УДК 519.23
DOI 10.34822/1999-7604-2021-3-46-50

ПРИМЕНЕНИЕ МЕТОДА АНТИРОБАСТНОГО ОЦЕНИВАНИЯ ПАРАМЕТРОВ ДЛЯ КЛАСТЕРИЗАЦИИ ВЫБОРКИ ДАННЫХ

С. И. Носков
Иркутский государственный университет путей сообщения, Иркутск, Россия
E-mail: sergey.noskov.57@mail.ru

В работе представлен алгоритм кластеризации данных, основанный на построении линейной регрессионной модели с применением метода антиробастного оценивания параметров. Он обладает свойством равенства числа максимальных по модулю ошибок аппроксимации числу параметров плюс единица. Сформированы три кластера выборки данных при моделировании объема погрузки на железнодорожном транспорте в Российской Федерации на основе статистической информации за 2005–2018 гг.

Ключевые слова: кластеризация, данные, выборка, регрессионная модель, метод антиробастного оценивания параметров.

APPLICATION OF METHOD OF ANTIROBUST ESTIMATION OF PARAMETERS FOR DATA SAMPLING CLUSTERING

S. I. Noskov
Irkutsk State Transport University, Irkutsk, Russia
E-mail: sergey.noskov.57@mail.ru

The article presents an algorithm of data clustering based on the construction of linear regression model with application of method of antirobust estimation of parameters. The algorithm has a property of equity of the number of maximum module approximation errors to the number of parameters plus one. Three clusters of data selecting, with modeling of loading volumes at the Russian Federation railways, are formulated based on the statistical data for the period of 2005–2018.

Keywords: clustering, data, sampling, regression model, method of antirobust estimation of parameters.

Введение

Одной из важных проблем анализа данных является разбиение исходной выборки данных на подвыборки, называемые кластерами, которые наделены некоторыми уникальными, присущими только им свойствами. Ее решению посвящено значительное количество работ. Так, в [1] предложена процедура генерирования иерархии кластеров для геохимических и геологических процессов, происходящих в разных пространственных масштабах. В работе [2] выполнена кластеризация данных о транспортных средствах, проведен анализ сгенерированных кластеров совместно с информацией о выбросах вредных веществ, вызванных сложными корреляциями их компонентов, предложен метод исследования данных кластера с репрезентативными атрибутами и определения его характеристик на основе отношений между данными о транспортном средстве. В [3] при моделировании внутреннего валового продукта стран исходная выборка разбивается на подвыборки на основе уровня их макроэкономического развития. В статье [4] исследуются распределительные свойства ряда схем для выбора регрессионной модели и оценки ее параметров с учетом разделения данных на подвыборки по пространству локальных параметров. В работе [5] предложена схема разбиения выборки на подвыборки для крупномасштабной многоклассовой логистической регрессии.

В работе [6] представлена так называемая гибридная нейро-фаззи система, предназначенная для решения задачи вероятностной и нечеткой классификации данных, представлен-

ных короткими выборками с пересекающимися классами произвольной формы. В [7] описан метод исследования панельных данных с использованием агломеративной иерархической кластеризации, т. е. группировки объектов на основании исследования сходства и различия их признаков. Применены два способа вычисления расстояний между объектами: расстояния между усредненными по интервалу наблюдений значениями и расстояния с использованием информации за весь изучаемый период. Произведено сравнение трех возможных методов вычисления расстояний между кластерами. В одном случае принято расстояние между ближайшими элементами из двух кластеров, в другом – среднее по парам элементов, в третьем – расстояние между наиболее удаленными элементами. Исследована эффективность применения индексов качества кластеризации Данна (Dunn Index, DI) и силуэта (Silhouette) для выбора оптимального числа кластеров и оценки статистической значимости полученных решений. В статье [8] предложена робастная модификация метода K-средних для решения задачи кластеризации при условии, что элементы обучающей выборки заданы в виде интервалов, а не точечно, как это делается в традиционных постановках. В работе [9] предложена модификация метода быстрого прототипирования систем нечеткого вывода на основе результатов обработки обучающей выборки эвристическим алгоритмом так называемой возможностной кластеризации для случая априори неизвестного числа классов. В [10] представлен алгоритм иерархической кластеризации, а также описана оценка качества разбиения по размеченной выборке. При этом основная идея такой кластеризации заключается в последовательном объединении меньших кластеров в большие с помощью агломеративных методов или разделении больших кластеров на меньшие посредством применения дивизивных методов (Divisive Analysis).

Следует обратить внимание и на другие работы, посвященные кластеризации данных [11–16].

Кластеризация информации на основе метода антиробастного оценивания параметров

Предположим, при моделировании некоторого сложного объекта методами регрессионного анализа исследователь полагает, что поведение выходного (зависимого) показателя (фактора, переменной) *y* определяется значениями входных (независимых) показателей x_1, x_2, …, x_m и это влияние имеет линейный характер, то есть правомерна регрессионная модель:

$$y_k = \sum_{i=1}^{m} a_i x_{ki} + \varepsilon_k, \; k = \overline{1,n}, \tag{1}$$

где k – номер наблюдения, n – длина обрабатываемой выборки, $a = (a_1, ..., a_m)^T$ – вектор оцениваемых параметров, ε_k – ошибки аппроксимации. Отметим, что модель (1) полностью детерминирована. Выборка (X, y) предполагается заданной, при этом $y = (y_1, ..., y_n)^T$, X – $(n \times m)$ – матрица с компонентами x_{ki}, $k = \overline{1,n}$, $i = \overline{1,m}$.

Задача состоит в разбиении (кластеризации) заданной выборки на классы (группы) наблюдений, каждый из которых обладает некоторыми уникальными по отношению к модели (1) свойствами. Одним из таких свойств может быть точность модели, определяемая величиной ошибок аппроксимации, одинаковой внутри каждого класса. В работе [17] подобная кластеризация произведена на основе метода наименьших модулей (МНМ) при оценивании параметров модели (1), обладающего свойством равенства нулю числа ошибок аппроксимации, не меньшего, чем m [18]. В регрессионном анализе значительно менее популярным, чем методы наименьших квадратов и модулей, является метод антиробастного оценивания (МАО) параметров, в соответствии с которым искомая оценка α^* является решением задачи:

$$\alpha^* = \arg\min J(\alpha),$$

где

$$J(\alpha) = \max_{k=\overline{1,n}} | \varepsilon_k | .$$

Она может быть сведена к задаче линейного программирования (см., например, [19]). В [20] показано, что при применении МАО число максимальных по модулю ошибок аппроксимации в модели (1) должно быть не меньше, чем $m + 2$.

Указанное свойство МАО позволяет кластеризовать все n наблюдений выборки (X, y) на d групп, где число d задается по правилу:

$$d = \begin{cases} n/m+1, \textit{ если число } n/m+1 \textit{ целое} \\ [n/m+1]+1, \textit{ в противном случае.} \end{cases}$$

Здесь $[c]$ – целая часть числа с.

При этом может быть реализован следующий алгоритм.

Обозначим через G множество номеров наблюдений выборки (X, y), т. е. $G = \{1, 2, \ldots, n\}$. Построим на ней регрессионную модель (1). Сформируем множество номеров наблюдений $S_1 = \{s_1, ..., s_{m+1}\}$, для элементов которого справедливо условие: $s_i \in S_1 \Leftrightarrow | \varepsilon_{s_i} | = \max_{k \in S} | \varepsilon_k |$. Следует иметь в виду, что во множество S_1 может быть включено больше, чем m + 1 элементов, как это показано в [20]. На следующем шаге сформируем множество номеров наблюдений $G_1 = G \setminus S_1$ и опять построим модель (1) на вновь образованной выборке с номерами из множества G_1. Далее сформируем множество S_2, включающее в свой состав номера наблюдений с максимальными по модулю ошибками аппроксимации, построим множество G_2 и т. д. Этот алгоритм завершается построением множества S_d, которое может содержать менее $m + 1$ элементов.

Множества S_i, $i = \overline{1,d}$ обладают следующим свойством – S_d содержит номера наблюдений выборки, на которых модель (1) максимально точно описывает исследуемый объект, для множества S_{d-1} характерно менее точное его описание, а номера наблюдений из множества S_1 обладают тем свойством, что на них модель (1) наименее адекватна в смысле величины ошибок аппроксимации.

Полученная в результате кластеризация множества номеров наблюдений выборки примет вид:

$$G = \bigcup_{i=1}^{d} S_i, \; S_i \bigcap S_j = \varnothing, \; i \neq j.$$

Развитие предложенного подхода возможно с применением аппарата построения линейно-неэлементарных [21], парно-множественных [22] и степенно-показательных [23] регрессионных моделей с возможностью сведения некоторых постановок задач к многокритериальной задаче линейного программирования [24].

Кластеризация данных при моделировании объема погрузки на железнодорожном транспорте

В работе [25] решена задача построения регрессионной модели объема погрузки на железнодорожном транспорте в Российской Федерации методом смешанного оценивания параметров на основе статистической информации за 2005–2018 гг. (всего 14 наблюдений). При этом в качестве переменных модели приняты следующие:

y – объем погрузки основных видов грузов на железнодорожном транспорте, тыс. т;

x_1 – добыча угля, млн т;

x_2 – вывозка древесины, млн $м^3$;

x_3 – рабочий парк груженых железнодорожных вагонов, тыс. шт.

На этих данных с помощью МАО было построено 5 линейных регрессий

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3$$

с последовательным уменьшением объема выборки на m + 2 = 5 наблюдений на каждом шаге работы алгоритма.

Число сформированных кластеров (групп номеров наблюдений) составляет [14/5] + 1 = 3.

Первый кластер:

$$S_1 = \{2, 4, 8, 9, 10\},$$
$$y = 1001045.8 - 185.4x_1 + 464.9x_2 + 1414.9x_3,$$
$$\text{E} = 4.9.$$

Здесь E – средняя относительная ошибка аппроксимации модели.

Второй кластер:

$$S_2 = \{1, 5, 6, 12, 14\},$$
$$y = 1041263.3 - 1533.9x_1 + 239.1x_2 + 5238.5x_3,$$
$$\text{E} = 2.1.$$

Третий кластер:

$$S_3 = \{3, 7, 11, 13\}.$$

Заключение

В работе предложен алгоритм кластеризации данных, основанный на построении линейной регрессионной модели с применением метода антиробастного оценивания параметров, который обладает свойством равенства числа максимальных по модулю ошибок аппроксимации числу параметров плюс единица. Проведена кластеризация выборки данных при моделировании объема погрузки на железнодорожном транспорте на основе статистической информации за 2005–2018 гг. по Российской Федерации.

Литература

1. Ellefsen K. J., Smith D. B. Manual Hierarchical Clustering of Regional Geochemical Data Using a Bayesian Finite Mixture Model // Applied Geochemistry. 2016. Vol. 75. P. 200–210.

2. Yun U., Ryang H., Kwon O. C. Monitoring Vehicle Outliers Based on Clustering Technique // Applied Soft Computing Journal. 2016. Vol. 49. P. 845–860.

3. Sumiyana S., Atmini S., Sugiri S. Predictive Power of Aggregate Corporate Earnings and their Components for Future GDP Growths: An International Comparison // Economics and Sociology. 2019. Vol. 12, Is. 1. P. 125–142.

4. Hirano K., Wright J. H. Forecasting With Model Uncertainty: Representations and Risk Reduction // Econometrica. 2017. Vol. 85. P. 617–643.

5. Han L., Tan K. M., Yang T., Zhang T. Local Uncertainty Sampling for Large-Scale Multiclass Logistic Regression // Annals of Statistics. 2020. Vol. 48, Is. 3. P. 1770–1788.

6. Бодянский Е. В., Винокурова Е. А., Пелешко Д. Д. Информационная технология кластеризации данных в условиях короткой обучающей выборки на основе ассоциативной вероятностной нейро-фаззи системы // Управляющие системы и машины. 2014. № 4. С. 73–76.

7. Кирилюк И. Л., Сенько О. В. Оценка качества кластеризации панельных данных с использованием методов Монте-Карло (на примере данных российской региональной экономики) // Компьютерные исследования и моделирование. 2020. Т. 12, № 6. С. 1501–1513.

8. Уткин Л. В., Жук Ю. А. Робастная модификация метода k-средних для кластеризации интервальных данных // Известия Санкт-Петербург. лесотехнич. акад. 2016. № 216. С. 255–266.

9. Анищенко В. В., Вятченин Д. А., Доморацкий А. В., Тати Р., Фисенко В. К. Метод быстрого прототипирования систем нечеткого вывода при неизвестном числе классов // Искусственный интеллект. 2013. № 3. С. 307–315.

10. Кирпичников А. П., Ризаев И. С., Тахавова Э. Г., Сафаров Н. И. Разработка эффективного алгоритма иерархической кластеризации // Вестник Технолог. ун-та. 2019. Т. 22, № 10. С. 117–122.

11. Островский А. А. Реализация параллельного выполнения алгоритма FCM-кластеризации // Прикладная информатика. 2009. № 2. С. 101–106.

12. Аверченков В. И., Казаков П. В. Эволюционный метод поиска оптимальных решений для задач со множеством экстремумов // Вестник компьютерных и информационных технологий. 2010. № 5. С. 3–11.

13. Емельянов В. В., Курейчик В. В., Курейчик В. М. Теория и практика эволюционного моделирования. М. : Физматлит, 2003. 431 с.

14. Курейчик В. М. Генетические алгоритмы и их применение. 2-е изд., перераб. и доп. Таганрог : Изд-во ТРТУ, 2002. 242 с.

15. Гладков Л. А., Зинченко Л. А., Курейчик В. В. и др. Методы генетического поиска. Таганрог : Изд-во ТРТУ, 2002. 122 с.

16. Пупков К. А., Феоктистов В. А. Алгоритм дифференциальной эволюции для задач технического проектирования // Информационные технологии. 2004. № 8. С. 25–31.

17. Носков С. И., Ильюшонок Д. М. Подход к кластеризации выборки данных на основе метода наименьших модулей // Южно-Сибир. науч. вестн. 2020. № 6. С. 255–259.

18. Лакеев А. В., Носков С. И. Метод наименьших модулей для линейной регрессии: число нулевых ошибок аппроксимации // Современные технологии. Системный анализ. Моделирование. 2012. № 2 (34). С. 48–50.

19. Носков С. И. Компромиссные паретовские оценки параметров линейной регрессии // Математическое моделирование. 2020. Т. 32, № 11. С. 70–78.

20. Носков С. И. Метод антиробастного оценивания параметров линейной регрессии: число максимальных по модулю ошибок аппроксимации // Южно-Сибир. науч. вестн. 2020. № 1. С. 51–54.

21. Базилевский М. П. Отбор информативных операций при построении линейно-неэлементарных регрессионных моделей // International Journal of Open Information Technologies. 2021. Т. 9, № 5. С. 30–35.

22. Базилевский М. П. Многокритериальный подход к построению моделей парно-множественной линейной регрессии // Известия Саратов. ун-та. Сер.: Математика. Механика. Информатика. 2021. Т. 21, № 1. С. 88–99.

23. Базилевский М. П. Построение степенно-показательных регрессионных моделей и их интерпретация // Вестник Воронеж. гос. ун-та. Сер. Системный анализ и информ. технологии. 2020. № 4. С. 19–28.

24. Носков С. И. Точечная характеризация множества Парето в линейной многокритериальной задаче // Современные технологии. Системный анализ. Моделирование. 2008. № 1. С. 99–101.

25. Носков С. И., Перфильева К. С. Моделирование объема погрузки на железнодорожном транспорте методом смешанного оценивания // Известия Тульск. гос. ун-та. Технич. науки. 2021. № 2. С. 148–153.