

УДК 303.717 + 004.93

СРАВНЕНИЕ АЛГОРИТМОВ КЛАСТЕРИЗАЦИИ K-MODES И CLOPE (НА ПРИМЕРЕ ДАННЫХ СОЦИОЛОГИЧЕСКОГО ОПРОСА)

П. И. Червоткина

Статья представлена научным руководителем – д. соц. н., доц. Малининой Т. Б.

В статье проводится сравнительный анализ двух алгоритмов кластеризации категориальных данных – K-modes и CLOPE – на примере данных социологического опроса ВЦИОМ по теме кибербезопасности (выборка 1561 респондент после предварительной обработки). Рассматриваются особенности применения каждого метода к бинарным переменным, отражающим практики обеспечения безопасности в интернете, управление паролями и осведомленность о государственных механизмах защиты персональных данных. Показано, что K-modes демонстрирует более высокую интерпретируемость результатов и значение силуэтной метрики (0,35 против 0,2895 у CLOPE), что делает его предпочтительным для выявления основных тенденций. CLOPE, в свою очередь, позволяет формировать более компактные и внутренне однородные кластеры, выявляя редкие и специфические группы респондентов.

Ключевые слова: кластеризация категориальных данных, алгоритмы K-modes и CLOPE, машинное обучение, метрика силуэта.

Введение.

Кластеризация как метод анализа особенно востребована в социологии, поскольку не требует разметки данных и способна работать с категориальными переменными, зачастую преобладающими в опросах. Алгоритмы кластеризации можно разделить на основании основных идей определения близости точек данных:

1. На основе центроида
2. На основе плотности
3. На основе распределения или модели
4. На основе иерархии
5. На основе сетки [5, с. 71]

К известным алгоритмам кластеризации относятся: K-modes и CLOPE (на основе центроида), ROCK и метод Уорда (иерархические).

Алгоритм K-modes – это обобщение K-means для категориальных данных. Он использует метрику Хэмминга для определения расстояния до центров кластеров, которые представляют собой наиболее часто встречающиеся значения характеристик объектов в кластере.

Алгоритм CLOPE (Clustering with sLOPE), предложенный в 2002 Y. Yang,

X. Guan, и J. You считается эффективным для больших наборов категориальных данных [6]. Коэффициент r (отталкивания) позволяет определить итоговое число кластеров. Чем больше r , тем ниже уровень сходства и тем больше кластеров будет создано.

Основная часть.

Предварительная обработка данных
Для сравнения алгоритмов используется массив данных от ВЦИОМ по теме кибербезопасности [1]. Выбор данных обоснован актуальностью темы цифровизации и рисков, связанных вместе с ней.

Наблюдения, где респонденты затруднились ответить хотя бы в одном из вопросов, были исключены в переменных EDU, S2, DONOD, FO, TIP, tq1, tq2_1-16, tq4_1-10, tq5. В переменных tq6, tq7_1-3, tq8_1-6, tq9 значения «затрудняюсь ответить» были оставлены, так как доля таких ответов составила чуть больше 8% для переменных tq8_1-6, около 6% для остальных. Предполагается, что такие ответы могут отражать недостаточную осведомленность респондентов в вопросах кибербезопасности и представляют ценность для анализа.

Так как переменные tq2_1-16, tq4_1-10, tq8_1-6 представляют собой множественные ответы, то каждое значение было перекодировано в бинарные переменные.

Объем выборки составил 1561.

Наиболее удобными для кластеризации оказались бинарные переменные. Для оценки связи бинарных признаков использовался критерий хи-квадрат с коэффициентом Крамера. Из анализа исключены переменные с сильной взаимосвязью (коэффициент Крамера > 0.2) и низкой долей положительных ответов ($< 14\%$), что свидетельствует об их недостаточной вариативности для кластеризации. Таким образом, оставлены переменные с ответами о практиках обеспечения безопасности в интернете, управления паролями: «Не пересылаю пароли по электронной почте и в личных сообщениях в социальных сетях» (ans_11), «Использую программы-антивирусы» (ans_1), «Не храню информацию о банковской карте – CVC и ПИН код карты в компьютере» (ans_10) «Иногда повторяю пароли, но для важных сервисов стараюсь использовать разные» (par_4), «Никогда не захожу в свои аккаунты через

чужие устройства» (par_7), «Использую двухфакторную аутентификацию (не только пароль, но и второй код, например, из СМС)» (par_2). Также включена переменная об осведомленности о способах государственной защиты персональных данных при финансовых операциях: «Персональная электронная подпись» (gos_2). Доля распределений положительных ответов идет от 14% до 44%.

Сравнение алгоритмов CLOPE, K-modes.

Инициализация K-modes шла с фиксированными значениями и инициализацией центроидов Huang, n_init=10, random_state = 42. Для сравнения алгоритмов использовался такой набор признаков – ans_11, ans_1, ans_10, par_4, par_7, par_2, gos_2, SEX. Оптимальными были выбраны 5 кластеров с самым высоким значением метрики силуэта (0.35). Метрика рассчитывалась на основе метрики Хэмминга и показывала хорошую разделяемость кластеров.

Кластеризация CLOPE была оценена по различным значениям r (отталкивания), наилучший результат был достигнут при значении $r = 4.5$. Подробное сравнение алгоритмов указано в табл. 1.

Таблица 1.
Метрики качества кластеризации k-modes и CLOPE

	k-modes	CLOPE			
	5 кластеров	R = 3 (10 кластеров)	R = 3.5 (9 кластеров)	R = 4 (9 кластеров)	R = 4.5 (10 кластеров)
Метрика силуэта	0.35	0.2411	0.2706	0.2832	0.2895
Cost	1581.0	341.96	360.69	335.76	322.43

Можно увидеть, что K-modes демонстрирует более высокое качество кластеризации по силуэтной метрике (0.35 против максимальных 0.2895 у CLOPE), что свидетельствует о лучшей разделимости кластеров и их содержательной интерпретируемости для социологических данных. Однако CLOPE при значении $r = 4.5$ показывает наименьшую сумму расстояний точек до центров

их кластеров (322.43), формируя более компактные и внутренне однородные группы респондентов. Важно отметить, что полученные значения силуэта можно считать хорошими для данных в социальных науках. Согласно некоторым работам, пороговое значение для того, чтобы считать, что данные имеют кластерную структуру считается 0.25 [3]. Однако это значение предлагается для

кластеризации метрических данных. Обзор исследований по кластеризации порядковых или номинальных данных в социальных науках показывает, что значения силуэтного коэффициента редко превышают 0.2-0.3 [4].

Почти все респонденты во всех кластерах не совершают никаких действий, чтобы обезопасить себя в сети. Согласно алгоритму K-modes, самые молодые респонденты наиболее высоко владеют цифровыми навыками: большинство из кластера 1 (297 респондентов, средний возраст 43, мужчины 76.1%) используют двухфакторную аутентификацию, а также знают о способе государственной защиты персональных данных – персональной электронной подписи. Кластеры 4 (244 респондентов, средний возраст 47.4, женщины 100%) и кластер 5 (262 респондентов, средний возраст 49.7, мужчины 100%) также знают о персональной электронной подписи. Для этих кластеров характерно проживание в городах миллиониках. Характеристики выявленных кластеров согласуются с исследованием Фазульяновой С. Н., Логиновой И. А., Яковлева А. В. [2], где продвинутый уровень осведомленности демонстрируют в основном мужчины лишь в отдельных практиках.

Алгоритм CLOPE указал мелкие кластеры с высокой однородностью, что позволяет выделить уникальные немногочисленные группы. Так кластер 8 (81 респондент, средний возраст 43.2, мужчины 54.3%), обладающий самым высо-

ким уровнем компетентности в сфере кибербезопасности, в большинстве использует программы-антивирусы, двухфакторную аутентификацию и знают о персональной электронной подписи. Также этот кластер и кластеры из K-modes с высоким уровнем компетентности знают о том, как частные лица (например, блогеры) используют собираемые данные в интернете (ответ «определенно да»), в отличие от менее осведомленных респондентов.

Заключение.

Результаты исследования показали преимущество алгоритма K-modes перед CLOPE для анализа социологических данных. Для социологического исследования интерпретируемость результатов зачастую важнее формальных метрик качества. За счет указания количества кластеров социолог может найти оптимальный вариант кластеризации опираясь не только на метрики, но и на теоретическую модель и существующие научные подходы. При этом алгоритм CLOPE позволяет выявить более тонкие особенности и уловить редкие группы респондентов со специфическими характеристиками, что может быть ценно для детального анализа отдельных сегментов. Однако для общего описания социальной реальности и выявления основных тенденций такое дробное разделение может оказаться излишним, затрудняя формирование целостной картины и интерпретацию результатов.

Список источников и литературы

1. Всероссийский центр изучения общественного мнения (ВЦИОМ). Цифровая среда без риска: возможно ли это? [Электронный ресурс] // Аналитический обзор. 2024. URL: <https://wciom.ru/analytical-reviews/analiticheskii-obzor/cifrovaja-sreda-bez-riska-vozmozno-li-eto> (дата обращения: 20.09.2025).
2. Фазульянова С. Н. Компетентность жителей Самарской области в вопросах кибербезопасности: опыт социологического измерения / Фазульянова С. Н., Логинова И. А., Яковлев А. В. [Электронный ресурс] // Социология. 2023. № 2. URL: <https://cyberleninka.ru/article/n/kompetentnost-zhiteley-samarskoy-oblasti-v-voprosah-kiberbezopasnosti-opyt-sotsiologicheskogo-izmereniya> (дата обращения: 20.09.2025).
3. Benaya R. Clustering Content Types and User Roles Based on Tweet Text Using K-Medoids Partitioning Based / Benaya R., Sibaroni Y., Ihsan A. F. // Journal of Informatics and Computer Science (JOSYC). 2023. Vol. 4, No. 4. P. 749-756. DOI: 10.47065/josyc.v4i4.3751.

4. *Dinh D. T.* Estimating the optimal number of clusters in categorical data clustering by silhouette coefficient / Dinh D. T., Fujinami T., Huynh V. N. // International Symposium on Knowledge and Systems Sciences. – Singapore : Springer Singapore, 2019. – P. 1-17.

5. *Ghosal A.* Short Review on Different Clustering Techniques and Their Applications / Ghosal A., Nandy A., Das A. K., Goswami S., Panday M. // Emerging Technology in Modelling and Graphics / ed. Mandal J., Bhattacharya D. – Singapore : Springer, 2020. – P. 69-81. – (Advances in Intelligent Systems and Computing ; vol. 937). – URL: <https://www.springerprofessional.de/a-short-review-on-different-clustering-techniques-and-their-appl/16938704> (last request: 20.09.2025).

6. *Yang Y.* CLOPE: A Fast and Effective Clustering Algorithm for Transactional Data / Yang Y., Guan X., You J. [electronic resource] // Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2002. P. 682-687. URL: <https://dl.acm.org/doi/10.1145/775047.775149> (last request: 20.09.2025).

Червоткина Полина Игоревна – бакалавр, кафедра социального анализа и математических методов в социологии, Факультет социологии, Санкт-Петербургский государственный университет, (г. Санкт-Петербург, Россия), bcu1119@gmail.com

Малинина Татьяна Борисовна – научный руководитель – д. соц. н., к. физ.-мат. н., доцент, профессор кафедры социального анализа и математических методов в социологии, Санкт-Петербургский государственный университет, (г. Санкт-Петербург, Россия), t.malinina@spbu.ru

COMPARISON OF K-MODES AND CLOPE CLUSTERING ALGORITHMS (USING THE DATA FROM A SOCIOLOGICAL SURVEY AS AN EXAMPLE)

P. I. Chervotkina

*The article is presented by the scientific supervisor, Doctor of Social Sciences,
Associate Professor Malinina T. B.*

This article provides a comparative analysis of two categorical data clustering algorithms – K-modes and CLOPE – using data from a VTsIOM sociological survey on cybersecurity (a sample of 1,561 respondents after preprocessing). The application of each method to binary variables reflecting internet security practices, password management, and awareness of government personal data protection mechanisms is discussed. It is shown that K-modes demonstrates higher interpretability of results and a higher silhouette metric value (0.35 versus 0.2895 for CLOPE), making it preferable for identifying key trends. CLOPE, in turn, allows for the formation of more compact and internally homogeneous clusters, identifying rare and specific groups of respondents.

Keywords: categorical data clustering, K-modes and CLOPE algorithms, machine learning, silhouette metric.

References

1. Vserossijskij centr izucheniya obshchestvennogo mneniya (VCIOM). Cifrovaya sreda bez riska: vozmozhno li ehto? [Russian Public Opinion Research Center (VCIOM). Digital Environment Without Risk: Is It Possible?] [Electronic resource] // Analiticheskij obzor [Analytical Review]. 2024. – URL: <https://wciom.ru/analytical-reviews/analiticheskii-obzor/cifrovaja-sreda-bez-riska-vozmozhno-li-ehto> (last request: 20.09.2025). (In Russ.)

2. *Fazulyanova S. N.* Kompetentnost' zhitelej Samarskoj oblasti v voprosakh kiberbezopasnosti: opyt sociologicheskogo izmereniya [Competence of Residents of the Samara Region in Cybersecurity Issues: The Experience of Sociological Measurement] / Fazulyanova S. N., Loginova I. A., Yakovlev A. V. [Electronic resource] // Sociologiya [Sociology]. 2023. No. 2. – URL: <https://cyberleninka.ru/article/n/kompetentnost-zhiteley-samarskoy-oblasti-v-voprosah-kiberbezopasnosti-opyt-sotsiologicheskogo-izmereniya> (last request: 20.09.2025). (In Russ.)

3. *Benaya R.* Clustering Content Types and User Roles Based on Tweet Text Using K-Medoids Partitioning Based / Benaya R., Sibaroni Y., Ihsan A. F. // *Journal of Informatics and Computer Science (JOSYC)*. 2023. Vol. 4, No. 4. P. 749-756. DOI: 10.47065/josyc.v4i4.3751.

4. *Dinh D. T.* Estimating the optimal number of clusters in categorical data clustering by silhouette coefficient / Dinh D. T., Fujinami T., Huynh V. N. // *International Symposium on Knowledge and Systems Sciences*. – Singapore : Springer Singapore, 2019. – P. 1-17.

5. *Ghosal A.* Short Review on Different Clustering Techniques and Their Applications / Ghosal A., Nandy A., Das A. K., Goswami S., Panday M. // *Emerging Technology in Modelling and Graphics* / ed. Mandal J., Bhattacharya D. – Singapore : Springer, 2020. – P. 69-81. – (Advances in Intelligent Systems and Computing ; vol. 937). – URL: <https://www.springerprofessional.de/a-short-review-on-different-clustering-techniques-and-their-appl/16938704> (last request: 20.09.2025).

6. *Yang Y.* CLOPE: A Fast and Effective Clustering Algorithm for Transactional Data / Yang Y., Guan X., You J. [electronic resource] // *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2002. P. 682-687. URL: <https://dl.acm.org/doi/10.1145/775047.775149> (last request: 20.09.2025).

Chervotkina Polina Igorevna – Bachelor's degree, Department of Social Analysis and Mathematical Methods in Sociology, Faculty of Sociology, St. Petersburg State University, (St. Petersburg, Russia), bcur1119@gmail.com

Malinina Tatiana Borisovna – scientific supervisor – Doctor of Social Sciences, Candidate of Physical and Mathematical Sciences, Associate Professor, Professor of the Department of Social Analysis and Mathematical Methods in Sociology, St. Petersburg State University, (St. Petersburg, Russia), t.malinina@spbu.ru

Статья поступила в редакцию: 21.04.2026; принята к публикации: 05.05.2026

ДЛЯ ЦИТИРОВАНИЯ:

Червоткина П. И. Сравнение алгоритмов кластеризации K-modes и CLOPE (на примере данных социологического опроса) // *Социогуманитарные коммуникации*. – 2026. – № 2(16). – С. 160-164.

FOR CITATION:

Chervotkina P. I. Sravnenie algoritmov klasterizacii K-modes i CLOPE (na primere dannykh sociologicheskogo oprosa) [A Comparison of K-Modes and CLOPE Clustering Algorithms (Using Sociological Survey Data)] // *Sociogumanitarnye kommunikacii* [Social and humanitarian communications]. 2026. № 2(16). P. 160-164.