УНИВЕРСИТЕТ ИТМО
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
сентябрь–октябрь 2020 Том 20 № 5 ISSN 2226-1494 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
September–October 2020 Vol. 20 No 5 ISSN 2226-1494 http://ntv.ifmo.ru/en/
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК УНИВЕРСИТЕТ ИТМО ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

УДК 519.712.2 doi: 10.17586/2226-1494-2020-20-5-755-760

ПРИМЕНЕНИЕ МЕТОДА *K*-СРЕДНИХ В ЗАДАЧЕ ОЦЕНКИ ХАРАКТЕРИСТИК ПРОЦЕССА ДЛЯ ВЕБ-ПРИЛОЖЕНИЙ

В.В. Евстратов, М.С. Ананьевский

Санкт-Петербургский государственный университет, Санкт-Петербург, 199034, Российская Федерация
Адрес для переписки: viktor.evst@gmail.com

Информация о статье
Поступила в редакцию 01.06.20, принята к печати 15.07.20
Язык статьи — русский

Ссылка для цитирования: Евстратов В.В., Ананьевский М.С. Применение метода *K*-средних в задаче оценки характеристик процесса для веб-приложений // Научно-технический вестник информационных технологий, механики и оптики. 2020. Т. 20. № 5. С. 755–760. doi: 10.17586/2226-1494-2020-20-5-755-760

Аннотация

Предмет исследования. Исследованы подходы к решению задачи оценки характеристик процесса на примере задачи прогнозирования характеристик активности пользователей в компьютерных онлайн-играх. Рассмотрены методы машинного обучения и определены потенциальные преимущества алгоритмов кластеризации в применении к рассматриваемой задаче. Исследованы различные метрики качества алгоритмов кластеризации. **Метод.** На основе гипотезы, возникшей в ходе предварительного анализа данных о пользовательской активности, разработан подход к оценке характеристик процесса, использующий кластеризацию. Собраны данные об активности пользователей, для которых уже известны значения прогнозируемых характеристик. Каждый пользователь представлен в виде пары векторов: первый вектор составлен из его характеристик в первые дни активности, второй – из прогнозируемых характеристик. Векторы, представляющие пользователей в первые дни активности, используются в качестве обучающей выборки для алгоритма *K*-средних. За подбор параметра *K* отвечает специально разработанный функционал энтропийного вида, адекватный исследуемой задаче. Выделенным кластерам ставятся в соответствие усредненные по попавшим в них пользователям векторы прогнозируемых характеристик. Эти соответствия используются в качестве прогнозов характеристик для новых пользователей. **Основные результаты.** Предложен ориентированный на рассмотренный тип данных метод оценки качества кластеризации, позволяющий выбрать наиболее подходящее для целевой задачи число кластеров. Проведен численный эксперимент, демонстрирующий применимость разработанного метода. **Практическая значимость.** Применение предложенного подхода позволяет прогнозировать одновременно несколько характеристик пользователей компьютерных онлайн-игр, и, таким образом, решать различные прикладные задачи планирования и аналитики, возникающие в ходе разработки. Например, изложенный в статье метод был использован в задачах анализа окупаемости разработки новых игровых элементов и прогнозирования нагрузки на серверы с целью заблаговременного наращивания мощностей. Его преимуществами являются отсутствие необходимости экспертной разметки обучающей выборки и относительно невысокие требования к вычислительным мощностям – в силу низкой вычислительной сложности функционала качества для подбора гиперпараметра *K*.

Ключевые слова
кластеризация, *K*-Means, алгоритм *K*-средних, оценка качества кластеризации, энтропия, машинное обучение, алгоритмы, веб

Благодарности
Исследование выполнено при финансовой поддержке Российского фонда фундаментальных исследований в рамках научного проекта (грант № 19-08-00865 А).

doi: 10.17586/2226-1494-2020-20-5-755-760

PROCESS CHARACTERISTICS ESTIMATION IN WEB APPLICATIONS USING *K*-MEANS CLUSTERING

V.V. Evstratov, M.S. Ananyevskiy

Saint Petersburg State University, Saint Petersburg, 199034, Russian Federation
Corresponding author: viktor.evst@gmail.com

Article info
Received 01.06.20, accepted 15.07.20
Article in Russian

For citation: Evstratov V.V., Ananyevskiy M.S. Process characteristics estimation in web applications using *K*-means clustering. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2020, vol. 20, no. 5, pp. 755–760 (in Russian). doi: 10.17586/2226-1494-2020-20-5-755-760

Abstract

Subject of Research. The paper presents the study of estimation problem of process characteristics for the particular case of user's activity prediction in computer online games. Various machine learning methods are considered, and the advantages of clustering-based approaches are identified. The variety of metrics for the estimation of clustering quality is studied. **Method.** A clustering-based approach to estimation of process characteristics was developed on the base of a hypothesis proposed during the preliminary analysis of user's activity data. Data on activity of users with the known predicted values was collected. Each user was represented as a pair of vectors: the first vector corresponded to his first days of activity, and the second one corresponded to the days with predicted performance. The vectors representing user's activity in the first days were used as training data for the *K*-means algorithm. A developed entropy-like loss function was used to find a value of *K* suitable for the problem under consideration. The clusters were matched with vectors of predicted process characteristics averaged over all users in the cluster. These matches were used as the prediction of new users' characteristics. **Main Results.** An approach to the determination of the suitable number of clusters is proposed, taking into account the specifics of the considered data. Numerical experiment is carried out, demonstrating the applicability of the developed method. **Practical Relevance.** The proposed approach application allows for the simultaneous prediction of multiple characteristics of online-game users, and, therefore, for solution of various planning and analytics problems during online-game development. For example, the method developed in the present work was used to analyze the development payback of new game elements, and to predict server load in order to increase available computational resources beforehand. The advantages of the developed method include no need for expert tagging of the training set and relatively low computational cost due to the low computational complexity of the proposed loss function used to estimate the hyperparameter *K*.

Keywords

clustering, *K*-means, *K*-means algorithm, clustering quality assessment, entropy, machine learning, algorithms, web

Acknowledgements

This study has been supported by the Russian Foundation for Basic Research, grant no. 19-08-00865 A.

Введение

В онлайн-сервисах часто возникают задачи, решение которых сводится к прогнозированию пользовательской активности. Например, оценка будущей нагрузки на серверы, построение системы обнаружения случаев мошенничества (fraud detection) или прогноз выручки компании. Предлагаемый в работе подход был использован при построении системы предсказания характеристик посетителей веб-сервиса.

В качестве *процесса* рассматривалась последовательность действий участников: вход в игру, победа или проигрыш, покупка внутриигровых предметов, выход из игры, и др. В качестве *характеристик процесса* рассматривались различные величины, описывающие пользовательскую активность — такие как время, проведенное в игре, и сумма внутриигровых покупок за различные периоды. По характеристикам процесса за прошедший период необходимо *оценить* их будущие значения, т. е. необходимо построить систему прогнозирования пользовательской активности.

Методы машинного обучения часто находят применение в различных веб-приложениях. В условиях растущих объемов данных они иногда оказываются единственным эффективным подходом. Рассматриваемая задача оценки характеристик процесса в дискретном времени может быть формализована несколькими способами, такими как задача линейной регрессии [1, 2], классификации [3–5] или кластеризации [6, 7]. Каждый из этих способов ранее применялся в отечественной и зарубежной практике; и выбор конкретного алгоритма, а также выбор представления данных позволяют исследовать задачу с разных сторон.

Подход на основе классификации хорошо исследован и успешно применяется в веб-приложениях [8, 9], но его серьезным ограничением является необходимость разметки данных, возможность которой не рассматривалась авторами настоящей статьи ввиду большого объема данных. Распространенный подход при таких ограничениях — применение методов обучения без учителя, к числу которых относится алгоритм кластеризации *K*-Means.

Дополнительным вызовом и мотивацией для авторов стали: необходимость использования легко интерпретируемого подхода и ограниченность доступных вычислительных ресурсов. Как правило, на ранних стадиях проекта внедрение новых подходов требует наглядного и убедительного обоснования для лиц, принимающих решения, поэтому авторами не рассматривалась возможность применения методов на основе искусственных нейронных сетей, ранее применявшихся в схожих задачах [10]. По аналогичным причинам не были использованы и готовые аналитические решения, например, [11].

Подход

Как было упомянуто выше, в результате предварительного анализа данных была выдвинута следующая гипотеза: если два дискретных во времени процесса «похожи» на своих первых *m* отсчетах (шагах), то они будут оставаться «похожими» и в будущем. В таком случае прогнозировать характеристики процессов можно было бы проведя кластеризацию по характеристикам их начальных отсчетов, а в качестве прогноза использовать усредненные по кластеру будущие характеристики.

Такой подход предоставляет возможность не только предсказывать характеристики процессов, но и одновременно производить разведочный анализ данных (exploratory data analysis) с помощью построенной

структуры (кластеров), а также снизить влияние шума в данных, часто встречающегося в веб-приложениях [12, 13].

Кластеризация

Кластеризация — это задача разбиения заданной выборки объектов на непересекающиеся подмножества, называемые кластерами, так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались.

В качестве алгоритма кластеризации использован алгоритм *K*-средних (*K*-Means) [14]. Коротко работу *K*-Means можно описать так: по заданному количеству кластеров происходит итеративный поиск набора точек (центроидов), вокруг которых наилучшим образом группируются кластеры. Новый объект относится к тому или иному кластеру на основании того, к какому центроиду он ближе согласно заданной мере сходства.

Чтобы воспользоваться *K*-Means, требуется подготовка данных: во-первых, необходимо представить обучающую выборку в виде набора векторов, представляющих объекты в пространстве признаков; во-вторых — выбрать меру сходства объектов. Пусть доступны m отсчетов для l характеристик каждого процесса. Например, могут быть известны данные m первых дней пользователя на сервисе, и l — число прогнозируемых величин. Так, если известны суммарная длительность сессий пользователя и его траты за день, то $l = 2$. Значит, процессы можно представить в виде $n = lm$-мерных векторов. В качестве метрики сходства процессов выбрана L_n-норма разности соответствующих векторов. Эксперименты с L_j нормами, где $j \leqslant 2$, показали неудовлетворительные результаты.

Формализация задачи

Пусть дан набор дискретных по времени процессов, каждый из которых представлен последовательностью из m векторов размерности l вида

$$\mathbf{x}^{(t)} = (\mathbf{x}_1{}^{(t)}, \ldots, \mathbf{x}_l{}^{(t)})^T, \; t \in [1, m].$$

Для удобства введем обозначение $n = ml$, и заметим, что каждый процесс тогда можно представить как вектор $\mathbf{x} \in R^n$, полученный путем конкатенации векторов $\mathbf{x}^{(t)}$ для всех $t \in [1, m]$.

Отметим также, что для каждого процесса известен l-мерный целевой вектор $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_l)^T$, представляющий собой будущие значения характеристик процесса, которое необходимо оценить. В результате кластеризации обучающей выборки из M процессов $train = \{\mathbf{x}_i, \mathbf{y}_i\}_1{}^M$ с количеством кластеров, равным K, получим набор центроидов

$$centroids = \{C_1, \ldots, C_K\},$$

где $C_i \in R^n$.

При необходимости определить кластер (отождествляя его со своим центроидом), в который попадает некий новый вектор $\mathbf{x}'$, полученный по формуле:

$$c(x) = \operatorname{argmin}_{C \in centroid} \|\mathbf{x} - C\|_n,$$

где $\|\mathbf{x}\|_n = (\sum_{i=1}^{n} |\mathbf{x}_i|^n)^{1/n} - L_n$ — норма вектора $\mathbf{x}$.

Далее, в соответствии с предложенным подходом, строится оценка $\mathbf{y}$ для вектора $\mathbf{x}'$ как усреднение целевых векторов из того же кластера:

$$\hat{\mathbf{y}}(\mathbf{x}) = \overline{\{\mathbf{y} | (\mathbf{x}, \mathbf{y}) \in train, c(\mathbf{x}) = c(\mathbf{x}')\}}.$$

Выбор, как слишком большого количества кластеров K, так и слишком малого, приведет к снижению предсказательной способности введенной оценки $\hat{\mathbf{y}}$ [15]. Таким образом, для эффективного решения изложенной задачи необходимо разработать функционал, который позволит численно оценить качество проведенной кластеризации.

Известные методы оценки качества кластеризации

Одним из наиболее часто применяемых на практике методов оценки качества кластеризации является эвристика Elbow [16]. В соответствии с этим методом оптимальным значением параметра K является «точка перегиба» графика зависимости суммы квадратов расстояний элементов кластеров до своих центроидов (sum of squared estimate of errors или SSE) от числа кластеров K, которую предполагается оценивать визуально. Этот метод обладает двумя существенными недостатками. Во-первых, для его использования необходимо участие эксперта. Во-вторых, описанные графики не всегда поддаются однозначной трактовке. Пример такого неоднозначного графика, полученного авторами в ходе экспериментов, приведен на рис. 1. На графике видно несколько точек-кандидатов, а правила выбора в таком случае не регламентированы методом.

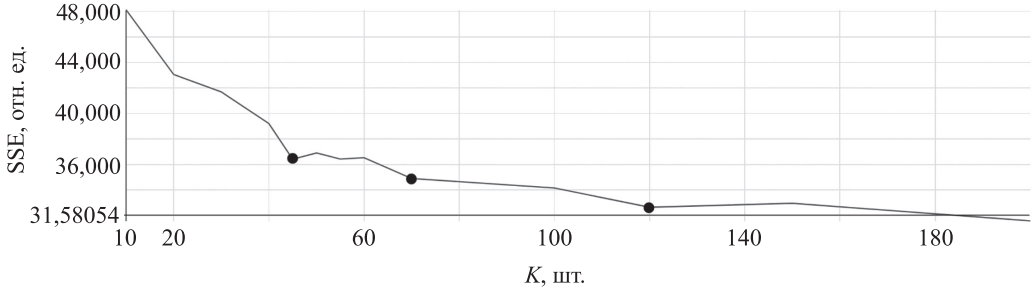

Рис. 1. График для эвристики Elbow. На графике выделены точки-кандидаты для выбора количества кластеров K. Жирным выделено минимальное значение SSE

Существуют и другие методы оценки качества кластеризации, один из них — Silhouette [17, 18]. К числу достоинств (данного метода) можно отнести необязательность участия эксперта и лучшая обоснованность. Существенным ограничением этого метода является его высокая вычислительная сложность: $O(n^2)$. В силу больших объемов данных, в настоящем исследовании от использования Silhouette пришлось также отказаться. По аналогичным причинам не были применены и некоторые другие известные методы.

Энтропия кластеризации

Построим функционал качества кластеризации, удовлетворяющий нуждам рассматриваемой задачи. Для этого проанализируем полученный в ходе кластеризации набор кластеров. Каждый из них может быть вписан в n-мерный шар. Чем больше объем такого шара относительно других, тем, в геометрическом смысле, выше вероятность того, что новый объект попадет в него. Заметим, что для введенной ранее оценки $\hat{\mathbf{y}}$ верно, что большой разброс радиусов таких шаров приведет к определенным неудобствам использования предложенной модели на практике. Подтверждение данному тезису приведено на рис. 2.

Точками показаны запуски *K*-Means с различными значениями параметра *K*. Как видно из графика, между данными величинами прослеживается линейная зависимость, т. е., чем меньше разнообразие радиусов полученных кластеров — тем меньше отличаются получаемые в них ошибки оценки $\hat{\mathbf{y}}$. На практике это означает, что прогнозы, полученные для различных кластеров, будут неравноценны.

Нормированные объемы таких шаров можно использовать в качестве вероятностей для построения функционала энтропийного вида. В дальнейшем, подобрав число кластеров *K*, максимизирующее такой функционал, можно получить кластеризацию, разбивающую исходное множество векторов на кластеры с наиболее близкими объемами описывающих их шаров — что и требовалось.

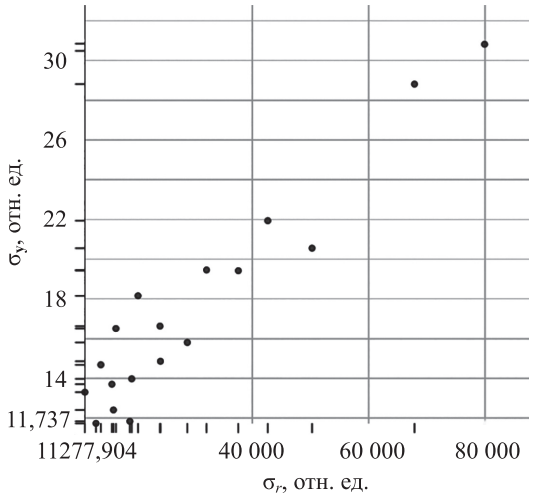

Рис. 2. Стандартные отклонения невязки прогноза $\sigma_{\mathbf{y}} = \|\mathbf{y} - \hat{\mathbf{y}}\|_n$ и радиусов полученных кластеров при различных значениях параметра *K*

Выпишем описанный функционал. Для этого введем энтропию кластеризации:

$$S = \sum_{i=1}^{K} -p_i \log(p_i),$$

где K — количество кластеров.

Вероятности p_i рассчитываются по формуле:

$$p_i = \overline{V_i} = V_i / \sum_{i=1}^{K} V_i,$$

где V_i — объем n-мерного шара, описывающего i-й кластер.

Для удобства расчетов радиусы кластеров предварительно нормированы:

$$\overline{R_i} = (R_i + 1) / \sum_{i=1}^{K} (R_i + 1).$$

Напомним, что объем n-мерного шара рассчитывается по формуле:

$$V_i = \frac{\pi^{n/2}}{\Gamma\left(\frac{n}{2} + 1\right)} \overline{R_i}^n,$$

где Γ — это эйлеровская гамма-функция.

Радиус же R_i n-мерного шара, описывающего кластер C_i, вычисляется как

$$R_i = \operatorname{argmax}_{x \in C_i} \|x - \overline{x_{C_i}}\|_n,$$

где $\|x\|_n = (\sum_{i=1}^{n} |x_i|^p)^{1/p}$ – L_n норма в R^n.

Как было продемонстрировано ранее, ключевым параметром в исследуемом подходе является то, насколько однородны размеры получаемых кластеров. Следовательно, количество кластеров, доставляющее максимум введенной энтропии кластеризации, будет в этом смысле оптимальным.

Эксперименты

Предложенное решение протестировано в задаче прогнозирования трат и суммарного времени активности пользователей компьютерной онлайн-игры на тридцатый день после регистрации по данным об их активности в первые пять дней. Таким образом, действия отдельно взятого пользователя рассматривались как процесс, дискретный по времени (с периодом дискретизации 1 день), а траты и суммарное время активности пользователя выступали в качестве оцениваемых характеристик. Каждый пользователь был представлен как десятимерный вектор, содержащий отсчеты времени, проведенного в приложении, и трат в каждый из первых пяти дней активности.

В данной задаче обучающая выборка представляла из себя набор пар векторов $\{(\mathbf{x}, \mathbf{y}) | \mathbf{x} \in R^{10}, \mathbf{y} \in R^2\}$, полученный по записям об активности 700 тысяч пользователей, зарегистрировавшихся в рассматриваемый период работы сервиса. Далее производился перебор значений параметра K в диапазоне от 1 до 200. На каждом шаге вычислялось значение предложенной выше энтропии.

Результаты моделирования представлены на рис. 3. Для удобства чтения график был сглажен скользящим

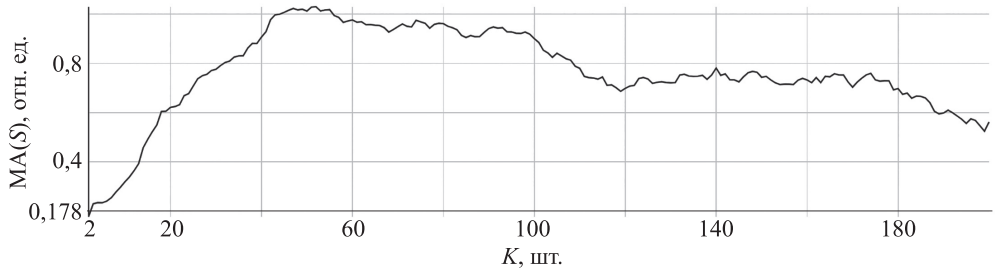

Рис. 3. Результаты моделирования, где K — количество кластеров; MA(S) — скользящее среднее значение энтропии кластеризации

средним. В окрестности 50 кластеров энтропия кластеризации достигает максимума, что и является искомым значением параметра K.

Заключение

В работе предложены подход к прогнозированию пользовательской активности и алгоритм выбора количества кластеров для алгоритма K-Means, лежащего в основе данного подхода. Выполнен численный эксперимент на реальных данных, продемонстрировавший применимость предложенного метода. Авторами не были использованы все доступные данные, например, информация о браузере, операционной системе и стране пребывания пользователя. Данное решение позволяет применять предложенный подход в аналогичных веб-приложениях в том случае, если доступ к подобным данным отсутствует, что делает метод более универсальным. В дальнейшем авторы планируют исследовать применимость и расширяемость подхода по отношению к дополнительным данным разного типа.

Литература

1. Zhang Z., Lai Z., Xu Y., Shao L., Wu J., Xie G.-S. Discriminative elastic-net regularized linear regression // IEEE Transactions on Image Processing. 2017. V. 26. N 3. P. 1466–1481. doi: 10.1109/TIP.2017.2651396
2. Olive D.J. Linear Regression. Springer, 2017. IX, 494 p. doi: 10.1007/978-3-319-55252-1
3. Xu J., Xu C., Zou B., Tang Y.Y., Peng J., You X. New incremental learning algorithm with support vector machines // IEEE Transactions on Systems, Man, and Cybernetics: Systems. 2019. V. 49. N 11. P. 2230–2241. doi: 10.1109/TSMC.2018.2791511
4. Press S.J., Wilson S. Choosing between logistic regression and discriminant analysis // Journal of the American Statistical Association. 1978. V. 73. N 364. P. 699–705. doi: 10.1080/01621459.1978.10480080
5. Friedman J., Hastie T., Tibshirani R. Additive logistic regression: A statistical view of boosting // Annals of Statistics. 2000. V. 28. N 2. P. 337–407. doi: 10.1214/aos/1016218223
6. Subramaniyaswamy V., Logesh R. Adaptive KNN based recommender system through mining of user preferences // Wireless Personal Communications. 2017. V. 97. N 2. P. 2229–2247. doi: 10.1007/s11277-017-4605-5
7. Cheung D.W., Kao B., Lee J. Discovering user access patterns on the World Wide Web // Knowledge-Based Systems. 1998. V. 10. N 7. P. 463–470. doi: 10.1016/S0950-7051(98)00037-9
8. Liu D.-S., Fan S.-J. A modified decision tree algorithm based on genetic algorithm for mobile user classification problem // The Scientific World Journal. 2014. P. 468324.
9. Santra A., Jayasudha S. Classification of web log data to identify interested users using Naïve Bayesian classification // International Journal of Computer Science Issues (IJCSI). 2012. V. 9. N 1. P. 381.
10. Park S., Suresh N.C., Jeong B.-K. Sequence-based clustering for web usage mining: A new experimental framework and ann-enhanced k-means algorithm // Data & Knowledge Engineering. 2008. V. 65. N 3. P. 512–543. doi: 10.1016/j.datak.2008.01.002
11. Medina-Ortiz D., Contreras S., Quiroz C., Asenjo J.A., Olivera-Nappa Á. DMAKit: A user-friendly web platform for bringing state-of-the-art data analysis techniques to non-specific users // Information Systems. 2020. V. 93. P. 101557. doi: 10.1016/j.is.2020.101557
12. Meroño-Peñuela A. Refining Statistical Data on the Web. CreateSpace Independent Publishing Platform, 2016. 252 p.

References

1. Zhang Z., Lai Z., Xu Y., Shao L., Wu J., Xie G.-S. Discriminative elastic-net regularized linear regression. *IEEE Transactions on Image Processing*, 2017, vol. 26, no. 3, pp. 1466–1481. doi: 10.1109/TIP.2017.2651396
2. Olive D.J. *Linear Regression.* Springer, 2017, IX, 494 p. doi: 10.1007/978-3-319-55252-1
3. Xu J., Xu C., Zou B., Tang Y.Y., Peng J., You X. New incremental learning algorithm with support vector machines. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2019, vol. 49, no. 11, pp. 2230–2241. doi: 10.1109/TSMC.2018.2791511
4. Press S.J., Wilson S. Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association*, 1978, vol. 73, no. 364, pp. 699–705. doi: 10.1080/01621459.1978.10480080
5. Friedman J., Hastie T., Tibshirani R. Additive logistic regression: A statistical view of boosting. *Annals of Statistics*, 2000, vol. 28, no. 2, pp. 337–407. doi: 10.1214/aos/1016218223
6. Subramaniyaswamy V., Logesh R. Adaptive KNN based recommender system through mining of user preferences. *Wireless Personal Communications*, 2017, vol. 97, no. 2, pp. 2229–2247. doi: 10.1007/s11277-017-4605-5
7. Cheung D.W., Kao B., Lee J. Discovering user access patterns on the World Wide Web. *Knowledge-Based Systems*, 1998, vol. 10, no. 7, pp. 463–470. doi: 10.1016/S0950-7051(98)00037-9
8. Liu D.-S., Fan S.-J. A modified decision tree algorithm based on genetic algorithm for mobile user classification problem. *The Scientific World Journal*, 2014, pp. 468324.
9. Santra A., Jayasudha S. Classification of web log data to identify interested users using Naïve Bayesian classification. *International Journal of Computer Science Issues (IJCSI)*, 2012, vol. 9, no. 1, pp. 381.
10. Park S., Suresh N.C., Jeong B.-K. Sequence-based clustering for web usage mining: A new experimental framework and ann-enhanced k-means algorithm. *Data & Knowledge Engineering*, 2008, vol. 65, no. 3, pp. 512–543. doi: 10.1016/j.datak.2008.01.002
11. Medina-Ortiz D., Contreras S., Quiroz C., Asenjo J.A., Olivera-Nappa Á. DMAKit: A user-friendly web platform for bringing state-of-the-art data analysis techniques to non-specific users. *Information Systems*, 2020, vol. 93, pp. 101557. doi: 10.1016/j.is.2020.101557
12. Meroño-Peñuela A. *Refining Statistical Data on the Web*. CreateSpace Independent Publishing Platform, 2016, 252 p.

13. Nithya P., Sumathi P. Novel pre-processing technique for web log mining by removing global noise and web robots // Proc. of the National Conference on Computing and Communication Systems (NCCCS 2012). 2012. P. 41–45. doi: 10.1109/NCCCS.2012.6412976
14. Kanungo T., Mount D.M., Netanyahu N.S., Piatko C.D., Silverman R., Wu A.Y. An efficient k-means clustering algorithm: Analysis and implementation // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2002. V. 24. N 7. P. 881–892. doi: 10.1109/TPAMI.2002.1017616
15. Yang S.-L., Li Y.-S., Hu X.-X., Pan R.-Y. Optimization study on k value of k-means algorithm // Xitong Gongcheng Lilun yu Shijian/ System Engineering Theory and Practice. 2006. V. 26. N 2. P. 97–101. (in Chinese)
16. Syakur M., Khotimah B., Rochman E.M.S., Satoto B.D. Integration k-means clustering method and elbow method for identification of the best customer profile cluster // IOP Conference Series: Materials Science and Engineering. 2018. V. 336. N 1. P. 012017. doi: 10.1088/1757-899X/336/1/012017
17. Thinsungnoen T., Kaoungku N., Durongdumronchai P., Kerdprasop K., Kerdprasop N. The clustering validity with silhouette and sum of squared errors // Proc. 3rd International Conference on Industrial Application Engineering (ICIAE 2015). 2015. P. 44–51. doi: 10.12792/iciae2015.012
18. Menardi G. Density-based Silhouette diagnostics for clustering methods // Statistics and Computing. 2011. V. 21. N 3. P. 295–308. doi: 10.1007/s11222-010-9169-0

13. Nithya P., Sumathi P. Novel pre-processing technique for web log mining by removing global noise and web robots. *Proc. of the National Conference on Computing and Communication Systems (NCCCS 2012)*, 2012, pp. 41–45. doi: 10.1109/NCCCS.2012.6412976
14. Kanungo T., Mount D.M., Netanyahu N.S., Piatko C.D., Silverman R., Wu A.Y. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002, vol. 24, no. 7, pp. 881–892. doi: 10.1109/TPAMI.2002.1017616
15. Yang S.-L., Li Y.-S., Hu X.-X., Pan R.-Y. Optimization study on k value of k-means algorithm. *Xitong Gongcheng Lilun yu Shijian/ System Engineering Theory and Practice*, 2006, vol. 26, no. 2, pp. 97–101. (in Chinese)
16. Syakur M., Khotimah B., Rochman E.M.S., Satoto B.D. Integration k-means clustering method and elbow method for identification of the best customer profile cluster. *IOP Conference Series: Materials Science and Engineering*, 2018, vol. 336, no. 1, pp. 012017. doi: 10.1088/1757-899X/336/1/012017
17. Thinsungnoen T., Kaoungku N., Durongdumronchai P., Kerdprasop K., Kerdprasop N. The clustering validity with silhouette and sum of squared errors. *Proc. 3rd International Conference on Industrial Application Engineering (ICIAE 2015)*, 2015, pp. 44–51. doi: 10.12792/iciae2015.012
18. Menardi G. Density-based Silhouette diagnostics for clustering methods. *Statistics and Computing*, 2011, vol. 21, no. 3, pp. 295–308. doi: 10.1007/s11222-010-9169-0

Авторы

Евстратов Виктор Владимирович — аспирант, Санкт-Петербургский государственный университет, Санкт-Петербург, 199034, Российская Федерация, ORCID ID: 0000-0002-0796-6559, viktor.evst@gmail.com

Ананьевский Михаил Сергеевич — кандидат физико-математических наук, доцент, Санкт-Петербургский государственный университет, Санкт-Петербург, 199034, Российская Федерация, Scopus ID: 14055405100, ORCID ID: 0000-0003-2355-9184, msaipme@yandex.ru

Authors

Victor V. Evstratov — Postgraduate, Saint Petersburg State University, Saint Petersburg, 199034, Russian Federation, ORCID ID: 0000-0002-0796-6559, viktor.evst@gmail.com

Mikhail S. Ananyevskiy — PhD, Associate Professor, Saint Petersburg State University, Saint Petersburg, 199034, Russian Federation, Scopus ID: 14055405100, ORCID ID: 0000-0003-2355-9184, msaipme@yandex.ru