
**АВТОМАТИЗАЦИЯ И УПРАВЛЕНИЕ
В МАШИНОСТРОЕНИИ**

УДК 519.6

**СОКРАЩЕНИЕ ОЦЕНКИ УСАДКИ ДЖЕЙМСА–ШТЕЙНА ДЛЯ
РЕШЕНИЯ ЗАДАЧИ АВТОМАТИЧЕСКОЙ ГРУППИРОВКИ ОДНОРОДНЫХ
ПРОИЗВОДСТВЕННЫХ ПАРТИЙ**© 2024 г. **Ф. Г. Ахматшин¹, И. А. Петрова¹, Л. А. Казаковцев², И. Н. Кравченко^{3, *}***¹Сибирский государственный университет науки и технологий им. академика
М. Ф. Решетнева, Красноярск, Россия**²Сибирский федеральный университет, Красноярск, Россия**³Институт машиноведения им. А. А. Благонравова РАН, Москва, Россия***e-mail: kravchenko-in71@yandex.ru*

Поступила в редакцию 01.12.2023 г.

После доработки 12.02.2024 г.

Принята к публикации 19.02.2024 г.

Сокращение оценки усадки Джеймса–Штейна может значительно повысить точность кластерного анализа k -средних для относительно широкого диапазона данных. В статье исследована эффективность использования оценки усадки Джеймса–Штейна при решении задачи автоматической группировки промышленной продукции в однородные производственные партии. Проведены испытания для партий интегральных схем путем сравнения полученных результатов усадки с традиционным алгоритмом k -средних. Набор данных нормализуется в соответствии со значениями допустимого дрейфа, приемлемого параметра и стандартного отклонения. С помощью индекса Rand установлено, что точность кластеризации существенно возрастает в задаче автоматической группировки промышленной продукции в однородные производственные партии, когда средние значения неинформативных параметров уменьшаются до нуля. Установлено, что использование сокращения оценки усадки Джеймса–Штейна позволяет снизить влияние неинформативных параметров нормализованных данных до приемлемых значений.

Ключевые слова: алгоритм k -средних, интегральные схемы, кластеризация, оценка усадки Джеймса–Штейна, эвристика, электронные радиоизделия, электротермическая подготовка

DOI: 10.31857/S0235711924030093, **EDN:** PGRPBV

Контроль качества промышленной продукции требует получения наиболее точных и стабильных результатов в отрасли с повышенными требованиями к качеству продукции. При этом точность относится к снижению доли ошибок автоматической группировки, а стабильность — к повторяемости результатов при многократном запуске алгоритма.

В специализированных испытательных центрах проводятся сотни тестов для комплектации бортового оборудования космических систем высоконадежной электронной компонентной базой и анализа каждого полупроводникового устройства. Одним из требований является, что отгружаемая партия продукции должна быть изготовлена из одной партии сырья (макросхем) и это гарантируется, если устройства

производятся только для использования в космической промышленности. На основе данных многомерных результатов тестирования реализованы различные алгоритмы кластеризации для решения задачи обнаружения однородных партий продукции [1–3]. Такие многомерные данные имеют несколько параметров и различаются единицами измерения [4]. Поэтому при проведении исследования набор данных сгруппируется по значениям допустимого дрейфа, допустимым значениям параметра и с использованием нормализации по стандартному отклонению [5].

Использование многомерных данных в модели автоматической группировки влияет на точность решения задачи. Между некоторыми характеристиками часто существуют явные корреляции. Однако эта задача эффективно решается с помощью алгоритма k -средних со специальными мерами расстояния, в которых учитываются эти зависимости.

В исходных многомерных данных обнаружены выбросы, которые влияют на интерпретацию полученных результатов. Для снижения влияния выбросов применяется нормализация значений исходных данных [6].

Представленные методы нормирования оказывают большое влияние на решение задачи автоматической группировки продуктов по однородным производственным партиям. При этом каждый из продуктов имеет большое количество входных характеристик. Характеристики с таким диапазоном значений также оказывают значительное влияние на группировку продуктов.

Постановка задачи. Исследование алгоритмов в рассматриваемой области связано с выбором методов нормализации данных, метрики расстояния и инициализации центроида [7–10].

Пусть исходный набор данных $D = \{x_1, x_2, \dots, x_N\}$ состоит из N точек. Обозначим центроиды, полученные после применения автоматической группировки k -средних $c_1, c_2, \dots, c_j, \dots, c_K$. Цель состоит в том, чтобы сумма квадратов расстояний от известных точек до ближайших центроидов достигла своего минимума

$$\operatorname{argmin} F(c_1, c_2, \dots, c_j, \dots, c_K) = \sum_{i=1}^n \min_{j \in \{1, K\}} x_i - c_j^2, \quad (1)$$

где c_j – центроид кластера C_j .

Для минимизации суммы квадратов расстояний в предложенном алгоритме центроиды алгоритма k -средних итеративно назначаются и обновляются [11, 12]. При этом алгоритм k -средних применяется для решения задачи автоматической группировки по минимизации целевой функции, которая эквивалентна задаче нахождения квадрата расстояния от точек данных до ближайшего центроида [13–15]. Наилучшим показателем минимизации целевой функции является сумма расстояний от точек в кластере до ближайших центроидов. Для улучшения средних значений $\|x_i - c_j\|$ в кластере используем сокращение оценки усадки Джеймса–Штейна [16, 17] для x_i к среднему значению μ_j всего набора данных

$$x_{JS} = \left(1 - \frac{(p-2)k\sigma^2}{x_i - \bar{\mu}^2} \right)^+ (x_i - \bar{\mu}) + \bar{\mu}, \quad (2)$$

где $\mu = (\mu_1, \dots, \mu_p)$ – p -мерный средний параметр; $x = (x_1, \dots, x_p)$ – p -мерное наблюдение; $\bar{\mu} = \frac{1}{N} \sum_{i=1}^N x_i$ – среднее значение выборки; $(\)^+$ – оценка с положительной частью, равной нулю при отрицательных значениях; σ^2 – квадрат дисперсии

значения x ; k – коэффициент сокращения оценки усадки к некоторому значимому ненулевому значению.

Улучшенные точки x_{js} используются в предложенном алгоритме. В статье исследуем эффективность использования сокращения оценки усадки после нормализации данных перед выполнением базового алгоритма k -средних. В отличие от результатов исследования [18], в котором выполняется сжатие центроидов относительно начала координат, рассматриваем процесс сжатия каждой точки относительно центра всего набора данных.

Методы исследования. Исходные данные представляют собой набор тестовых испытаний смешанной партии двух интегральных схем (ИС) 140UD25A и 140UD26A до и после электротермической подготовки (ЭТТ). Общее количество микросхем равно 807 и 532 для первой и второй интегральных схем. Затем дополнительно отбираем каждые четвертые тестовые данные и формируем дополнительный набор данных в количестве 201 и 132 для первой и второй микросхем соответственно. Интегральные схемы в каждой партии описываются 18 входными измеряемыми параметрами.

Набор данных сгруппирован с использованием нормализации по стандартному отклонению (A), для которого использовались данные от 3 до 18 параметров, причем параметры 1 и 2 являлись неинформативными и в дальнейших расчетах не учитывались.

Вектор характерных данных с использованием нормализации по стандартному отклонению

$$x_{std} = (0, 0, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{16}, x_{17}). \quad (3)$$

Набор данных с использованием нормализации по значениям допустимого дрейфа (B) рассчитан только для 3–6 параметров, для которых установлены соответствующие нормы (остальные данные были обнулены). Разница в изменениях 3–6 параметров до и после ЭТТ учтена путем добавления 19–22 параметров. Вектор характерных данных с использованием нормализации по значениям допустимого дрейфа

$$x_{vpd} = (0, 0, x_2, x_3, x_4, x_5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, x_{18}, x_{19}, x_{20}, x_{21}). \quad (4)$$

Набор данных с использованием нормализации по допустимым значениям параметров (C) рассчитан для 3–16 параметров, для которых были установлены соответствующие нормы, остальные данные были обнулены. Разница в изменении 3–6 параметров до и после ЭТТ учтена путем добавления 19–22 параметров. Вектор характерных данных с использованием нормализации по приемлемым значениям параметров

$$x_{apv} = (0, 0, x_2, x_3, x_4, x_5, x_6, \dots, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, 0, 0, x_{18}, x_{19}, x_{20}, x_{21}). \quad (5)$$

Таким образом, сформировано 24 набора данных, для каждого из которых проведено по 30 экспериментов с использованием алгоритма кластеризации k -средних с евклидовыми мерами расстояния. При этом исследовали влияние коэффициента $k = (0, \dots, 100)$ сокращения оценки усадки Джеймса–Штейна к некоторому значимому ненулевому значению.

Результаты вычислительных экспериментов. Результаты кластеризации приведены в табл. 1–4, где показаны максимум (Max), минимум (Min), средние значения (Mean), медиана (Median) и стандартное отклонение (Std) для индекса Rand (RI) и целевой функции (F). При этом для целевой функции F вычисляются значения коэффициентов вариации (Var) и охвата (Spn). Средние результаты кластеризации

Таблица 1. Результаты кластеризации для ИС 140UD25A с использованием нормализации по стандартному отклонению с коэффициентом k сокращения оценки усадки Джеймса–Штейна

JS	Значение целевой функции F							Индекс Ранда RI				
k	Max	Min	Mean	Median	StD	Var	Spn	Max	Min	Mean	Median	StD
0	77.428	66.811	68.663	66.813	2.746	0.040	10.617	0.605	0.559	0.592	0.564	0.011
1	24.803	20.604	21.491	20.604	0.829	0.039	4.198	0.573	0.417	0.468	0.419	0.055
2	12.597	10.602	11.005	10.606	0.376	0.034	1.995	0.562	0.337	0.456	0.339	0.059
3	6.856	6.130	6.392	6.167	0.185	0.029	0.726	0.622	0.368	0.467	0.381	0.057
4	5.324	4.855	5.030	4.862	0.110	0.022	0.470	0.622	0.396	0.496	0.401	0.086
5	4.635	3.960	4.205	4.007	0.148	0.035	0.675	0.645	0.386	0.553	0.393	0.093
6	3.858	3.303	3.491	3.310	0.146	0.042	0.555	0.637	0.372	0.547	0.374	0.097
7	3.626	2.842	3.033	2.854	0.192	0.063	0.784	0.669	0.351	0.594	0.358	0.067
8	3.360	2.527	2.682	2.530	0.177	0.066	0.833	0.710	0.348	0.602	0.461	0.057
9	2.763	2.218	2.396	2.236	0.134	0.056	0.546	0.720	0.547	0.633	0.563	0.038
10	2.339	1.920	2.070	1.921	0.128	0.062	0.419	0.701	0.378	0.620	0.381	0.073
20	0.879	0.579	0.651	0.584	0.074	0.113	0.300	0.690	0.569	0.654	0.587	0.027
30	0.901	0.448	0.596	0.449	0.141	0.237	0.452	0.671	0.511	0.618	0.547	0.032
40	1.163	0.380	0.538	0.388	0.172	0.320	0.783	0.647	0.299	0.582	0.383	0.065
50	0.805	0.343	0.451	0.346	0.087	0.193	0.462	0.636	0.466	0.571	0.468	0.041
60	0.694	0.290	0.438	0.296	0.128	0.291	0.404	0.605	0.405	0.518	0.405	0.058
70	0.738	0.248	0.380	0.256	0.124	0.326	0.490	0.573	0.299	0.489	0.345	0.056
80	0.662	0.236	0.399	0.238	0.125	0.313	0.426	0.547	0.299	0.444	0.299	0.063
90	0.493	0.196	0.311	0.198	0.089	0.288	0.296	0.573	0.375	0.457	0.375	0.058
100	2.339	1.920	2.070	1.921	0.128	0.062	0.419	0.701	0.378	0.620	0.381	0.073

k -средних с использованием нормализации по стандартному отклонению (А), значениям допустимого дрейфа (В), а также по допустимым значениям параметров (С) с различным коэффициентом k сокращения оценки усадки Джеймса–Штейна для ИС 140UD25A (размер набора данных $n = 201$) представлены в табл. 1, 2 и на рис. 1–4.

Первый набор данных для ИС 140UD25A при $k = 20$ показывает наилучшее значение точности кластеризации по индексу Rand при минимальном значении стандартного отклонения целевой функции. Вектор характеристик улучшенной оценки усадки по Джеймсу–Штейна

$$x_{std} = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, x_{16}, 0), \quad (6)$$

показывает наибольшее влияние параметра 16.

Для второго набора данных для ИС 140UD25A при $k = (4, \dots, 100)$ использование коэффициента сокращения оценки усадки Джеймса–Штейна неэффективно. Характеристический вектор сокращения оценки усадки Джеймса–Штейна

$$x_{std} = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0), \quad (7)$$

показывает уменьшение параметров для всего набора данных. Эффективное уменьшение второго набора данных находится в диапазоне $k = (0, \dots, 1)$.

Таблица 2. Результаты кластеризации для ИС 140UD25A с использованием нормализации по допустимому значению параметра с коэффициентом k сокращения оценки усадки Джеймса–Штейна

JS	Значение целевой функции F							Индекс Ранда RI				
k	Max	Min	Mean	Median	StD	Var	Spn	Max	Min	Mean	Median	StD
0	46.240	40.127	41.271	40.130	1.696	0.041	6.113	0.603	0.573	0.593	0.575	0.006
1	12.072	10.582	11.037	10.628	0.304	0.028	1.490	0.554	0.409	0.452	0.411	0.034
2	5.577	4.807	5.012	4.807	0.234	0.047	0.771	0.487	0.359	0.429	0.360	0.033
3	2.655	2.295	2.360	2.296	0.064	0.027	0.360	0.581	0.397	0.478	0.400	0.044
4	1.951	1.674	1.739	1.682	0.054	0.031	0.277	0.586	0.387	0.467	0.393	0.057
5	1.674	1.387	1.487	1.390	0.069	0.046	0.287	0.616	0.386	0.484	0.388	0.064
6	1.790	1.205	1.299	1.216	0.104	0.080	0.585	0.625	0.383	0.471	0.385	0.082
7	1.498	1.061	1.149	1.067	0.086	0.075	0.437	0.614	0.380	0.480	0.392	0.081
8	1.288	0.965	1.026	0.965	0.078	0.076	0.322	0.598	0.406	0.462	0.406	0.069
9	1.137	0.864	0.955	0.870	0.064	0.067	0.273	0.596	0.370	0.463	0.371	0.080
10	1.074	0.754	0.864	0.762	0.073	0.085	0.320	0.637	0.399	0.505	0.399	0.088
20	0.577	0.374	0.441	0.377	0.056	0.127	0.202	0.770	0.541	0.687	0.544	0.062
30	0.518	0.360	0.406	0.361	0.045	0.110	0.158	0.764	0.530	0.677	0.539	0.056
40	0.477	0.344	0.395	0.344	0.041	0.103	0.133	0.760	0.556	0.679	0.559	0.044
50	0.513	0.341	0.392	0.343	0.047	0.120	0.172	0.753	0.467	0.659	0.497	0.066
60	0.494	0.324	0.375	0.324	0.044	0.117	0.170	0.747	0.467	0.666	0.507	0.052
70	0.471	0.318	0.367	0.318	0.036	0.097	0.153	0.721	0.536	0.656	0.545	0.046
80	0.422	0.313	0.355	0.313	0.028	0.079	0.108	0.744	0.521	0.655	0.548	0.039
90	0.523	0.306	0.356	0.306	0.051	0.143	0.217	0.706	0.511	0.641	0.518	0.044
100	1.074	0.754	0.864	0.762	0.073	0.085	0.320	0.637	0.399	0.505	0.399	0.088

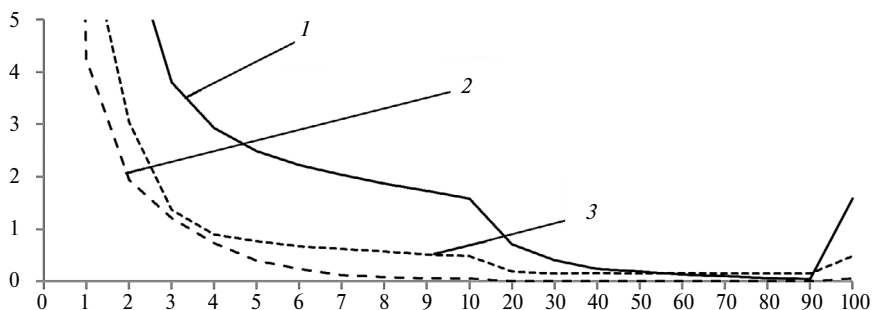


Рис. 1. Сравнительные результаты для ИС 140UD25A с другим коэффициентом k сокращения оценки усадки Джеймса–Штейна по значению целевой функции: 1 — нормализация по стандартному отклонению; 2 — нормализация по значениям допустимого дрейфа; 3 — по допустимым значениям параметров.

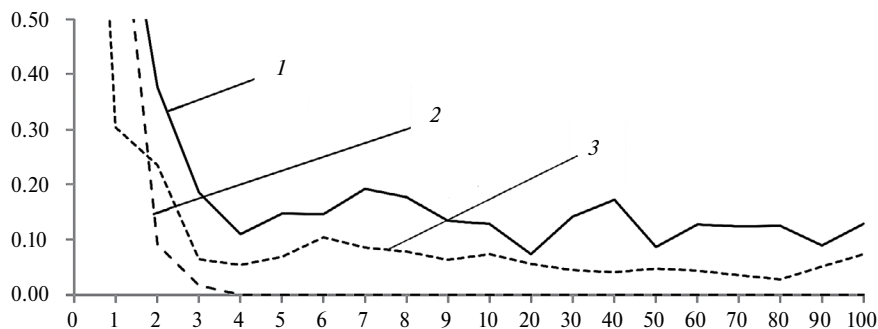


Рис. 2. Сравнительные результаты для ИС 140UD25A с другим коэффициентом k сокращения оценки усадки Джеймса–Штейна на стандартное отклонение значения целевой функции: 1 – нормализация по стандартному отклонению; 2 – нормализация по значениям допустимого дрейфа; 3 – по допустимым значениям параметров.

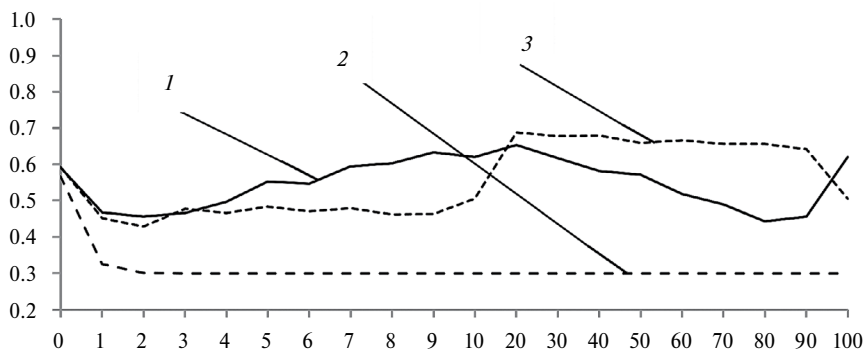


Рис. 3. Сравнительные результаты для ИС 140UD25A с другим коэффициентом k сокращения оценки усадки Джеймса–Штейна по индексу Rand: 1 – нормализация по стандартному отклонению; 2 – нормализация по значениям допустимого дрейфа; 3 – по допустимым значениям параметров.

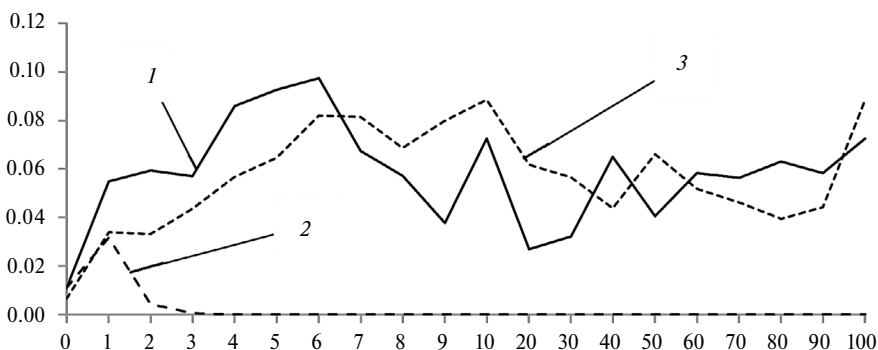


Рис. 4. Сравнительные результаты для ИС 140UD25A с другим коэффициентом k сокращения оценки усадки Джеймса–Штейна на стандартное отклонение индекса Рэнда: 1 – нормализация по стандартному отклонению; 2 – нормализация по значениям допустимого дрейфа; 3 – по допустимым значениям параметров.

Таблица 3. Сравнительные результаты для ИС 140UD26A с использованием нормализации по стандартному отклонению с другим коэффициентом k сокращения оценки усадки Джеймса–Штейна

JS	Значение целевой функции F							Индекс Ранда RI				
k	Max	Min	Mean	Median	StD	Var	Spn	Max	Min	Mean	Median	StD
0	63.737	57.724	59.685	57.725	1.642	0.028	6.013	0.662	0.580	0.615	0.581	0.019
1	19.854	12.967	13.832	12.967	1.375	0.099	6.888	0.576	0.376	0.483	0.398	0.053
2	7.244	6.018	6.461	6.029	0.299	0.046	1.227	0.596	0.332	0.486	0.349	0.057
3	3.966	3.655	3.796	3.659	0.085	0.022	0.311	0.626	0.482	0.562	0.490	0.032
4	3.452	2.794	2.927	2.798	0.164	0.056	0.657	0.657	0.399	0.568	0.432	0.053
5	2.910	2.406	2.482	2.417	0.092	0.037	0.504	0.703	0.490	0.573	0.492	0.061
6	2.363	2.122	2.215	2.128	0.059	0.027	0.241	0.677	0.497	0.610	0.500	0.058
7	2.145	1.941	2.030	1.947	0.049	0.024	0.204	0.684	0.479	0.603	0.482	0.064
8	2.043	1.770	1.867	1.784	0.072	0.038	0.273	0.687	0.496	0.632	0.502	0.050
9	1.956	1.609	1.730	1.616	0.088	0.051	0.348	0.687	0.497	0.626	0.500	0.050
10	1.798	1.464	1.582	1.487	0.072	0.045	0.334	0.688	0.511	0.622	0.512	0.057
20	0.939	0.607	0.704	0.611	0.076	0.108	0.332	0.615	0.412	0.552	0.414	0.046
30	0.623	0.251	0.406	0.260	0.085	0.209	0.372	0.582	0.272	0.482	0.323	0.074
40	0.401	0.147	0.239	0.152	0.065	0.274	0.254	0.518	0.272	0.422	0.272	0.073
50	0.276	0.100	0.187	0.108	0.051	0.272	0.176	0.478	0.272	0.356	0.272	0.058
60	0.189	0.070	0.126	0.071	0.040	0.317	0.119	0.440	0.272	0.354	0.272	0.057
70	0.126	0.050	0.098	0.050	0.029	0.296	0.076	0.413	0.272	0.314	0.272	0.046
80	0.078	0.020	0.058	0.023	0.022	0.371	0.059	0.381	0.272	0.304	0.272	0.035
90	0.051	0.017	0.039	0.017	0.015	0.385	0.033	0.359	0.272	0.288	0.272	0.022
100	1.798	1.464	1.582	1.487	0.072	0.045	0.334	0.688	0.511	0.622	0.512	0.057

Третий набор данных для ИС 140UD25A при $k = 40$ показывает наилучшее значение точности кластеризации по индексу Rand и минимальное значение стандартного отклонения целевой функции.

Вектор характеристик сокращения оценки усадки Джеймса–Штейна

$$x_{std} = (0, 0, 0, 0, x_4, x_5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0), \quad (8)$$

показывает наибольшее влияние параметров 4–5.

Средние результаты кластеризации k -средних с использованием нормализации по стандартному отклонению (А), значениям допустимого дрейфа (В), а также по допустимым значениям параметров (С) с различным коэффициентом сокращения оценки усадки Джеймса–Штейна, оцененного для ИС 140UD26A (размер набора данных $n = 132$), представлены в табл. 3, 4 и на рис. 5–8.

Первый набор данных для ИС 140UD26A при $k = 8$ показывает наилучшее значение точности кластеризации по индексу Rand и минимальное значение стандартного отклонения целевой функции. Вектор характеристик сокращения оценки усадки Джеймса–Штейна

$$x_{std} = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, x_{16}, 0) \quad (9)$$

показывает наибольшее влияние параметра 16.

Таблица 4. Сравнительные результаты для ИС 140UD26А с использованием нормализации по приемлемому значению параметра с другим коэффициентом k сокращения оценки усадки Джеймса–Штейна

JS	Значение целевой функции F							Индекс Ранда RI				
k	Max	Min	Mean	Median	StD	Var	Spn	Max	Min	Mean	Median	StD
0	34.203	26.052	26.844	26.058	1.546	0.058	8.152	0.611	0.584	0.603	0.586	0.006
1	8.612	6.320	6.731	6.320	0.590	0.088	2.292	0.544	0.392	0.448	0.396	0.042
2	3.585	2.829	3.035	2.830	0.161	0.053	0.756	0.559	0.310	0.440	0.320	0.057
3	1.694	1.225	1.367	1.226	0.113	0.082	0.469	0.683	0.375	0.488	0.395	0.080
4	1.013	0.835	0.888	0.839	0.037	0.042	0.178	0.715	0.406	0.549	0.415	0.107
5	0.915	0.698	0.766	0.700	0.057	0.075	0.218	0.731	0.409	0.584	0.410	0.122
6	0.792	0.607	0.664	0.607	0.051	0.077	0.185	0.726	0.385	0.629	0.388	0.122
7	0.792	0.550	0.620	0.554	0.051	0.082	0.242	0.724	0.402	0.666	0.402	0.083
8	0.799	0.507	0.571	0.507	0.056	0.097	0.293	0.824	0.380	0.632	0.381	0.127
9	0.594	0.459	0.514	0.461	0.036	0.070	0.135	0.732	0.380	0.650	0.386	0.101
10	0.552	0.430	0.479	0.430	0.041	0.085	0.122	0.901	0.647	0.733	0.664	0.056
20	0.286	0.153	0.182	0.154	0.035	0.193	0.133	0.939	0.723	0.872	0.723	0.057
30	0.210	0.126	0.149	0.126	0.020	0.136	0.084	0.957	0.788	0.879	0.793	0.049
40	0.235	0.120	0.156	0.120	0.038	0.245	0.115	0.932	0.626	0.856	0.627	0.090
50	0.234	0.120	0.144	0.122	0.026	0.180	0.113	0.930	0.632	0.872	0.685	0.068
60	0.240	0.123	0.156	0.124	0.030	0.190	0.116	0.927	0.757	0.857	0.770	0.046
70	0.195	0.122	0.145	0.122	0.019	0.134	0.073	0.927	0.782	0.865	0.784	0.044
80	0.220	0.125	0.148	0.126	0.018	0.123	0.094	0.924	0.782	0.862	0.782	0.041
90	0.234	0.124	0.150	0.124	0.029	0.193	0.111	0.918	0.684	0.853	0.730	0.055
100	0.552	0.430	0.479	0.430	0.041	0.085	0.122	0.901	0.647	0.733	0.664	0.056

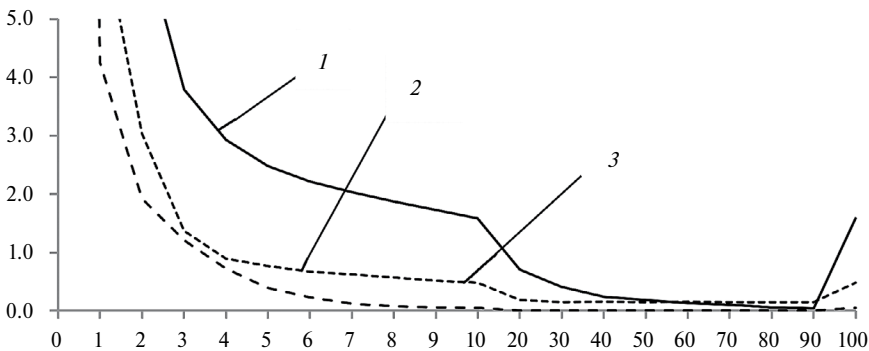


Рис. 5. Сравнительные результаты для ИС 140UD26А с другим коэффициентом k сокращения оценки усадки Джеймса–Штейна на значение целевой функции: 1 — нормализация по стандартному отклонению; 2 — нормализация по значениям допустимого дрейфа; 3 — по допустимым значениям параметров.

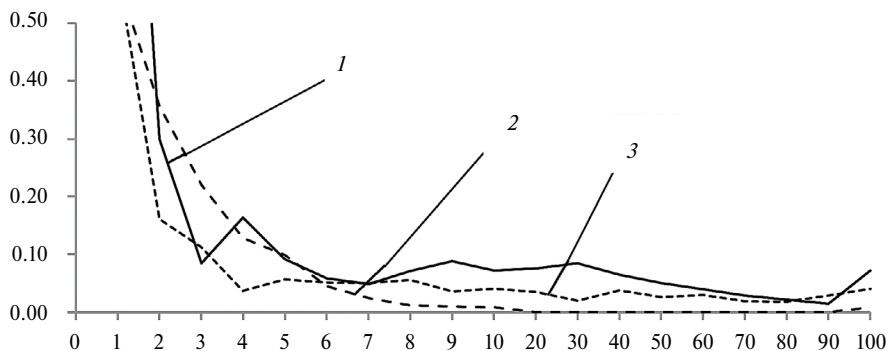


Рис. 6. Сравнительные результаты для ИС 140UD26A с другим коэффициентом k сокращения оценки усадки Джеймса–Штейна на стандартное отклонение значения целевой функции: 1 – нормализация по стандартному отклонению; 2 – нормализация по значениям допустимого дрейфа; 3 – по допустимым значениям параметров.

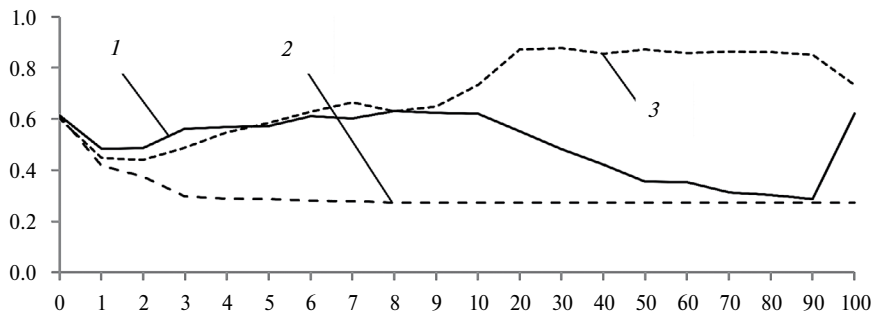


Рис. 7. Сравнительные результаты для ИС 140UD26A с другим коэффициентом k сокращения оценки усадки Джеймса–Штейна по индексу Rand: 1 – нормализация по стандартному отклонению; 2 – нормализация по значениям допустимого дрейфа; 3 – по допустимым значениям параметров.

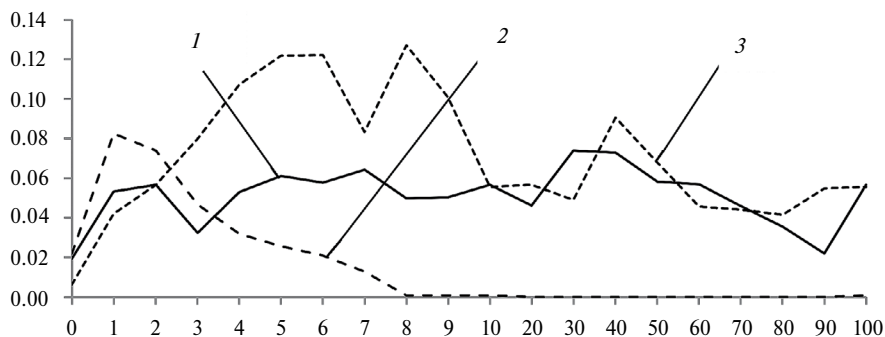


Рис. 8. Сравнительные результаты для ИС 140UD26A с другим коэффициентом k сокращения оценки усадки Джеймса–Штейна на стандартное отклонение индекса Рэнда: 1 – нормализация по стандартному отклонению; 2 – нормализация по значениям допустимого дрейфа; 3 – по допустимым значениям параметров.

Для второго набора данных для ИС 140UD26A при $k = (1, \dots, 100)$ использование сокращения оценки усадки Джеймса–Штейна неэффективно. Характерный вектор сокращения оценки усадки Джеймса–Штейна

$$x_{std} = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0), \quad (10)$$

показывает уменьшение параметров для всего набора данных. Эффективное уменьшение второго набора данных находится в диапазоне $k = (0, \dots, 1)$.

Для третьего набора данных для ИС 140UD26A при $k = 70$ показано наилучшее значение точности кластеризации по индексу Rand и минимальное значение стандартного отклонения целевой функции. При этом характерный вектор сокращения оценки усадки Джеймса–Штейна

$$x_{std} = (0, 0, 0, 0, x_4, x_5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0), \quad (11)$$

показывает наибольшее влияние параметров 4–5.

Выводы. 1. Использование сокращения оценки усадки Джеймса–Штейна уменьшает влияние неинформативных параметров нормализованных данных. При сокращении оценки усадки Джеймса–Штейна значение целевой функции уменьшается, а затем увеличивается. **2.** Наилучшее значение точности кластеризации по индексу Rand наблюдается при минимальном значении стандартного отклонения средних результатов кластеризации целевой функции при одновременном сокращении оценки усадки Джеймса–Штейна. **3.** Исходя из результатов проведенных экспериментов, предпочтительно нормализовать допустимые значения параметра для повышения точности кластеризации, используя сокращения оценки усадки Джеймса–Штейна.

Финансирование работы. Результаты исследования получены в рамках реализации гранта НШ-421.2022.4 Совета по грантам Президента РФ по государственной поддержке ведущих научных школ РФ и гранта 075-15-2022-1121 Минобрнауки РФ («Мегагрант»).

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

СПИСОК ЛИТЕРАТУРЫ

1. Ершов И. А., Воскобойникова О. Б., Стукач О. В. Кластерный анализ процессов в полупроводниковом производстве // Динамика систем, механизмов и машин. 2016. Т. 2. № 1. С. 178.
2. Rozhnov I., Orlov V., Kazakovtsev L. Ensembles of clustering algorithms for problem of detection of homogeneous production batches of semiconductor devices // School-Seminar on Optimization Problems and their Applications. 2018. V. 2098. P. 338.
<http://ceur-ws.org/Vol-2098/paper29.pdf>
3. Oti E. U., Olusola M. O., Eze F. C., Enogwe S. U. Comprehensive Review of K-Means Clustering Algorithms // Int. J. of Advances in Scientific Research and Engineering. 2021. V. 7 (8). P. 64.
4. Федосов В. В., Орлов В. И. Минимально необходимый объем испытанных изделий микроэлектроники на этапе входного контроля // Известия высших учебных заведений. Приборостроение. 2011. Т. 54. № 4. С. 58.
5. Ahmatshin F. Selection of free parameter forel-2 algorithm in the problem of automatic grouping of industrial products by homogeneous production batches // Системы управления и информационные технологии. 2021. P. 28.
<https://doi.org/10.36622/Vstu.2021.86.4.006>
6. Mathai A., Provost S., Haubold H. Factor Analysis // Multivariate Statistical Analysis in the Real and Complex Domains. 2022. P. 679. https://doi.org/10.1007/978-3-030-95864-0_111
7. Na S., Xumin L., Yong G. Research on k-means clustering algorithm: an improved k-means clustering algorithm // In: 2010 Third Int. Symposium on Intelligent Inf. Technology and Security Informatics, Jínggāngshān. P. 63.

8. *Patel V.R., Mehta R.G.* Modified k-Means Clustering Algorithm // Computational Intelligence and Inf. Technology. 2011. V. 250. P. 307.
https://doi.org/10.1007/978-3-642-25734-6_46
9. *Li Y., Wu H.* A clustering method based on K-means algorithm // Physics Procedia. 2012. V. 25. P. 1104.
<https://doi.org/10.1016/j.phpro.2012.03.206>
10. *Perez-Ortega J., Almanza-Ortega N.N., Romero D.* Balancing effort and benefit of K-means clustering algorithms in Big Data realms // PLoS ONE. 2018. V. 13 (9). e0201874.
<https://doi.org/10.1371/journal.pone.0201874>
11. *Aggarwal C.C., Reddy C.K.* Data Clustering Algorithms and Applications. Publisher: CRC Press, 2013.
<https://www.researchgate.net/publication/331534089>
12. *Kazakovtsev L.A., Antamoshkin A.N., Masich I.S.* Fast deterministic algorithm for EEE components classification // IOP Conf. Series: Materials Science and Engineering. 2015. V. 94. P. 012015.
<https://doi.org/10.1088/1757-899X/04/1012015>
13. *Ansari S.A., Darmawan N., Robbi R., Rahmat H.* Using K-means clustering to cluster provinces in Indonesia // J. of Physics Conf. Series. 2018. V. 1028 (1). P. 012006.
14. *Hossain Md., Akhtar Md.N., Ahmad R.B., Rahman M.* A dynamic K-means clustering for data mining // Indonesian J. of Electrical Engineering and Computer Science. 2019. V. 13 (2). P. 521.
<https://doi.org/10.11591/ijeecs.v13.i2.pp521-526>
15. Шкаберина Г.Ш., Казаковцев Л.А., Лу Ж. Модели и алгоритмы автоматической группировки объектов на основе модели k-средних // Сибирский журнал науки и технологий. 2020. Т. 21. № 3. С. 347.
<https://doi.org/10.31772/2587-6066-2020-21-3-347-354>
16. *Kumar S., Tripathi Yo. M., Misra N.* James–Stein type estimators for ordered normal means // J. of Statistical Computation and Simulation. 2006. V. 75. P. 501.
<https://doi.org/10.1080/00949650412331272877>
17. *Tong T., Jang H., Wang Y.* James–Stein type estimators of variances // J. of Multivariate Analysis. 2012. V. 107. P. 232.
<https://doi.org/10.1016/j.jmva.2012.01.019>
18. *Gao J., Hitchcock D.B.* James–Stein shrinkage to improve k-means cluster analysis // Computational Statistics & Data Analysis. 2010. V. 54. P. 2113.
<https://doi.org/10.1016/j.csda.2010.03.018>