УДК 518.852

С.И.Носков

Иркутский государственный университет путей сообщения, г. Иркутск, Российская Федерация

НЕКОТОРЫЕ ФОРМЫ КЛАСТЕРНЫХ КУСОЧНО-ЛИНЕЙНЫХ РЕГРЕССИЙ

Аннотация. В работе рассмотрены три формы кластерной регрессии: кластерная кусочно-линейная регрессионная функция Леонтьева, кластерная кусочно-линейная регрессионная функция риска, кластерная смешанная кусочно-линейная регрессия. Указано, что при определенных условиях задачи их построения могут быть сведены к задачам линейно-булева программирования.

Ключевые слова: кластерная регрессия, оценки параметров, L_v-оценки, нелинейное программирование, булевы и индикаторные переменные.

S.I. Noskov[1]

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

SOME FORMS OF CLUSTER PIECEWISE LINEAR REGRESSIONS

Abstract. The paper considers three forms of cluster regression: cluster piecewise linear Leontief regression function, cluster piecewise linear regression risk function, cluster mixed piecewise linear regression. It is indicated that under certain conditions the problems of their construction can be reduced to linear-Boolean programming problems.

Keywords: cluster regression, parameter estimates, L_v-estimates, nonlinear programming, Boolean and indicator variables.

Построение кластерной линейной регрессии (КЛР) — хорошо известный метод аппроксимации данных с использованием более чем одной линейной функции, основанный на сочетании методов кластеризации и множественной линейной регрессии (см., например, [1]). Формально он может быть представлен следующим образом [1, 2]. Пусть задана выборка данных (X,y) длины n, где $X-$ $(n \times m)-$ матрица значений независимых переменных с компонентами x_{ki}, $k=\overline{1,n}$, $i=\overline{1,m}$, $\mathrm{y}=(\mathrm{y}_1,\ldots,\mathrm{y}_\mathrm{n})^\mathrm{T}$ – вектор значений зависимой переменной. Предположим, что характер влияния независимых переменных x_i, $i=\overline{1,m}$ на зависимую переменную y меняется на различных r участках выборки (кластерах). В этом случае имеет смысл разделить ее (кластеризировать) на непересекающиеся подвыборки (X^j,y^j), $j=\overline{1,r}$, где в матрицы X^j и векторы y^j войдут соответственно строки и компоненты с номерами из индексных множеств $P^j \subset\{1,2,\ldots,n\}$. При этом должны выполняться обязательные условия:

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

Тогда кластерная линейная регрессия примет вид:

$$y_k=\alpha_0^j+\sum_{i=1}^{m}\alpha_i^j x_{ki}+\varepsilon_k^j,\ j=\overline{1,r},\ k \in P^j. \qquad (1)$$

Введем матрицу оценок параметров $A=||\alpha_i^j||$, $j=\overline{1,r}$, $i=\overline{0,m}$ и матрицу булевых индикаторных переменных $\Omega=||\omega_{kj}||$, $k=\overline{1,n}$, $j=\overline{1,r}$ следующим образом:

$$\omega_{kj}=\begin{cases} 1, & k \in P^j \\ 0, & \text{в противном случае} \end{cases}.$$

Тогда оценка параметров и формирование составов индексных множеств P^j, $j=\overline{1,r}$ кластерной линейной регрессии (1) осуществляется посредством решения следующей задачи в общем случае нелинейного программирования с булевыми переменными:

$$G(A,\Omega)=\sum_{k=1}^{n}\sum_{j=1}^{r}\omega_{kj}|\varepsilon_k^j|^{\nu}, \quad (2)$$

где фиксированное значение показателя степени $\nu \geq 1$, как и при использовании L_ν-оценок [3], определяет способ задания расстояния (метрики) между фактическими и расчетными значениями зависимой переменной.

Вопросам построения, исследования и применения КЛР посвящена весьма обширная литература. Можно, в частности, отметить работы [4-8], в которых описаны некоторые алгоритмы решения задач, связанных с построением КЛР. В ряде публикаций описано практическое применение КЛР: [9] (задача сегментации потребительских выгод), [10] (моделирование процесса сварки металлов в среде инертного газа), [11] (разработка системы управления дорожным покрытием), [12] (прогнозирование осадков) и др.

При кластеризации данных могут быть также использованы и различные кусочно-линейные регрессионные формы (см., в частности, [13, 14]):

- кластерная кусочно-линейная регрессионная функция Леонтьева

$$y_k=\alpha_0^j+min\{\alpha_1^j x_{k1},\alpha_2^j x_{k2},\dots,\alpha_m^j x_{km}\}+\varepsilon_k^j, j=\overline{1,r}, k\in P^j, \quad (3)$$

- кластерная кусочно-линейная регрессионная функция риска

$$y_k=\alpha_0^j+max\{\beta_1^j x_{k1},\beta_2^j x_{k2},\dots,\beta_m^j x_{km}\}+\varepsilon_k^j, j=\overline{1,r}, k\in P^j, \quad (4)$$

- кластерная смешанная кусочно-линейная регрессия

$$y_k=\alpha_0^j+min\{\alpha_1^j x_{k1},\alpha_2^j x_{k2},\dots,\alpha_m^j x_{km}\}+max\{\beta_1^j x_{k1},\beta_2^j x_{k2},\dots,\beta_m^j x_{km}\}+\varepsilon_k^j,$$
$$j=\overline{1,r}, k\in P^j. \quad (5)$$

Задача (2) по отношению к регрессии (5) примет вид:

$$G(A,B,\Omega)=\sum_{k=1}^{n}\sum_{j=1}^{r}\omega_{kj}|\varepsilon_k^j|^{\nu},$$

где $B=||\beta_i^j||$, $j=\overline{1,r}$, $i=\overline{1,m}$.

Нетрудно убедиться, что при определенных условиях задачи (2) построения кластерных кусочно-линейных моделей (3) – (5) могут быть сведены к задачам линейно-булева программирования.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. 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.

2. Носков С.И., Беляев С.В. Способ кластеризации выборки данных на основе применения критерия согласованности поведения // Информационные технологии и математическое моделирование в управлении сложными системами. – 2024. – № 4.

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

4. Kin-nam Lau, Pui-lam Leung, and Ka-kit Tse. A mathematical programming approach to clusterwise regression model and its extensions // European Journal of Operational Research. – 1999. V. 116. – No. 3. – P. 640–652.

5. Dimitris Bertsimas, and Romy Shioda. Classification and regression via integer optimization // Operations Research. – 2007. – V. 55. – No. 2. – P. 252–271.

6. Real A. Carbonneau, Gilles Caporossi, and Pierre Hansen. Extensions to the repetitive branch and bound algorithm for globally optimal clusterwise regression // Computers & operations research. – 2012. V. 39. – No. 11. – P. 2748–2762.

7. Wayne S. DeSarbo, Richard L. Oliver, and Arvind Rangaswamy. A simulated annealing methodology for clusterwise linear regression // Psychometrika. – 1989. – V. 54. – No. 4. – P. 707–736.

8. Adil M. Bagirov and Julien Ugon. Nonsmooth DC programming approach to clusterwise linear regression: optimality conditions and algorithms // Optimization Methods and Software. – 2018. – V. 33. – No. 1. – P. 194–219.

9. Michel Wedel and Cor Kistemaker. Consumer benefit segmentation using clusterwise linear regression // International Journal of Research in Marketing. – 1989. – V. 6. – No. 1. – P. 45–59.

10. Jagadeesh P. Ganjigatti, Dilip K. Pratihar, and A. Roy Choudhury. Global versus clusterwise regression analyses for prediction of bead geometry in MIG welding process // Journal of materials processing technology – 2007. – V. 189. – No. 1-3. – P. 352–366.

11. Mukesh Khadka and Alexander Paz. Comprehensive clusterwise linear regression for pavement management systems // Journal of Transportation Engineering, Part B: Pavements. – 2017. – V. 143. – No. 4. – P. 1-13.

12. Adil M. Bagirov, Arshad Mahmood, and Andrew Barton. Prediction of monthly rainfall in Victoria, Australia: clusterwise linear regression approach // Atmospheric Research. – 2017. – V. 188. – P. 20–29.

13. Носков С.И., Хоняков А.А. Программный комплекс построения некоторых типов кусочно-линейных регрессий // Информационные технологии и математическое моделирование в управлении сложными системами. – 2019. – № 3 (4). – С. 47-55.

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

REFERENCES

1. 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.

2. Noskov S.I., Belyaev S.V. Method of clustering a data sample based on the application of the behavior consistency criterion // Information technologies and mathematical modeling in the management of complex systems. – 2024. – No. 4.

3. Demidenko E.Z. Linear and nonlinear regressions. – Moscow: Finance and statistics, 1981. – 302 p.

4. Kin-nam Lau, Pui-lam Leung, and Ka-kit Tse. A mathematical programming approach to clusterwise regression model and its extensions // European Journal of Operational Research. – 1999. V. 116. – No. 3. – P. 640–652.

5. Dimitris Bertsimas, and Romy Shioda. Classification and regression via integer optimization // Operations Research. – 2007. – V. 55. – No. 2. – P. 252–271.

6. Real A. Carbonneau, Gilles Caporossi, and Pierre Hansen. Extensions to the repetitive branch and bound algorithm for globally optimal clusterwise regression // Computers & operations research. – 2012. V. 39. – No. 11. – P. 2748–2762.

7. Wayne S. DeSarbo, Richard L. Oliver, and Arvind Rangaswamy. A simulated annealing methodology for clusterwise linear regression // Psychometrika. – 1989. – V. 54. – No. 4. – P. 707–736.

8. Adil M. Bagirov and Julien Ugon. Nonsmooth DC programming approach to clusterwise linear regression: optimality conditions and algorithms // Optimization Methods and Software. – 2018. – V. 33. – No. 1. – P. 194–219.

9. Michel Wedel and Cor Kistemaker. Consumer benefit segmentation using clusterwise linear regression // International Journal of Research in Marketing. – 1989. – V. 6. – No. 1. – P. 45–59.

10. Jagadeesh P. Ganjigatti, Dilip K. Pratihar, and A. Roy Choudhury. Global versus clusterwise regression analyses for prediction of bead geometry in MIG welding process // Journal of materials processing technology – 2007. – V. 189. – No. 1-3. – P. 352–366.

11. Mukesh Khadka and Alexander Paz. Comprehensive clusterwise linear regression for pavement management systems // Journal of Transportation Engineering, Part B: Pavements. – 2017. – V. 143. – No. 4. – P. 1-13.

12. Adil M. Bagirov, Arshad Mahmood, and Andrew Barton. Prediction of monthly rainfall in Victoria, Australia: clusterwise linear regression approach // Atmospheric Research. – 2017. – V. 188. – P. 20–29.

13. Noskov S.I., Khonyakov A.A. Software package for constructing some types of piecewise linear regressions // Information technologies and mathematical modeling in the management of complex systems. – 2019. – No. 3 (4). – P. 47-55.

14. 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.

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

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

Для цитирования

Носков С.И. Некоторые формы кластерных кусочно-линейных регрессий // «Информационные технологии и математическое моделирование в управлении сложными системами»: электрон. науч. журн. – С. 41-44. – 2024. – №4. (дата обращения: 02.12.2024)

For citations

Noskov S.I. Some forms of cluster piecewise linear regressions // 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. 41-44. – 2024. – No. 4. (Accessed 02.12.2024)