УДК 330.4
DOI: 10.17277/vestnik.2025.01.pp.025-033

ГРУППИРОВКА ПЕРЕМЕННЫХ В РЕГРЕССИОННОЙ КОМБИНИРОВАННОЙ ФУНКЦИИ РИСКА

С. И. Носков[✉]**, Е. С. Попов, А. А. Бутин**

Кафедра «Информационные системы и защита информации», sergey.noskov.57@mail.ru; ФГБОУ ВО «Иркутский государственный университет путей сообщения», Иркутск, Россия

Ключевые слова: группировка; задача линейно-булева программирования; индексное множество; комбинированная регрессионная модель риска; мощность множества.

Аннотация: Дан краткий обзор публикаций, посвященных группировке переменных в регрессионных моделях. В частности, рассмотрены: задача выбора сгруппированных переменных для точного прогнозирования в регрессии; оценка в многомерной линейной модели с сильно коррелирующими переменными; новый метод решения проблемы выбора групповой переменной в модели пропорциональных рисков Кокса; проблемы выбора переменных в модели логистической регрессии; расширение группового Лассо на модели логистической регрессии; способ применения одномерных моделей отдельно для каждого уровня группирующей переменной; проблема выбора сгруппированных переменных в моделях Пуассона; модификация эластичной сети и ее адаптивного аналога, позволяющая учитывать при группировке малую и среднюю корреляцию между объясняющими переменными. Разработан алгоритмический способ идентификации параметров регрессионной комбинированной функции риска и группировки независимых переменных, входящих в состав ее линейной и кусочно-линейной частей. При этом число элементов в соответствующих индексных множествах считается известным. Показано, что если в качестве метода определения параметров применяется метод наименьших модулей, то эта задача сводится к задаче линейно-булева программирования. Рассмотрен численный пример.

Введение

Методы математического моделирования, в частности, регрессионного, давно и успешно применяются при исследовании сложных технических и социально-экономических систем. Одной из подлежащих решению задач при этом является правильная верификация разрабатываемых моделей, состоящая, в том числе, в подборе адекватных форм связи между задействованными переменными (факторами) и их группировке. Так, в работе [1] рассматривается задача выбора сгруппированных переменных для точного прогнозирования в регрессии. Такая проблема естественным образом возникает во многих практических ситуациях, наиболее важным и известным примером которых является задача многофакторного дисперсионного анализа. Вместо выбора факторов путем поэтапного обратного исключения внимание концентрируется на точности оценки в рамках расширения Лассо и алгоритма LARS (*англ.* Least Angle Regression). Предлагаются эффективные способы модификации этих методов и показывается, что это обеспечивает превосходную производительность по сравнению с традиционным методом пошагового обратного исключения в задачах выбора факторов. В работе [2] рассматривается оценка в многомерной линейной модели с сильно коррелирующими переменными. Предлагается сначала сгруппировать переменные в кластеры, а затем выполнить разреженную оценку, например Лассо, для представителей кластеров или групповое Лассо на основе структуры кластеров. Описывается новый восходящий алгоритм агломеративной кластеризации, основанный на канонических корреляциях, и показывается, что он находит оптимальное решение и является статистически последовательным. В статье [3] разрабатывается новый метод решения проблемы выбора групповой переменной в модели пропорциональных рисков Кокса. Данный метод не только эффективно удаляет незначимые группы, но и сохраняет гибкость выбора переменных внутри идентифицированных групп. Исследование [4] посвящено изучению проблемы выбора переменных в модели логистической регрессии. Предлагается новый метод выбора переменных, основанный на логистической эластичной сети. Доказывается, что он имеет эффект группировки, то есть сильно коррелированные предикторы имеют тенденцию входить в модель или выходить из нее вместе. Логистическая эластичная сеть особенно полезна, когда число предикторов намного превышает количество наблюдений. В работе [5] проводится расширение группового Лассо на модели логистической регрессии и описывается эффективный алгоритм, который особенно подходит для задач большой размерности и может применяться к обобщенным линейным моделям для решения соответствующей задачи выпуклой оптимизации. В публикации [6], используя идеи уменьшения размерности в регрессионных моделях, описывается способ применения одномерных моделей отдельно для каждого уровня группирующей переменной. Эти модели менее общие, чем обычно изучаемые модели взаимодействия, но они могут привести к очень простым результатам, а также к простым и полезным сводным графикам. В исследовании [7] рассматривается проблема выбора сгруппированных переменных в моделях Пуассона с нулевым расширением посредством регуляризации группового моста. Штрафные коэффициенты вычисляются с использованием аппроксимации наименьших квадратов и уточняются с помощью эффективного алгоритма группового спуска. В статье [8] предложена модель логистической регрессии с $l_{p,q}$-регуляризацией, которая могла бы быть успешно применена к задачам выбора переменных с разреженной групповой структурой. Работа [9] посвящена разработке модификации эластичной сети и ее адаптивного аналога, позволяющей учитывать при группировке малую и среднюю корреляцию между объясняющими переменными и одновременно обеспечивать согласованность выбора результирующих переменных.

Постановка задачи

Пусть на выборке данных длины n заданы значения зависимой y_k и независимых x_{ki} переменных, $k = \overline{1, n}$, $i = \overline{1, m}$.

В работе [10], в предположении, что все переменные детерминированы, сформулирована задача оценивания параметров комбинированной кусочно-линейной регрессии (модели)

$$y_k = \sum_{i \in I_1} \alpha_i x_{ki} + \min_{j \in I_2} \{\beta_j x_{kj}\} + \varepsilon_k, \quad k = \overline{1, n}. \tag{1}$$

Здесь ε_k, $k = \overline{1, n}$, – ошибки аппроксимации; I_1, I_2 – наперед заданные индексные множества такие, что

$$I_1 \bigcup I_2 = \{1, 2, ..., m\}.$$

При этом выполнение традиционно накладываемого при подобных построениях условия

$$I_1 \bigcap I_2 = \varnothing$$

не требуется, так как допускается включение некоторых (или всех) независимых переменных одновременно и в линейную, и в кусочно-линейную компоненты модели (1). Следовательно, в общем случае $|I_1| + |I_2| \geq m$, где $|C|$ – число элементов в множестве C, его мощность.

Идентификация параметров $\alpha_i, i \in I_1, \beta_j, \; j \in I_2$ в [10] проводится с применением метода наименьших модулей (**МНМ**) путем минимизации функции $J_1(\alpha,\beta)$

$$J_1(\alpha,\beta) = \sum_{k=1}^{n} \left| y_k - \sum_{i \in I_1} \alpha_i x_{k\,i} - \min_{j \in I_2} \left\{ \beta_j x_{k\,j} \right\} \right| \to \min. \tag{2}$$

Задача (2) сводится к следующей задаче линейно-булева программирования (**ЛБП**):

$$\sum_{i \in I_1} \alpha_i x_{k\,i} + z_k + u_k - v_k = y_k, \quad k = \overline{1, n}; \tag{3}$$

$$z_k \leq \beta_j x_{k\,j}, \quad k = \overline{1, n}, \quad j \in I_2; \tag{4}$$

$$\beta_j x_{k\,j} - z_k \leq \left(1 - \sigma_{k\,j}\right) M, \; k = \overline{1, n}, \quad j \in I_2; \tag{5}$$

$$\sum_{j \in I_2} \sigma_{k\,j} = 1, \;\; u_k \geq 0, v_k \geq 0, \quad k = \overline{1, n}; \tag{6}$$

$$\sigma_{k\,j} \in \{0, 1\}, \quad k = \overline{1, n}, \quad j \in I_2; \tag{7}$$

$$J_1(\alpha,\beta) = \sum_{k=1}^{n} (u_k + v_k) \to \min. \tag{8}$$

Здесь M – заранее выбранное большое положительное число.

Ранее одним из авторов (см., например, [11]) предложен способ идентификации параметров регрессионной функции риска

$$y_k = \max \left\{ \beta_1 x_{k\,1}, \beta_2 x_{k\,2}, ..., \beta_m x_{k\,m} \right\} + \varepsilon_k, \quad k = \overline{1, n}, \tag{9}$$

противоположной по своей содержательной интерпретации по отношению к кусочно-линейной составляющей модели (1). В качестве как независимых, так и зависимой переменных в модели (9) задействованы негативные для анализируемого объекта факторы, такие как убытки, загрязнения, риски, технические сбои, отказы оборудования, ущерб, техногенные аварии и т.д. При этом, в соответствии с (9), значение переменной y определяется максимальным значением одного из локальных негативных независимых факторов, а любое уменьшение значений других предикторов на данный исход не влияет. Параметры функции риска при использовании МНМ оцениваются с помощью решения задачи ЛБП.

Рассмотрим противоположную по смыслу по отношению к модели (1) комбинированную кусочно-линейную функцию риска

$$y_k = \sum_{i \in I_1} \alpha_i x_{k\,i} + \max_{j \in I_2} \{\beta_j x_{k\,j}\} + \varepsilon_k, \quad k = \overline{1, n}. \tag{10}$$

Поставим более общую, чем (2), задачу оценивания параметров модели (10)

$$J_2(\alpha, \beta, I_1, I_2) = \sum_{k=1}^{n} \left| y_k - \sum_{i \in I_1} \alpha_i x_{k\,i} - \max_{j \in I_2} \{\beta_j x_{k\,j}\} \right| \to \min. \tag{11}$$

Необходимо отметить, что в результате решения задачи (11) следует не только вычислить вектора оценок параметров α и β, но и сформировать состав индексных множеств I_1 и I_2 при наперед заданных их мощностях: $|I_1| = p_1$, $|I_2| = p_2$, то есть сгруппировать эти множества.

Сведение задачи группировки предикторов в регрессионной функции риска с линейной составляющей к задаче линейно-булева программирования

Решение задачи (11), как и задачи (2), также сводится к несколько расширенной, по сравнению с (3) – (8), задаче ЛБП с использованием соответствующих вычислительных приемов [10 – 12]:

$$\sum_{i=1}^{m} \alpha_i x_{k\,i} + z_k + u_k - v_k = y_k, \quad k = \overline{1, n}; \tag{12}$$

$$-\delta_i M_1 \le \alpha_i \le \delta_i M_1, \ \ i = \overline{1, m}; \tag{13}$$

$$\sum_{i=1}^{m} \delta_i = p_1; \tag{14}$$

$$z_k \ge \beta_j x_{k\,j}, \quad k = \overline{1, n}, \quad j = \overline{1,\ m};; \tag{15}$$

$$\beta_j x_{k\,j} - z_k \ge (\sigma_{k\,j} - 1) M_2, \ k = \overline{1, n}, \quad j = \overline{1,\ m}; \tag{16}$$

$$\sum_{j=1}^{m} \sigma_{k\,j} = 1, \ \ u_k \ge 0, \ \ v_k \ge 0, \quad k = \overline{1, n}; \tag{17}$$

$$\beta_j \le \gamma_j M_3, \quad j = \overline{1,\ m}; \tag{18}$$

$$\sum_{j=1}^{m} \gamma_j = p_2; \tag{19}$$

$$\sigma_{k\,i} \in \{0, 1\}, \quad k = \overline{1, n}, \ \ i = \overline{1,\ m}; \tag{20}$$

$$\delta_i \in \{0, 1\}, \quad i = \overline{1,\ m}; \tag{21}$$

$$\gamma_j \in \{0, 1\}, \quad j = \overline{1,\ m}; \tag{22}$$

$$J_2(\alpha, \beta, I_1, I_2) = \sum_{k=1}^{n} (u_k + v_k) \to \min. \tag{23}$$

Здесь M_1, M_2, M_3 – заранее выбранные большие положительные числа.

После решения задачи (12) – (23) состав множеств I_1 и I_2 формируется по следующим правилам:

$$\delta_i = 1 \Rightarrow i \in I_1,\ i = \overline{1,m};$$

$$\gamma_j = 1 \Rightarrow j \in I_2,\ j = \overline{1,m}.$$

Разумеется, предикторы с нулевыми значениями параметров α_i и β_i, не войдут соответственно в состав множеств I_1 и I_2, а также в линейную и кусочно-линейную части комбинированной функции риска (10). Заметим, что вызывает интерес задача построения регрессии (10) с интервальным заданием исходных данных и использованием соответствующих методов (см., например, [13, 14]).

Рассмотрим простой численный пример. Будем строить регрессионную комбинированную функцию риска (10).

Пусть дана выборка данных: $n = 6$, $m = 5$,

$$X = \begin{pmatrix} 1 & 9 & 3 & 1 & 5 \\ 8 & 2 & 8 & 4 & 6 \\ 2 & 7 & 3 & 7 & 5 \\ 6 & 3 & 9 & 4 & 6 \\ 9 & 7 & 4 & 1 & 5 \\ 4 & 1 & 8 & 5 & 7 \end{pmatrix}, \quad y = \begin{pmatrix} 2 \\ 6 \\ 2 \\ 5 \\ 7 \\ 4 \end{pmatrix}.$$

Зададим мощность множеств I_1 и I_2:

$$|I_1| = 2\,;\ |I_2| = 2\,.$$

В результате решения задачи ЛБП (12) – (23) получим значения неизвестных:

$$\sigma = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}; \quad \delta = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}; \quad \gamma = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{pmatrix};$$

$$\alpha = (0{,}71 \quad 0 \quad 0 \quad 0 \quad -0{,}38)\,;$$

$$\beta = (0 \quad 0{,}36 \quad 0{,}34 \quad 0 \quad 0)\,;$$

$$z = \begin{pmatrix} 3{,}21 \\ 2{,}69 \\ 2{,}50 \\ 3{,}03 \\ 2{,}50 \\ 3.84 \end{pmatrix}; \quad u = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}; \quad v = \begin{pmatrix} 0 \\ 0{,}09 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix};$$

$$I_1 = \{1,\ 5\}\,;\quad I_2 = \{2,\ 3\}\,.$$

При этом суммарная абсолютная ошибка аппроксимации равна 0,09.

Таким образом, регрессионная комбинированная функция риска примет вид

$$y_k = 0{,}71x_{k1} - 0{,}38x_{k5} + \max\{\,0{,}36x_{k2}\quad 0{,}34x_{k3}\,\} + \varepsilon_k,\ \ k = \overline{1,\ 6}.$$

Заметим, что независимая переменная x_4 в нее не вошла.

Заключение

Предложен способ идентификации параметров регрессионной комбинированной функции риска и группировки предикторов, входящих в состав ее линейной и кусочно-линейной составляющих. При этом мощности соответствующих индексных множеств считаются заданными. Показано, что в случае, когда в качестве метода оценивания неизвестных параметров используется метод наименьших модулей, эта задача сводится к задаче линейно-булева программирования.

Список литературы

1. Yuan, M. Model Selection and Estimation in Regression with Grouped Variables / M. Yuan, Y. Lin // Journal of the Royal Statistical Society Series B: Statistical Methodology. – 2006. – Vol. 68, No. 1. – P. 49 – 67. doi: 10.1111/j.1467-9868.2005.00532.x

2. Bühlmann, P. Correlated variables in regression: Clustering and sparse estimation / P. Bühlmann, S. van de Geer, C.-H. Zhang // Journal of Statistical Planning and Inference. – 2013. – Vol. 143, No. 11. – P. 1835 – 1858. doi: 10.1016/j.jspi.2013.05.019

3. Wang, S. Hierarchically penalized Cox regression with grouped variables / S. Wang, B. Nan, N. Zhu, J. Zhu // Biometrika. – 2009. – Vol. 96, No. 2. – P. 307 – 322. doi: 10.1093/biomet/asp016

4. Variable Selection in Logistic Regression Model / S. Zhang, L. Zhang, K. Qiu, Y. Lu, B. Cai // Chinese Journal of Electronics. – 2015. – Vol. 24, No. 4. – P. 813 – 817. doi: 10.1049/cje.2015.10.025

5. Meier, L. The Group Lasso for Logistic Regression / L. Meier, S. van de Geer, P. Bühlmann // Journal of the Royal Statistical Society Series B: Statistical Methodology. – 2008. – Vol. 70, No. 1. – P. 53 – 71. doi: 10.1111/j.1467-9868.2007.00627.x

6. Cook, R. D. Partial One-Dimensional Regression Models / R. D. Cook, S. Weisberg // The American Statistician. – 2004. – Vol. 58, No. 2. – P. 110 – 116. doi: 10.1198/0003130043240

7. Group Regularization for Zero-Inflated Poisson Regression Models with an Application to Insurance Ratemaking / S. Chowdhury, S. Chatterjee, H. Mallick, P. Banerjee, B. Garai // Journal of Applied Statistics. – 2019. – Vol. 46, No. 9. – P. 1567 – 1581. doi: 10.1080/02664763.2018.1555232

8. Group Logistic Regression Models with $l_{p,q}$ Regularization / Y. Zhang, C. Wei, C. Wei, X. Liu // Mathematics. – 2022. – Vol. 10, No. 13. – P. 2227. doi: 10.3390/math10132227

9. Algamal, Z. Penalized Poisson Regression Model using adaptive modified Elastic Net Penalty / Z. Algamal, M. H. Lee // Electronic Journal of Applied Statistical Analysis. – 2015. – Vol. 8, No. 2. doi: 10.1285/i20705948v8n2p236

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

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

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

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

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

Grouping of Variables in the Regression Combined Risk Function

S. I. Noskov✉, E. S. Popov, A. A. Butin

Department of Information Systems and Information Security, sergey.noskov.57@mail.ru; Irkutsk State Transport University, Irkutsk, Russia

Keywords: grouping; linear Boolean programming problem; index set; combined risk regression model; set cardinality.

Abstract: The paper provides a brief overview of publications devoted to grouping variables in regression models. In particular, we consider: the problem of selecting grouped variables for accurate forecasting in regression; estimation in a multivariate linear model with highly correlated variables; a new method for solving the problem of group variable selection in the Cox proportional hazards model; problems of selecting variables in a logistic regression model; extending group Lasso to logistic regression models; the method of applying univariate models separately for each level of the grouping variable; the problem of selecting grouped variables in Poisson models; modification of the elastic network and its adaptive analogue, allowing for small and medium correlations between explanatory variables to be taken into account when grouping. An algorithmic method has been developed for identifying the parameters of the regression combined risk function and grouping the independent variables included in its linear and piecewise linear parts. In this case, the number of elements in the corresponding index sets is considered known. It is shown that if the least modulus method is used as a method for determining parameters, this problem is reduced to a linear Boolean programming problem. A numerical example is given.

References

1. Yuan M., Lin Y. Model Selection and Estimation in Regression with Grouped Variables, *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2006, vol. 68, no. 1, pp. 49-67. doi: 10.1111/j.1467-9868.2005.00532.x

2. Bühlmann P., Geer S. van de, Zhang C.-H. Correlated variables in regression: Clustering and sparse estimation, *Journal of Statistical Planning and Inference*, 2013, vol. 143, no. 11, pp. 1835-1858. doi: 10.1016/j.jspi.2013.05.019

3. Wang S., Nan B., Zhu N., Zhu J. Hierarchically penalized Cox regression with grouped variables, *Biometrika*, 2009, vol. 96, no. 2, pp. 307-322. doi: 10.1093/biomet/asp016

4. Zhang S., Zhang L., Qiu K., Lu Y., Cai B. Variable Selection in Logistic Regression Model, *Chinese Journal of Electronics*, 2015, vol. 24, no. 4, pp. 813-817. doi: 10.1049/cje.2015.10.025

5. Meier L., Geer S. van de, Bühlmann P. The Group Lasso for Logistic Regression, *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2008, vol. 70, no. 1, pp. 53-71. doi: 10.1111/j.1467-9868.2007.00627.x

6. Cook R.D., Weisberg S. Partial One-Dimensional Regression Models, *The American Statistician*, 2004, vol. 58, no. 2, pp. 110 – 116. doi: 10.1198/0003130043240

7. Chowdhury S., Chatterjee S., Mallick H., Banerjee P., Garai B. Group Regularization for Zero-Inflated Poisson Regression Models with an Application to Insurance Ratemaking, *Journal of Applied Statistics*, 2019, vol. 46, no. 9, pp. 1567-1581. doi: 10.1080/02664763.2018.1555232

8. Zhang Y., Wei C., Wei C., Liu X. Group Logistic Regression Models with $l_{p,q}$ Regularization, *Mathematics*, 2022, vol. 10, no. 13, pp. 2227. doi: 10.3390/math10132227

9. Algamal Z., Lee M.H. Penalized Poisson Regression Model using adaptive modified Elastic Net Penalty, *Electronic Journal of Applied Statistical Analysis*, 2015, vol. 8, no. 2. doi: 10.1285/i20705948v8n2p236

10. Noskov S.I. [Identification of parameters of a combined piecewise linear regression model], *Vestnik Yugorskogo gosudarstvennogo universiteta* [Bulletin of the Yugra State University], 2022, no. 4 (67), pp. 115-119 (In Russ., abstract in Eng.)

11. Noskov S.I. [Identification of parameters of a piecewise linear risk function], *Transportnaya infrastruktura Sibirskogo regiona* [Transport infrastructure of the Siberian region], 2017, vol. 1, pp. 417-421 (In Russ., abstract in Eng.)

12. Noskov S.I., Protopopov V.A. [Assessing the level of vulnerability of transport infrastructure objects: a formalized approach], *Sovremennyye tekhnologii. Sovremennyy analiz. Modelirovaniye* [Modern technologies. Modern analysis. Modeling], 2011, no. 4(32), pp. 241-244. (In Russ., abstract in Eng.)

13. Kreinovich V., Lakeyev A.V., Noskov S.I. Approximate linear algebra is intractable, *Linear Algebra and its Applications*, 1996, vol. 232, no. 1-3, pp. 45-54.

14. Noskov S.I. [Point characterization of sets of solutions of interval systems of linear algebraic equations], *Informatsionnyye tekhnologii i matematicheskoye modelirovaniye v upravlenii slozhnymi sistemami* [Information technologies and mathematical modeling in the management of complex systems], 2018, no. 1, pp. 8-13. (In Russ., abstract in Eng.)

Gruppierung von Variablen in einer kombinierten Regressionsrisikofunktion

Zusammenfassung: Es ist ein kurzer Überblick über Veröffentlichungen gegeben, die sich mit der Gruppierung von Variablen in Regressionsmodellen befassen. Insbesondere werden folgende Punkte berücksichtigt: das Problem der Auswahl gruppierter Variablen für eine genaue Prognose bei der Regression; Schätzung in einem multivariaten linearen Modell mit stark korrelierten Variablen; eine neue Methode zur Lösung des Problems der Auswahl einer Gruppenvariablen im Cox-Proportional-Hazards-Modell; Probleme der Variablenauswahl in logistischen Regressionsmodellen; Erweiterung von Group Lasso auf logistische Regressionsmodelle; eine Methode zum separaten Anwenden invarianter Modelle für jede Ebene der Gruppierungsvariablen; das Problem der Auswahl von Gruppierungsvariablen in Poisson-Modellen; eine Modifikation des elastischen Netzwerks und seines adaptiven Analogons, die es ermöglicht, kleine und mittlere Korrelationen zwischen erklärenden Variablen bei der Gruppierung zu berücksichtigen. Es ist eine algorithmische Methode zur Identifizierung der Parameter einer kombinierten Regressionsrisikofunktion und zur Gruppierung unabhängiger Variablen entwickelt, die in ihren linearen und stückweise linearen Teilen enthalten sind. In diesem Fall gilt die Anzahl der Elemente in den entsprechenden

Indexsätzen als bekannt. Es ist gezeigt, dass dieses Problem auf ein linear-Boolesches Programmierproblem reduziert wird, wenn die Methode der kleinsten absoluten Werte als Methode zur Bestimmung von Parametern verwendet wird. Es ist ein Zahlenbeispiel betrachtet.

Regroupement des variables dans la fonction de régression combinée du risqué

Résumé: Est donnée la révue des publications sur le regroupement des variables dans les modèles de régression. En particulier, sont examinées les questions de la sélection des variables groupées pour une prédiction précise en régression; de l'évaluation dans un modèle linéaire multivarié avec des variables fortement corrélées; d'une nouvelle méthode pour résoudre le problème de la sélection de variables de groupe dans le modèle de risque proportionnel de Cox; des problèmes de sélection de variables dans le modèle de régression logistique; de l'extension du Lasso de groupe au modèle de régression logistique; de la méthode d'application des modèles unidimensionnels séparément pour chaque niveau de la variable de groupe; le problème du choix des variables groupées dans les modèles de Poisson; de la modification du réseau élastique et de son analogue adaptatif permettant de prendre en compte la corrélation faible et moyenne entre les variables explicatives lors du regroupement. Est mis au point le moyen algorithmique pour identifier les paramètres d'une fonction de risque combinée de régression et pour regrouper les variables indépendantes qui composent ses parties linéaires et linéaires par morceaux. Dans ce cas, le nombre d'éléments dans les ensembles d'indices correspondants est considéré comme connu. Est démontré que si la méthode des plus petits modules est utilisée comme méthode de définition des paramètres, cette tâche est réduite à une tâche de programmation linéaire-mais booléenne. Est considéré un exemple numérique.

Авторы: ***Носков Сергей Иванович*** – доктор технических наук, профессор кафедры «Информационные системы и защита информации»; ***Попов Егор Сергеевич*** – магистрант; ***Бутин Александр Алексеевич*** – кандидат физико-математических наук, доцент кафедры «Информационные системы и защита информации», ФГБОУ ВО «Иркутский государственный университет путей сообщения», Иркутск, Россия.