оценка. В соответствии с заданными показателями произведена группировка регионов в рамках кластеров.

Основная часть. Пусть задана выборка данных (X, y) длины n, где $X-$ $(n \times m)-$ матрица значений независимых переменных с компонентами x_{ki}, $k = \overline{1,n}$, $i = \overline{1,m}$, $y = (\mathrm{y}_1, \ldots, \mathrm{y}_\mathrm{n})^\mathrm{T}$ – вектор значений зависимой переменной. Предположим, что характер влияния независимых переменных x_i, $i = \overline{1,m}$ на зависимую переменную y может меняться на различных r участках выборки. В этом случае имеет смысл разбить, разделить ее (подвергнуть кластеризации) на непересекающиеся подвыборки (X^j, y^j), $j = \overline{1,r}$, где в

матрицы X^j и векторы y^j входят соответственно строки и компоненты с номерами из индексных множеств $P^j \subset \{1,2,\ldots, n\}$. При этом выполняются естественные условия:

$$\cup_{j=1}^{r} P^j = \{1,2,\ldots,n\},\ P^i \cap P^j = \emptyset,\ i \neq j.$$

Заметим, что в работах [15-17] предлагаются способы кластеризации выборки, основанные на свойствах методов оценивания параметров регрессионных моделей.

Для разбиения выборки будем использовать т.н. разделяющие регрессии (т.е. регрессии, разделяющие исходную выборку на подвыборки)

$$y_k = F(\alpha^j; x_{k1}, x_{k2},\ldots, x_{km}) + \varepsilon_k^j,\ j = \overline{1,r},\ k \in P^j. \quad (1)$$

Здесь F– аппроксимирующая вещественная функция, α^j – векторы оценок параметров. Заметим, что при линейности F регрессии (1) называют также (см., например, [8]) частными кластерными линейными регрессиями, переключающимися или типологическими регрессиями.

В качестве функции F могут быть, в частности, использованы модельные конструкции, содержащие различные преобразования независимых переменных, как это сделано, например, в [18].

Постановка задачи кластеризации выборки данных с помощью разделяющих регрессий (1), как и в методе КЛР [8], формализуется следующим образом:

$$G(\alpha^1, \alpha^2,\ldots, \alpha^r, P^1, P^2,\ldots, P^r) = \sum_{j=1}^{r} \sum_{k \in P^j} |\varepsilon_k^j|^{\nu} \to min, \quad (2)$$

где фиксированное значение показателя степени $\nu \geq 1$, как и при использовании L_ν-оценок [19], задает способ расчета расстояния (метрики) между фактическими и вычисленными значениями зависимой переменной.

Эту задачу можно поставить несколько по-иному, с использованием обобщенного критерия согласованности поведения (КСП) K^j [20]. Обозначим через $\hat{y}_k^j$ расчетные значения зависимой переменной в (1):

$$\hat{y}_k^j = F(\alpha^j; x_{k1}, x_{k2},\ldots, x_{km}).$$

Тогда K^j имеет вид:

$$K^j = \sum_{\substack{k>s \\ k,s \in P^j}} sign[(y_k - y_s)(\hat{y}_k^j - \hat{y}_s^j)],$$

где

$$sign[a] = \begin{cases} 1, & a \geq 0 \\ 0, & a < 0. \end{cases}$$

С учетом КСП задача кластеризации выборки примет вид:

$$K(\alpha^1, \alpha^2,\ldots, \alpha^r, P^1, P^2,\ldots, P^r) = \sum_{j=1}^{r} K^j \to max. \quad (3)$$

Поскольку КСП имеет дискретный характер, решение задачи (3) может быть неединственным, особенно для относительно коротких выборок. Поэтому иногда вместо КСП имеет смысл использовать его непрерывный аналог НКСП [21]:

$$\widetilde{K}^j = \sum_{\substack{k>s \\ k,s \in P^j}} l_{ks}^j,$$

где

$$l_{ks}^j = \begin{cases} |\hat{y}_k^j - \hat{y}_s^j|, & (y_k - y_s)(\hat{y}_k^j - \hat{y}_s^j) < 0 \\ 0, & (y_k - y_s)(\hat{y}_k^j - \hat{y}_s^j) \geq 0. \end{cases}$$

В этом случае вместо задачи (3) следует решить задачу:

$$\widetilde{K}(\alpha^1, \alpha^2,\ldots, \alpha^r, P^1, P^2,\ldots, P^r) = \sum_{j=1}^{r} \widetilde{K}^j \to min. \quad (4)$$

Таким образом, в результате решения задач (2), (3) или (4) должны быть как сформированы составы индексных множеств $P^j, j = \overline{1,r}$, так и определены векторы оценок параметров $\alpha^j, j = \overline{1,r}$ разделяющих регрессий (1). Эти задачи имеют переборный характер и распадаются на серию задач нелинейного программирования. При этом механизм перебора может быть формализован введением булевых переменных, как это сделано, например, в работах [8, 22, 23].

Рассмотрим иллюстративный пример. Пусть задана выборка данных длины n = 6:

Таблица 1.

Исходные данные

x_1	x_2	y
2	1	9
7	3	4
8	9	5
6	7	2
9	2	1
3	5	7

Необходимо разбить ее на две подвыборки одинаковой длины (т.е. при r = 2, $|P^1|=|P^2|=3$) линейными разделяющими регрессиями без свободного члена:

$$y_k = \alpha_1^j x_{k1} + \alpha_2^j x_{k2} + \varepsilon_k^j,\ j = \overline{1,2},\ k \in P^j.$$

Число комбинаций составов индексных множеств P^1 и P^2 будет равно $C_6^3/2=10$, поскольку $P^2 = \{1,2,\ldots,n\} \setminus P^1$.

В результате решения задачи (2) при $\nu = 1$ получим две разделяющие линейные регрессии

$$y_k = 7.6x_{k1} - 6.2x_{k2} + \varepsilon_k^1,\ k \in P^1 = \{1,3,4\},$$
$$y_k = -0.23x_{k1} + 1.54x_{k2} + \varepsilon_k^2,\ k \in P^2 = \{2,5,6\},$$
$$G = 1.2.$$

При этом наихудший вариант разбиения следующий:

$$y_k = -4.33x_{k1} + 4.0x_{k2} + \varepsilon_k^1,\ k \in P^1 = \{1,4,6\},$$
$$y_k = -0.02x_{k1} + 0.57x_{k2} + \varepsilon_k^1,\ k \in P^2 = \{2,3,5\},$$
$$G = 10.07.$$

Заметим, что полученное решение совпадает с решением задачи (4), что является следствием короткой длины выборки.

Очевидно, что разделяющие регрессии (1) имеют и самостоятельное значение при анализе исследуемого объекта и решении различных прогнозных задач.

Отметим, что наряду с рассмотренной интерес вызывает также задача кластеризации интервальной выборки (см, в частности, [24]). Естественно, при ее решении необходимо использовать аппарат интервальной математики (см, например, [25]).

Заключение. В работе рассмотрен встречающийся в литературе способ разбиения (кластеризации) выборки данных с помощью кластерных (разделяющих) регрессий, приводящий к задаче минимизации сумм ошибок аппроксимации для всех выделенных подвыборок. Предложен также способ решения задачи кластеризации с помощью обобщенного критерия согласованности поведения и его непрерывного аналога. Решен численный иллюстративный пример.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Ntani, G., Inskip, H., Osmond, C. et al. Consequences of ignoring clustering in linear regression // BMC Med Res Methodol. – 2021. – No. 21. – P. 1-13.

2. Desai M, Begg MD. A comparison of regression approaches for analyzing clustered data // Am J Public Health. – 2008. – V. 98. – No. 8. – P. 1425-1429.

3. Bertan Ari, H.Altay Güvenir. Clustered linear regression // Knowledge-Based Systems. – 2002. – V. 15. – No. 3. – P. 169-175.

4. N. Yamaguchi, J. Han, G. Ghatikar, S. Kiliccote, M. A. Piette and H. Asano. Regression models for demand reduction based on cluster analysis of load profiles // 2009 IEEE PES/IAS Conference on Sustainable Alternative Energy (SAE). – 2009. – P. 1-7.

5. Motoyoshi, Masahiro & Miura, Takao & Shioya, Isamu. Clustering Stream Data by Regression Analysis / Australasian Workshop on Data Mining and Web Intelligence. – 2004. – V. 32. – P. 115-120.

6. Bouwmeester, W., Twisk, J.W., Kappen, T.H. et al. Prediction models for clustered data: comparison of a random intercept and standard regression model // BMC Med Res Methodol. – 2013. – No. 13. – P. 1-10.

7. Ting Li, Xinyuan Song, Yingying Zhang, Hongtu Zhu, Zhongyi Zhu. Clusterwise functional linear regression models // Computational Statistics & Data Analysis. – 2021. – Vol. 158. – P. 1-15.

8. Qiang Long, Adil Bagirov, Sona Taheri, Nargiz Sultanova, and Xue Wu. Methods and Applications of Clusterwise Linear Regression: A Survey and Comparison // ACM Trans. Knowl. Discov. Data. – 2023. – V. 17. – No. 3. – P. 1-54.

9. Jain A., Murty M., Flynn P. Data Clustering: A Review. // ACM Computing Surveys. – 1999. – V. 31. – no. 3. – P. 264-323.

10. Мандель И.Д. Кластерный анализ. – М: Финансы и статистика, 1988. – 176 с.

11. Таскин А.С., Миркес Е.М. Линейная регрессия с кластеризацией по признаку на данных с действительными величинами // Сибирский аэрокосмический журнал. – 2012. – №3 (43). – С. 71-76.

12. Марков Л.С. Теоретико-методологические основы кластерного подхода. – Новосибирск: ИЭОПП СО РАН, 2015. – 300 с.

13. Неслухов Д.С. Использование кластерного и регрессионного анализа в изучении экономической деятельности судостроительных и судоремонтных предприятий // Интернет-журнал «НАУКОВЕДЕНИЕ». – 2016. – Т. 8. – №4. – С. 1-11.

14. Ерофеев А.А. Регрессионное моделирование на кластерах как средство исследования региональной специфики закономерностей информатизации общества // Экономические науки. – 2010. – № 12 (73). – С. 357-367.

15. Носков С.И. О кластеризации данных на основе свойств методов идентификации параметров линейной регрессии // Информационные технологии и математическое моделирование в управлении сложными системами. – 2022. – № 4 (16). – С. 82-85.

16. 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.

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

18. Носков С.И., Протопопов В.А. Оценка уровня уязвимости объектов транспортной инфраструктуры: формализованный подход // Современные технологии. Системный анализ. Моделирование. – 2011. – №4 (32). – С. 241-244.

19. Демиденко Е.З. Линейная и нелинейная регрессии. – М.: Финансы и статистика, 1981. – 302 с.

20. Носков С.И. Обобщенный критерий согласованности поведения в регрессионном анализе // Информационные технологии и математическое моделирование в управлении сложными системами. – 2018. – № 1 (1). – С. 14-20.

21. Носков С.И. Применение непрерывного критерия согласованности поведения при построении регрессионных моделей // Известия Тульского государственного университета. Технические науки. – 2021. – № 6. – С. 74-78.

22. Носков С.И. Идентификация параметров кусочно-линейной функции риска // Транспортная инфраструктура Сибирского региона. – 2017. – Т. 1. – С. 417-421.

23. Носков С.И. Идентификация параметров комбинированной кусочно-линейной регрессионной модели // Вестник Югорского государственного университета. – 2022. – № 4 (67). – С. 115-119.

24. Носков С.И. Точечная характеризация множеств решений интервальных систем линейных алгебраических уравнений // Информационные технологии и математическое моделирование в управлении сложными системами. – 2018. – № 1 (1). – С. 8-13.

25. Kreinovich V., Lakeyev A.V., Noskov S.I. Approximate linear algebra is intractable // Linear Algebra and its Applications. – 1996. – Vol. 232. – № 1-3. – P. 45-54.

REFERENCES

1. Ntani, G., Inskip, H., Osmond, C. et al. Consequences of ignoring clustering in linear regression // BMC Med Res Methodol. – 2021. – No. 21. – P. 1-13.

2. Desai M, Begg MD. A comparison of regression approaches for analyzing clustered data // Am J Public Health. – 2008. – V. 98. – No. 8. – P. 1425-1429.

3. Bertan Ari, H.Altay Güvenir. Clustered linear regression // Knowledge-Based Systems. – 2002. – V. 15. – no. 3. – P. 169-175.

4. N. Yamaguchi, J. Han, G. Ghatikar, S. Kiliccote, M. A. Piette and H. Asano. Regression models for demand reduction based on cluster analysis of load profiles // 2009 IEEE PES/IAS Conference on Sustainable Alternative Energy (SAE). – 2009. – P. 1-7.

5. Motoyoshi, Masahiro & Miura, Takao & Shioya, Isamu. Clustering Stream Data by Regression Analysis / Australasian Workshop on Data Mining and Web Intelligence. – 2004. – V. 32. – P. 115-120.

6. Bouwmeester, W., Twisk, J.W., Kappen, T.H. et al. Prediction models for clustered data: comparison of a random intercept and standard regression model // BMC Med Res Methodol. – 2013. – No. 13. – P. 1-10.

7. Ting Li, Xinyuan Song, Yingying Zhang, Hongtu Zhu, Zhongyi Zhu. Clusterwise functional linear regression models // Computational Statistics & Data Analysis. – 2021. – Vol. 158. – P. 1-15.

8. Qiang Long, Adil Bagirov, Sona Taheri, Nargiz Sultanova, and Xue Wu. Methods and Applications of Clusterwise Linear Regression: A Survey and Comparison // ACM Trans. Knowl. Discov. Data. – 2023. – V. 17. – No. 3. – P. 1-54.

9. Jain A., Murty M., Flynn P. Data Clustering: A Review. // ACM Computing Surveys. – 1999. – V. 31. – No. 3. – P. 264-323.

10. Mandel I.D. Cluster analysis. – Moscow: Finance and statistics, 1988. – 176 p.

11. Taskin A.S., Mirkes E.M. Linear regression with clustering by the feature of data with real values // Siberian aerospace journal. – 2012. – No. 3 (43). – P. 71-76.

12. Markov L.S. Theoretical and methodological foundations of the cluster cluster. – Novosibirsk: IEIE SB RAS, 2015. – 300 p.

13. Neslukhov D.S. Use of cluster and regression analysis in the study of economic activities of shipbuilding and ship repair enterprises // Internet journal "NAUKOVEDENIE". – 2016. – V. 8. – No. 4. – P. 1-11.

14. Erofeev A. A. Regression modeling on clusters as a study of regional specifics of patterns of informatization of society // Economic sciences. – 2010. – No. 12 (73). – P. 357-367.

15. Noskov S. I. On data clustering based on methods for determining linear regression parameters // Information technologies and mathematical modeling in management work consistently. – 2022. – No. 4 (16). – P. 82-85.

16. Noskov S. I., Ilyushonok D. M. An approach to clustering a data sample based on the least absolute values method // South Siberian Scientific Bulletin. – 2020. – No. 6. – P. 255-259.

17. Noskov S.I. Application of the method of antirobust parameter measurement for clustering a data sample // Bulletin of Cybernetics. - 2021. - No. 3 (43). - P. 46-50.

18. Noskov S.I., Protopopov V.A. Vulnerability assessment of transport employment facilities: a formalized approach // Modern technologies. Systems analysis. Modeling. - 2011. - No. 4 (32). - P. 241-244.

19. Demidenko E.Z. Linear and nonlinear regression. - M. : Finance and Statistics, 1981. - 302 p.

20. Noskov S.I. Generalized criterion for behavior consistency in regression analysis // Information technology and mathematical modeling in sequential management. - 2018. - No. 1 (1). – P. 14-20.

21. Noskov S.I. Application of a continuous behavior consistency criterion in the construction of regression models // Bulletin of Tula State University. Technical sciences. – 2021. – No. 6. – P. 74-78.

22. Noskov S.I. Identification of parameters of a piecewise linear risk function // Transport infrastructure of the Siberian region. – 2017. – Vol. 1. – P. 417-421.

23. Noskov S.I. Identification of parameters of a combined piecewise linear regression model // Bulletin of Yugra State University. – 2022. – No. 4 (67). – P. 115-119.

24. Noskov S.I. Point characterization of multiple solutions of interval systems of linear algebraic methods // Information technologies and mathematical modeling in the control of mechanisms sequentially. - 2018. - No. 1 (1). - P. 8-13. 25. Kreynovich V., Lakeev A.V., Noskov S.I. Approximate linear algebra is intractable // Linear algebra and its applications. - 1996. - V. 232. - No. 1-3. - P. 45-54.

Информация об авторах

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

Сергей Вячеславович Беляев – магистрант кафедры «Информационные системы и защита информации», Иркутский государственный университет путей сообщения, г. Иркутск, e-mail: bsv2001@list.ru

Authors

Sergey Ivanovich Noskov – Doctor of Technical Sciences, Professor, Professor of the Department of Information Systems and Information Security, Irkutsk State Transport University, Irkutsk, e-mail: sergey.noskov.57@mail.ru

Sergey Vyacheslavovich Belyaev – Master's student of the Department of Information Systems and Information Security, Irkutsk State Transport University, Irkutsk, e-mail: bsv2001@list.ru

Для цитирования

Носков С.И., Беляев С.В. Способ кластеризации выборки данных на основе критерия согласованности поведения // «Информационные технологии и математическое моделирование в управлении сложными системами»: электрон. науч. журн. – С.1-7. – 2024. – №4. (дата обращения: 29.11.2024)

For citations

Noskov S.I., Belyaev S.V. A method for clustering a data sample based on a behavioral consistency criterion // Informacionnye tehnologii i matematicheskoe modelirovanie v upravlenii slozhnymi sistemami: ehlektronnyj nauchnyj zhurna [Information technology and mathematical modeling in the management of complex systems: electronic scientific journal]. – P.1-7. – 2024. – No. 4. (Accessed 29.11.2024)