УДК 519.852 *DOI 10.55421/3034-4689_2026_29_3_142*

С. И. Носков, А. С. Вергасов, И. Д. Кириллов

ОСОБЕННОСТИ ПОСТРОЕНИЯ КЛАСТЕРНОЙ ЛИНЕЙНОЙ РЕГРЕССИИ С ОБЩЕЙ ЧАСТЬЮ

Ключевые слова: кластерная линейная регрессия с общей частью, метод наименьших модулей, задача линейно-булева программирования, средняя процентная ошибка, ГЭС, гидроагрегат.

В работе дан краткий обзор публикаций по различным вопросам разработки методов построения кластерных моделей различного типа и прикладной направленности. В частности, упомянута задача прогнозирования оставшегося срока службы при техническом обслуживании авиадвигателя в различных условиях эксплуатации, решаемая с помощью самоадаптивного подхода к динамической кластеризации для отбора полезных мультимодальных данных в различные кластеры, каждый из которых имеет постоянную тенденцию к деградации. Рассмотрено расширение известного алгоритма k-средних на многокритериальную структуру, которое основано на определении многокритериального расстояния на основе предпочтений, определенных лицом, принимающим решения. Приведено краткое изложение методологии кластеризации для прямой группировки временных рядов и автоматической стратегии рекурсивного поиска для анализа энергетической периодичности в этих рядах, а также новой модели нечетких временных рядов, которая может интерполировать исторические данные для эффективного прогнозирования на будущее и в рамках которой после нормализации исходных данных создается автоматический алгоритм для определения подходящего количества кластеров и нахождения нечетких связей каждого элемента в последовательности с установленными кластерами. Представлено сжатое описание нового метода инкрементного нечеткого кластерного ансамбля, основанного на идее объединения задачи кластерного анализа с методами классификации, а также подхода к групповой кластеризации для улучшения процесса категоризации при планировании рабочих нагрузок в облачных центрах обработки данных. В статье предложен способ идентификации параметров кластерной линейной регрессии с общей частью и группировки переменных в соответствующих множествах индексов, предполагающий сведение этой задачи к задаче линейно-булева программирования. Построены кластерная и обычная линейная модель функционирования гидроагрегата одной из ГЭС Сибири. Проведен их краткий анализ.

S. I. Noskov, A. S. Vergasov, I. D. Kirillov

FEATURES OF CONSTRUCTING A CLUSTER LINEAR REGRESSION WITH A COMMON PART

Keywords: cluster linear regression with a common part, method of least modules, linear-Boolean programming problem, average percentage error, hydroelectric power station, hydroelectric unit.

The paper provides a brief overview of publications on various issues of developing methods for constructing cluster models of various types and applications. In particular, the problem of predicting the remaining service life during maintenance of an aircraft engine under various operating conditions is mentioned, which is solved using a self-adaptive approach to dynamic clustering to select useful multimodal data into various clusters, each of which has a constant tendency to degradation. An extension of the well-known k-means algorithm to a multi-criteria structure is considered, which is based on determining the multi-criteria distance based on preferences determined by the decision maker. A summary of the clustering methodology for direct grouping of time series and an automatic recursive search strategy for analyzing energy periodicity in these series is given, as well as a new fuzzy time series model that can interpolate historical data for effective forecasting for the future and, after normalization of the initial data, creates an automatic algorithm to determine the appropriate number of clusters and find fuzzy ones. the relationships of each element in the sequence with the established clusters. A concise description of a new incremental fuzzy cluster ensemble method is presented, based on the idea of combining cluster analysis tasks with classification methods, as well as an approach to group clustering to improve the categorization process when planning workloads in cloud data centers. A method is proposed for identifying the parameters of a cluster linear regression with a common part and grouping variables in the corresponding sets of indexes, suggesting reducing this problem to a linear Boolean programming problem. A cluster and conventional linear model of the operation of a hydroelectric unit of one of the Siberian hydroelectric power plants are constructed. Their brief analysis is carried out.

Рассмотрим кластерную линейную регрессию (КЛР):

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

Здесь $x_i,\ i=\overline{1,m}$ – независимые переменные, y – зависимая переменная, k – номер наблюдения, r – заданное число кластеров, P^j – индексные множества, при этом

$P^j \subset \{1,2,\dots,n\},\quad \bigcup_{j=1}^{r} P^j=\{1,2,\dots,n\},\ P^i \cap P^j=\emptyset,\ i \neq j,\ n$ – длина выборки данных, $\alpha_i^j,\ i=\overline{0,m}, j=\overline{1,r}$ – оцениваемые параметры, $\varepsilon_k^j,\ j=\overline{1,r},\ k \in P^j$ – ошибки аппроксимации. Все переменные в модели (1) детерминированы.

Разработка КЛР (1) сопряжена с вычислением оценок ее параметров и формированием составов кластеров, т.е. в случае, если в качестве расстояния между расчетными и фактическими значениями зависимой переменной принято городское расстояние, с решением следующей задачи (см., в частности, [1,2]):

$$G(A,\Xi)=\sum_{k=1}^{n}\sum_{j=1}^{r}\sigma_{kj}|\varepsilon_k^j| \to min, \qquad (2)$$

где $A=||\alpha_i^j||,\ j=\overline{1,r},\ i=\overline{0,m},\ \Xi=||\sigma_{kj}||,\ k=\overline{1,n},\ j=\overline{1,r}$,

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

Вопросам разработки методов построения кластерных моделей различного типа и прикладной направленности уделяется серьезное внимание в научной литературе. Так, в работе [3] рассматривается проблема прогнозирования оставшегося срока службы при техническом обслуживании авиадвигателя в различных условиях эксплуатации. Разработан самоадаптивный подход к динамической кластеризации для отбора полезных мультимодальных данных в различные кластеры, каждый из которых имеет постоянную тенденцию к деградации. В исследовании [4] предлагается расширить хорошо известный алгоритм k-средних на многокритериальную структуру. Это расширение основано на определении многокритериального расстояния на основе структуры предпочтений, определенной лицом, принимающим решения. В этом случае две альтернативы будут похожи, если они предпочтительны, безразличны и несравнимы с более или менее одинаковыми действиями. Такой подход позволяет разделить множество альтернатив на классы, которые имеют смысл с многокритериальной точки зрения. Статья [5] посвящена описанию методологии кластеризации для прямой группировки временных рядов и автоматической стратегии рекурсивного поиска и анализа энергетической периодичности в этих рядах, что позволяет извлекать информацию на разных уровнях детализации проблемы. В [6] предлагается новая модель нечетких временных рядов, которая может интерполировать исторические данные для эффективного прогнозирования на будущее. В этой модели, после нормализации исходных данных, создается автоматический алгоритм для определения подходящего количества кластеров и нахождения нечетких связей каждого элемента в последовательности с установленными кластерами. В публикации [7] для решения проблемы уменьшения неопределенности и нечеткости в наборах данных предлагается новый метод инкрементного нечеткого кластерного ансамбля, основанный на идее объединения задачи кластерного анализа с методами классификации. Показано, что качество конечного решения слабо коррелирует с размером ансамбля, настройка параметров при построении грубых приближений является приемлемой, а предлагаемый метод устойчив к разнообразию элементов жесткой кластеризации.

В статье [8] представлен новый подход к групповой кластеризации для улучшения процесса категоризации при планировании рабочих нагрузок в облачных центрах обработки данных. Этот подход сочетает в себе различные методы нормализации и преобразования, включая анализ основных компонентов, для создания нескольких конвейеров предварительной обработки данных. Информация, полученная из этих конвейеров, служит входными данными для различных моделей базовой кластеризации. В [9] предлагаются методы построения матрицы различий объектов с разных сайтов с сохранением конфиденциальности. Эти матрицы могут быть использованы для кластеризации с сохранением конфиденциальности, а также для объединения баз данных, привязки записей и других операций, требующих парного сравнения отдельных объектов личных данных, горизонтально распределенных по нескольким сайтам. В исследовании [10] представлен алгоритм, который обнаруживает кластеры в подпространствах, охватываемых различными комбинациями измерений, с помощью локальных взвешиваний объектов. Такой подход позволяет привязать к каждому кластеру вектор веса, значения которого отражают значимость объектов в этом кластере.

В приведенных выше весьма интересных работах не рассматриваются вопросы группировки переменных и идентификации параметров в кластерной линейной модели с общей компонентой, входящей во все выделенные частные регрессии. Их решению и посвящена настоящая работа.

Оценивание параметров и группировка переменных в КЛР с общей частью

Рассмотрим теперь кластерную линейную регрессию с общей частью:

$$y_k=\alpha_0^j + \sum_{i\in I^1}\alpha_i^j x_{ki}+\sum_{i\in I^2}\beta_i x_{ki} +\varepsilon_k^j,\ j=\overline{1,r},\ k\in P^j, \quad (3)$$

в которой компонента $\sum_{i\in I^2}\beta_i x_{ki}$ входит в каждый из кластеров, т.е. является общей для них. Подобные модели рассматриваются, в частности, в работах по одновременному компонентному анализу (SCA) (см., например, [11-13]).

По отношению к КЛР (3) задача (2) должна быть уточнена следующим образом:

$$\tilde{G}\left(\alpha_0^1,\dots,\alpha_0^r,\tilde{A},\beta,\Xi,I^1,I^2\right)=\sum_{k=1}^n\sum_{j=1}^r\sigma_{kj}|\varepsilon_k^j|\to min, \quad (4)$$

где $\tilde{A}=||\alpha_i^j||$, $j=\overline{1,r}$, $i\in I^1$, $\beta=(\beta_i)$, $i\in I^2$.

Таким образом, в соответствии с задачей (4), необходимо не только вычислить оценки параметров КЛР (3), но и сформировать индексные множества $I^1\subset\{1,2,\dots,m\}$, $I^2\subset\{1,2,\dots,m\}$. При этом, очевидно, $I^1\cap I^2=\emptyset$ и, в общем случае, $I^1\cup I^2\subseteq\{1,2,\dots,m\}$, т.е. допускается невхождение некоторых независимых переменных в модель (3). Число элементов (мощность) в индексных множествах I^1 и I^2 ($|I^1|$ и $|I^2|$) считается заданной:

$|I^1|=d_1$, $|I^2|=d_2$.

В работах [1,2] показано, что задача (2) построения КЛР (1) сводится к следующей задаче линейно-булева программирования (ЛБП):

$$\alpha_0^j+\sum_{i=1}^m\alpha_i^j x_{ki}+\sum_{i=1}^m\beta_i x_{ki}- M\sigma_{kj}+u_k\geq y_k- M,\ j=\overline{1,r},\ k=\overline{1,n}, \quad (5)$$

$$\alpha_0^j+\sum_{i=1}^m\alpha_i^j x_{ki}+\sum_{i=1}^m\beta_i x_{ki}+ M\sigma_{kj}-u_k\leq y_k+ M,\ j=\overline{1,r},\ k=\overline{1,n}, \quad (6)$$

$$\sum_{j=1}^r\sigma_{kj}=1,\ k=\overline{1,n}, \quad (7)$$

$$\sigma_{kj}\in\{0,1\},\ j=\overline{1,r},\ k=\overline{1,n}, \quad (8)$$

$$u_k\geq 0, k=\overline{1,n}, \quad (9)$$

$$\sum_{i=1}^n u_k\to min. \quad (10)$$

Для решения на основе задачи ЛБП (5) – (10) задачи (4) воспользуемся вычислительным приемом, примененным в монографии [14] при построении аддитивной по параметрам регрессионной зависимости. Введем в рассмотрение булевы переменные $\delta_i, \gamma_i, i=\overline{1,m}$ по следующим правилам:

$$\delta_i = \begin{cases} 1, & i \in I^1 \\ 0, & \text{в противном случае,} \end{cases}$$

$$\gamma_i = \begin{cases} 1, & i \in I^2 \\ 0, & \text{в противном случае.} \end{cases}$$

Тогда для решения задачи (4) необходимо дополнить ограничения задачи ЛБП (5) – (10) следующими:

$$-N_i\delta_i \le \alpha_i^j \le N_i\delta_i,\ i=\overline{1,m},\ j=\overline{1,r}, \quad (11)$$

$$-N_i\gamma_i \le \beta_i \le N_i\gamma_i,\ i=\overline{1,m}, \quad (12)$$

$$\sum_{i=1}^{m}\delta_i = d_1, \quad (13)$$

$$\sum_{i=1}^{m}\gamma_i = d_2, \quad (14)$$

$$\delta_i \in \{0,1\},\ i=\overline{1,m}, \quad (15)$$

$$\gamma_i \in \{0,1\},\ i=\overline{1,m}. \quad (16)$$

Здесь N_i, $i=\overline{1,m}$ – наперед заданные большие положительные числа. Целевая функция (10) останется неизменной.

Проиллюстрируем описанный способ построения КЛР с общей частью на статистических данных функционирования гидроагрегата одной из ГЭС Сибири, приведенных в работе [15]. Используем обозначения: y - активная мощность, МВт; x_1 - реактивная мощность, МВт; x_2 - напряжение статора, В; x_3 - ток ротора, А; x_4 - ток статора, А; x_5 - температура горячего воздуха воздухоохладителя, °С. Таким образом, m=5, n=20. Зададим мощность (относительно равную) индексных множеств I^1, I^2 и число кластеров:

$$d_1 = 2,\ d_2 = 3,\ r = 2.$$

Большее число кластеров задавать нецелесообразно из-за ограниченности выборки двадцатью наблюдениями.

Будем строить КЛР с общей частью без свободного члена, т.е. $\alpha_0^1 = \alpha_0^2 = 0$. В результате решения задачи ЛБП (5) – (9), (11) – (16), (10) получим следующую модель:

$$y_k = -2.2x_{k1} + 0.31x_{k5} - 42.87x_{k2} + 623.3x_{k3} + 0.89x_{k4} + \varepsilon_k^1, \quad (17)$$

$k \in P^1 = \{2,3,4,5,6,7,8,14,15,16,17,19\}$,

$$y_k = -2.11x_{k1} + 0.11x_{k5} - 42.87x_{k2} + 623.3x_{k3} + 0.89x_{k4} + \varepsilon_k^2, \quad (18)$$

$k \in P^2 = \{1,9,10,11,12,13,18,20\}$, $\tilde{G}$ = 5.48, E = 0.13%.

Здесь E – средняя процентная ошибка, вычисляемая по формуле:

$$E = 100\% \sum_{i=1}^{20} |\varepsilon_k| \,/\, \sum_{i=1}^{20} y_k.$$

Построим теперь по тем же данным обычную линейную регрессию методом наименьших модулей (см, в частности, [16]):

$$y_k = -2.1x_{k1} - 40.3x_{k2} + 582.2x_{k3} + 3.13x_{k4} + 0.3x_{k5} + \varepsilon_k, \quad (19)$$

$\sum_{i=1}^{20} \varepsilon_k$ = 15, E = 0.33%.

Анализ КЛР с общей частью (17), (18) и линейной модели (19) позволяет сделать следующие выводы.

1. Индексное множество P^1 содержит двенадцать элементов, а P^2 – восемь.

2. Знаки параметров в кластерных частях модели (17), (18) совпадают, хотя сами оценки параметров несколько различаются.

3. Совершенно естественно, что точность кластерной модели существенно выше, чем линейной, и для второй она, судя по критерию E, вполне приемлема.

Заключение

Оценивая в целом полученные результаты, отметим следующее. Реализация задачи (4) путем сведения ее к задаче линейно-булева программирования (5) – (9), (11) – (16), (10) с последующим решением позволяет достичь нескольких важных целей. Во-первых, обеспечивается возможность выделения общей для всех кластеров линейной компоненты. Во-вторых, производится группировка независимых переменных в ней и в частных линейных регрессиях. Наконец, в-третьих, рассчитываются оценки параметров во всей сформированной таким образом кластерной модели. Разработаны обычная и кластерная линейная модель функционирования гидроагрегата одной из ГЭС Сибири, обладающая высокими аппроксимационными характеристиками. Проведен их краткий содержательный анализ.

Литература

1. Носков С. И., Беляев С. В. *Оценка непротиворечивости кластерной линейной регрессионной модели* // Вестник Технологического университета. 2025. Т. 28, № 2. С. 88-91.
2. D. Bertsimas, R. Shioda, Classification and regression via integer optimization, *Operations Research*, 55, 2, 252–271 (2007). DOI: 10.1287/opre.1060.0360.
3. J. Chen, D. Li, R. Huang, Z. Chen, W. Li, *Aero-engine remaining useful life prediction method with self-adaptive multimodal data fusion and cluster-ensemble transfer regression*, Reliability Engineering & System Safety, 234, Article 109151 (2023), DOI: 10.1016/j.ress.2023.109151.
4. Y. De Smet, L. Montano Guzmán, *Towards multicriteria clustering: An extension of the k-means algorithm*, European Journal of Operational Research, 158, 2, 390–398 (2004). DOI: 10.1016/j.ejor.2003.06.012.
5. L.G.B. Ruiz, M.C. Pegalajar, M. Molina-Solana, *A time-series clustering methodology for knowledge extraction in energy consumption data*, Expert Systems with Applications, 160, Article 113731 (2020). DOI: 10.1016/j.eswa.2020.113731.
6. T. Vovan, N. Ledai, *A new fuzzy time series model based on cluster analysis problem*, International Journal of Fuzzy Systems, 21, 852–864 (2019). DOI: 10.1007/s40815-018-0589-x.
7. J. Hu, T. Li, C. Luo, H. Fujita, Y. Yang, *Incremental fuzzy cluster ensemble learning based on rough set theory*, Knowledge-Based Systems, 132, 144–155 (2017). DOI: 10.1016/j.knosys.2017.06.020.
8. M. Daraghmeh, A. Agarwal, Y. Jararweh, *An ensemble clustering approach for modeling hidden categorization perspectives for cloud workloads*, Cluster Computing, 27, 4779–4803 (2024). DOI: 10.1007/s10586-023-04205-5.
9. A. İnan, S.V. Kaya, Y. Saygın, E. Savaş, A.A. Hintoğlu, A. Levi, *Privacy preserving clustering on horizontally partitioned data*, Data & Knowledge Engineering, 63, 3, 646–666 (2007). DOI: 10.1016/j.datak.2007.03.015.
10. C. Domeniconi, D. Gunopulos, S. Ma, B. Yan, M. Al-Razgan, D. Papadopoulos, *Locally adaptive metrics for*

clustering high dimensional data, Data Mining and Knowledge Discovery, 14, 63–97 (2007). DOI: 10.1007/s10618-006-0060-8.
11. K. De Roover, M.E. Timmerman, B. Mesquita, E. Ceulemans, *Common and cluster-specific simultaneous component analysis*, The PLOS ONE Staff, 8, e62280 (2013). DOI: 10.1371/journal.pone.0062280.
12. M. Schouteden, K. Van Deun, S. Pattyn, I. Van Mechelen, *SCA with rotation to distinguish common and distinctive information in linked data*, Behavior Research Methods, 45, 3, 822–833 (2013). DOI: 10.3758/s13428-012-0295-9.
13. O. Alter, P.O. Brown, D. Botstein, *Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms*, Proceedings of the National Academy of Sciences, 100, 6, 3351–3356 (2003). DOI: 10.1073/pnas.0530258100.
14. С.И. Носков, *Технология моделирования объектов с нестабильным функционированием и неопределенностью в данных*, Облинформпечать, Иркутск (1996). 320 с.
15. С.И. Носков, В.А. Лисицын, *Подход к заполнению пропусков в данных о функционировании гидроагрегата*, Инженерный вестник Дона, 2024, № 10 (118), 567–575.

References

1. Noskov S. I., Belyaev S. V. Evaluation of the consistency of the cluster linear regression model // Bulletin of the Technological University. 2025. Vol. 28, No. 2. pp. 88-91.
2. D. Bertsimas, R. Shioda, Classification and regression via integer optimization, Operations Research, 55, 2, 252–271 (2007). DOI: 10.1287/opre.1060.0360.
3. J. Chen, D. Li, R. Huang, Z. Chen, W. Li, Aero-engine remaining useful life prediction method with self-adaptive multimodal data fusion and cluster-ensemble transfer regression, Reliability Engineering & System Safety, 234, Article 109151 (2023), DOI: 10.1016/j.ress.2023.109151
4. Y. De Smet, L. Montano Guzmán, Towards multicriteria clustering: An extension of the k-means algorithm, European Journal of Operational Research, 158, 2, 390–398 (2004). DOI: 10.1016/j.ejor.2003.06.012
5. L.G.B. Ruiz, M.C. Pegalajar, M. Molina-Solana, A time-series clustering methodology for knowledge extraction in energy consumption data, Expert Systems with Applications, 160, Article 113731 (2020). DOI: 10.1016/j.eswa.2020.113731
6. T. Vovan, N. Ledai, A new fuzzy time series model based on cluster analysis problem, International Journal of Fuzzy Systems, 21, 852–864 (2019). DOI: 10.1007/s40815-018-0589-x
7. J. Hu, T. Li, C. Luo, H. Fujita, Y. Yang, Incremental fuzzy cluster ensemble learning based on rough set theory, Knowledge-Based Systems, 132, 144–155 (2017). DOI: 10.1016/j.knosys.2017.06.020.
8. M. Daraghmeh, A. Agarwal, Y. Jararweh, An ensemble clustering approach for modeling hidden categorization perspectives for cloud workloads, Cluster Computing, 27, 4779–4803 (2024). DOI: 10.1007/s10586-023-04205-5.
9. A. İnan, S.V. Kaya, Y. Saygın, E. Savaş, A.A. Hintoğlu, A. Levi, Privacy preserving clustering on horizontally partitioned data, Data & Knowledge Engineering, 63, 3, 646–666 (2007). DOI: 10.1016/j.datak.2007.03.015.
10. C. Domeniconi, D. Gunopulos, S. Ma, B. Yan, M. Al-Razgan, D. Papadopoulos, Locally adaptive metrics for clustering high dimensional data, Data Mining and Knowledge Discovery, 14, 63–97 (2007). DOI: 10.1007/s10618-006-0060-8.
11. K. De Roover, M.E. Timmerman, B. Mesquita, E. Ceulemans, Common and cluster-specific simultaneous component analysis, The PLOS ONE Staff, 8, e62280 (2013). DOI: 10.1371/journal.pone.0062280.
12. M. Schouteden, K. Van Deun, S. Pattyn, I. Van Mechelen, SCA with rotation to distinguish common and distinctive information in linked data, Behavior Research Methods, 45, 3, 822–833 (2013). DOI: 10.3758/s13428-012-0295-9.
13. O. Alter, P.O. Brown, D. Botstein, Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms, Proceedings of the National Academy of Sciences, 100, 6, 3351–3356 (2003). DOI: 10.1073/pnas.0530258100.
14. S.I. Noskov, Technology of modeling objects with unstable functioning and uncertainty in data, Oblinformpechat, Irkutsk (1996). 320 p.
15. S.I. Noskov, V.A. Lisitsyn, An approach to filling in gaps in data on the operation of a hydraulic unit, Engineering Bulletin of the Don, 2024, № 10 (118), 567-575.

© С. И. Носков – д-р техн. наук, проф., профессор кафедры «Информационные системы и защита информации» (ИСЗИ), Иркутский государственный университет путей сообщения (ИГУПС), Иркутск, Россия, sergey.noskov.57@mail.ru; **А. С. Вергасов** – старший преподаватель кафедры ИСЗИ, ИГУПС, Tluck@inbox.ru; **И. Д. Кириллов** – студент кафедры ИСЗИ, ИГУПС, iliakirillov07@gmail.com.

© S. I. Noskov – Doctor of Sciences, Prof., Professor of the department of Information Systems and Information Security (ISIS), Irkutsk State Transport University (ISTU), Irkutsk, Russia, sergey.noskov.57@mail.ru; **A. S. Vergasov** – Senior Lecturer at the ISIS department, ISTU, Tluck@inbox.ru; **I. D. Kirillov** – Student of the ISIS department, ISTU, iliakirillov07@gmail.com.

Дата поступления рукописи в редакцию – 14.01.26.
Дата принятия рукописи в печать – 02.03.26.