УДК 519.23
DOI 10.34822/1999-7604-2021-4-78-82

УСТОЙЧИВОСТЬ АЛГОРИТМОВ ОТБОРА ПРИЗНАКОВ К ОШИБКАМ ВТОРОГО РОДА

А. Д. Черемухин
Нижегородский государственный инженерно-экономический университет, Княгинино, Россия
E-mail: ngieu.cheremuhin@yandex.ru

Предметом исследования является эффективность работы алгоритмов отбора признаков применительно к задачам регрессии в контексте частоты выявления ими ложных статистически значимых зависимостей. Целью стало построение соответствующей методики и апробация ее на сгенерированных данных, а также проверка гипотезы о наличии частоты появления ошибок второго рода от распределения зависимой переменной. Проведено изучение 7 методов отбора признаков: Simulated Annealing, Select Difference, Hill-Climbing, Las Vegas, Sequential Forward Selection, Select Slope, Whale Optimization. В качестве зависимых переменных выбраны переменные, которые подчинялись 8 видам распределений (бета, Коши, экспоненциальное, гамма, логнормальное, нормальное, равномерное, Вейбулла). Установлено, что при строгом подходе к оценке качества моделей вероятность использования в практической деятельности ложнозначимых моделей невелика.

Ключевые слова: отбор признаков, регрессия, ложноположительный результат, статистическая значимость.

RESISTANCE OF THE ALGORITHMS FOR FEATURE SELECTION TO TYPE II ERRORS

A. D. Cheremukhin
Nizhny Novgorod State University of Engineering and Economics, Knyaginino, Russia
E-mail: ngieu.cheremuhin@yandex.ru

The subject of the study is the efficiency of feature selection algorithms in relation to regression problems in the context of frequency of detecting false statistically significant dependencies. The aim of the study is to build a suitable methodology and to test it on generated data, to test the hypothesis of frequency of occurrence of type II errors from distribution of dependent variable. In total, 7 methods of feature selection were studied in the work: Simulated Annealing, Select Difference, Hill-Climbing, Las Vegas, Sequential Forward Selection, Select Slope, Whale Optimization. Variables distributed according to 8 laws (Beta, Cauchy, exponential, Gamma, log-normal, normal, uniform, Weibull) were chosen as dependent variables. As a result of the study, it was found that the probability of using practical false-valued models is small using a rigorous approach in assessing the quality of models.

Keywords: feature selection, regression, false-positive result, statistical significance.

Введение

Значительная часть задач по анализу больших данных в экономике, медицине, общественных науках, технике решается сегодня с применением различных алгоритмов регрессии, кластеризации и классификации, при этом практическая ориентированность этих задач зачастую вынуждает жертвовать теоретической обоснованностью их решения.

Частое отсутствие точной и глубокой содержательной модели исследуемого явления обусловливает неопределенность в части факторов, влияющих на рассматриваемые показа-

тели. Это привело к развитию такого специфического направления анализа данных, как отбор признаков (feature selection), который в условиях их большого количества помогает выбрать существенно влияющие показатели без детального теоретического описания явления.

Методы анализа данных очень полезны при решении актуальных задач отбора факторов, влияющих на скорость распространения Covid-19 [1–2], их активно применяют при выборе фактора, влияющего на экспрессию гена из сотен или тысяч вариантов, в генетике [3–4], для прогнозирования урожая разных культур в Индии [5], прогнозирования продаж в целом [6], а также в психологических и социологических исследованиях [7–8]. Кроме того, использование данных методов позволяет существенно снизить количество вычислительных ресурсов при построении моделей регрессии и классификации [9, 10].

Большая часть современных исследований, посвященных разработке и анализу эффективности методов анализа данных, подходит к измерению их результативности с точки зрения true positive в терминах классификации: насколько эффективно метод обнаруживает уже существующую зависимость. В работе предлагается иная, дополняющая формулировка задачи: анализ методов анализа данных с точки зрения ошибок второго рода (насколько часто этот алгоритм находит зависимость, которой нет).

Активно совершенствуются и методы отбора признаков, из последних работ можно отметить, например:

- развитие методов lasso [11], scad [12], lars [13] и их применение в задачах конечной смеси регрессий (когда в разных субпопуляциях возможен разный состав влияющих факторов) [14];
- модифицированный генетический алгоритм отбора признаков с одновременным применением методов уменьшения количества атрибутов из исходного набора данных и feature selection [5];
- модификацию алгоритма типа simulated annealing [jdss], а также сравнение его эффективности с классическими методами [15] на основе специфичного набора тестовых функций;
- применение метода «случайного леса» для изменения метода многоцелевого поиска ENORA [6];
- совместное применение методов сокращения размерности (например, метода главных компонент) и отбора признаков [16].

Однако в настоящее время существует проблема доступа ко многим модификациям алгоритмов для сравнения их эффективности. Авторы используют разные языки программирования, зачастую не размещают исходный код в открытом доступе, что ограничивает воспроизводимость расчетов.

Материалы и методы

Для проведения исследования были выбраны следующие классические методы отбора признаков, реализованные в пакете FSinR [17] языка R:

1. Simulated Annealing – вероятностный метод аппроксимации глобального оптимума функции в дискретном пространстве, который применяется, если примерный поиск глобального оптимума приоритетнее, чем точный поиск локального оптимума [18].

2. Select Difference – метод прямого поиска, который выбирает признаки в модель на основе определенной метрики качества, пока разница оценок для какой-то пары отсортированных признаков не превысит определенного значения.

3. Hill-Climbing – метод локального поиска, разработанный в рамках численного анализа, имеющий ограничения в нахождении алгоритмом локального оптимума для невыпуклых задач [19].

4. Las Vegas – вид вероятностного алгоритма, гарантирующий достижение заданного результата на основе применения проверки исходного алгоритма на корректность [20].

5. Sequential Forward Selection – метод, основанный на постепенном добавлении признаков во множество выбранных для наибольшего увеличения метрик качества [21].

6. Select Slope – метод прямого поиска, который выбирает множество признаков, пока величина прироста определенной метрики качества не превысит определенного значения.

7. Whale Optimization – метод, основанный на технологии «пузырчатой сети», на основе которой строят охоту горбатые киты [22].

Основной метрикой качества для всех этих методов выбран коэффициент детерминации. Задачей исследования являлся анализ устойчивости вышеописанных алгоритмов отбора признаков к ошибкам 2-го рода и зависимость параметров устойчивости от распределения зависимой переменной. Все расчеты проводились с использованием языка R (находится в открытом доступе: URL: https://github.com/acheremuhin/fs_R_error2type_article).

Методика проведения исследования состояла из следующих этапов:

1. Генерировали 16 переменных, которые подчинялись 1 из 8 видов распределений (бета, Коши, экспоненциальное, гамма, логнормальное, нормальное, равномерное, Вейбулла) со случайными параметрами данных распределений таким образом, чтобы две переменных подчинялись каждому распределению. Каждая переменная включала в себя 1 000 наблюдений.

2. Затем выбирали одну из переменных в качестве зависимой, а остальные – в качестве независимых и строили 7 уравнений регрессии после применения каждого из методов отбора признаков.

3. Каждое из уравнений регрессии тестировали на достоверность с помощью интегральной оценки, которая складывалась из следующих параметров:

- доля значимых на 5 %-м уровне коэффициентов в полученном уравнении регрессии;
- наличие значимости на 5 %-м уровне всей модели в целом.

Следовательно, каждая построенная модель регрессии получала индикатор достоверности – долю статистически значимых параметров от общего их числа.

4. Далее алгоритм повторялся 10 раз, и значение индикатора достоверности усреднялось.

Таким образом, в результате применения методики получены средние значения индикаторов достоверности в системе координат «вид распределения зависимой переменной – используемый метод отбора признаков» (табл.).

Таблица

Средние значения индикаторов достоверности

Распределение	Метод отбора признаков						
	Simulated Annealing	**Select Difference**	**Hill-Climbing**	**Las Vegas**	**Sequential Forward Selection**	**Select Slope**	**Whale Optimization**
Бета	0,104	0,076	0,076	0,128	0,094	0,129	0,146
Коши	0,042	0,053	0,065	0,111	0,029	0,071	0,082
Экспоненциальное	0,174	0,088	0,118	0,098	0,082	0,094	0,141
Гамма	0,171	0,112	0,106	0,142	0,135	0,106	0,184
Логнормальное	0,072	0,041	0,065	0,057	0,059	0,071	0,059
Нормальное	0,146	0,059	0,088	0,101	0,071	0,047	0,104
Равномерное	0,107	0,129	0,106	0,094	0,059	0,047	0,104
Вейбулла	0,131	0,076	0,1	0,152	0,094	0,112	0,107

Примечание: составлено авторами на основании собственных расчетов.

Анализ таблицы позволяет сделать вывод, что при определенном распределении зависимой переменной наименьшие показатели индикаторов достоверности (т. е. наименьшую вероятность ошибок второго рода) дают следующие алгоритмы:

- при бета-распределении – алгоритмы Select Difference и Hill-Climbing;
- при распределении Коши – алгоритм Sequential Forward Selection;
- при экспоненциальном распределении – алгоритм Sequential Forward Selection;

- при гамма-распределении – алгоритм Hill-Climbing и Select Slope (при этом для данного типа распределений велика вероятность ошибок второго рода);
- при логнормальном распределении – алгоритм Select Difference;
- при нормальном или равномерном распределении – алгоритм Select Slope;
- при распределении Вейбулла – алгоритм Select Difference.

В целом можно констатировать, что для всех рассмотренных алгоритмов вероятность построения ошибочно статистически значимой модели невелика. При этом сами вероятности различаются значительно и зависят как от распределения зависимого показателя, так и от применяемого метода.

Обсуждение и заключение

По результатам исследования методов отбора признаков на вероятность ошибки второго рода (т. е. на построение ошибочно значимых статистических моделей) установлено, что существует незначительная вероятность ошибки второго рода выбранных методов отбора признаков и видов распределений.

Однако при решении реальной задачи, значительно отличной от 0, остается вероятность, что в условиях истинной независимости и некореллированности переменных между собой может быть построена статистически значимая модель с одним или двумя статистически значимыми коэффициентами при незначимости всех остальных. В этих условиях практическое применение моделей после фазы отбора признаков должно соответствовать принципу использования моделей только со всеми статистически значимыми коэффициентами. Для этого при использовании метода наименьших квадратов как способа оценки коэффициентов необходимо проводить исследования и на выполнимость условий Гаусса – Маркова.

В качестве дальнейших направлений развития данного исследования можно выделить иную трактовку сущности ошибок второго рода для алгоритмов отбора признаков, построение методики полного исследования алгоритмов отбора признаков на наличие таких ошибок, а также других алгоритмов отбора признаков с разными входными условиями.

Литература

1. Djordjevic M., Salom I., Markovic S., Rodic A., Milicevic O., Djodrjevic M. Inferring the Main Drivers of SARS-Cov-2 Global Transmissibility by Feature Selection Methods // GeoHealth. 2021. Vol. 5, Is. 9. P. e2021GH000432.

2. Kaliappan J., Srinivasan K., Qaisar S. M., Sundararajan K., Chang C.-Y., C S. Performance Evaluation of Regression Models for the Prediction of the COVID-19 Reproduction Rate // Front Public Health. 2021. Vol. 9. P. 729795.

3. Conlon E. M., Liu X. S., Lieb J. D., Liu J. S. Integrating Regulatory Motif Discovery and Genome-Wide Expression Analysis // Proc Natl Acad Sci USA. 2003. Vol. 100, No. 6. P. 3339–3344.

4. Zhong W., Zeng P., Ma P., Liu J. S., Zhu U. RSIR: Regularized Sliced Inverse Regression for Motif Discovery // Bioinformatics. 2005. Vol. 21, No. 22. P. 4169–4175.

5. Shastry K. A., Sanjay H. A. A Modified Genetic Algorithm and Weighted Principal Component Analysis Based Feature Selection and Extraction Strategy in Agriculture // Knowledge-Based Systems. 2021. Vol. 232. P. 107460.

6. Jiménez F., García J. M., Sciavicco G., Pechuán L. M. Multi-Objective Evolutionary Feature Selection for Online Sales Forecasting // Neurocomputing. 2017. Vol. 234. P. 75–92. URL: http://dx.doi.org/10.1016/j.neucom.2016.12.045 (дата обращения: 02.10.2021).

7. Blesser B. A., Kuklinski T. T., Shillman R. J. Empirical Tests for Feature Selection Based on a Psychological Theory of Character Recognition // Pattern Recognition. 1976. Vol. 8, Is. 2. P. 77–85.

8. Tang J., Liu. H. Feature Selection for Social Media Data // ACM Trans Knowl Discov Data. 2014. Vol. 8, Is. 4. P. 19.

9. Gao Y., Xu A., Hu P. J.-H., Cheng T.-H. Incorporating Association Rule Networks in Feature Category-Weighted Naïve Bayes Model to Support Weaning Decision Making // Decis Support Syst. 2017. Vol. 96. P. 27–38.

10. Yuan H., Lau R. Y. K., Xu W. The Determinants of Crowdfunding Success: A Semantic Text Analytics Approach // Decis Support Syst. 2016. Vol. 91. P. 67–76.

11. Tibshirani R. Regression Shrinkage and Selection via the Lasso // J R Statist Soc B. 1996. Vol. 58, No. 1. P. 267–288.

12. Fan J., Li R. New Estimation and Model Selection Procedures for Semiparametric Modeling in Longitudinal Data Analysis // J Amer Statist Assoc. Vol. 99, No. 467. P. 710–723.

13. Efron B., Hastie T., Johnstone I., Tibshirani R. Least Angle Regression (with Discussion) // Ann Statist. 2004. Vol. 32, Is. 2. P. 407–499.

14. Khalili A. An Overview of the New Feature Selection Methods in Finite Mixture of Regression Models // JIIRS. 2011. Vol. 10, Is. 2. P. 201–235.

15. Zhang L., Mistry K., Lim C. P., Neoh S. C. Feature Selection Using Firefly Optimization for Classification and Regression Models // Decis Support Syst. 2018. Vol. 106. P. 64–85.

16. Shang R., Chang J., Jiao L., Xue Y. Unsupervised Feature Selection Based on Self-Representation Sparse Regression and Local Similarity Preserving // International Journal of Machine Learning and Cybernetics. 2019. Vol. 10. P. 757–770.

17. Aragón-Royón F., Jiménez-Vílchez A., Arauzo-Azofra A., Benitez J. M. FSinR: An Exhaustive Package for Feature Selection // arXiv:2002.10330 [cs.LG]. 2020. URL: https://arxiv.org/abs/2002.10330 (дата обращения: 02.10.2021).

18. Posario F., Thangadurai K. Simulated Annealing Algorithm for Feature Selection // International Journal of Computers & Technology. 2016. Vol. 15, No. 2. P. 6471–6479.

19. Gelbart D., Morgan N., Tsymbal A. Hill-Climbing Feature Selection for Multi-Stream ASR // INTERSPEECH 2009. URL: https://www.icsi.berkeley.edu/pubs/speech/gelbart-2009.pdf (дата обращения: 02.10.2021).

20. Nandy G. An Enhanced Approach to Las Vegas Filter (LVF) Feature Selection Algorithm // 2nd National Conference on Emerging Trends and Applications in Computer Science. 2011. P. 1–3. URL: https://ieeexplore.ieee.org/document/5751392 (дата обращения: 02.10.2021).

21. Marcano-Cedeño A., Quintanilla J., Cortina-Januchs G., Andina D. Feature Selection Using Sequential Forward Selection and Classification Applying Artificial Metaplasticity Neural Network // 36th Annual Conference on IEEE Industrial Electronics Society. 2010. P. 2845–2850. URL: https://ieeexplore.ieee.org/document/5675075 (дата обращения: 02.10.2021).

22. Zamani H., Nadimi-Shahraki M. H. Feature Selection Based on Whale Optimization Algorithm for Diseases Diagnosis // International Journal of Computer Science and Information Security. 2016. Vol. 14. P. 1243–1247.