УДК 004.8
DOI: 10.17586/0021-3454-2020-63-11-1003-1011

ИССЛЕДОВАНИЕ АЛГОРИТМОВ ПОТОКОВОЙ КЛАСТЕРИЗАЦИИ ПРИ РЕШЕНИИ ЗАДАЧИ АНАЛИЗА ДАННЫХ ТЕЛЕМЕТРИИ МАЛЫХ КОСМИЧЕСКИХ АППАРАТОВ

В. Ю. Скобцов, Н. А. Новоселова

Объединенный институт проблем информатики НАН Беларуси, 220012, Минск, Беларусь
E-mail: vasko_vasko@mail.ru

Рассматривается задача анализа данных телеметрии бортовой аппаратуры малых космических аппаратов с целью идентификации состояний ее функционирования. Выполнено исследование алгоритмов потоковой кластеризации при решении данной задачи. Использование таких алгоритмов позволяет выделить кластерную структуру данных, а также проследить ее динамику совместно с автоматическим обнаружением резких изменений, связанных как со сменой состояний процесса функционирования систем бортовой аппаратуры, так и с возможным появлением отказов в их работе.

Ключевые слова: *бортовая аппаратура малых космических аппаратов, данные телеметрии, потоковая кластеризация, микрокластеры, макрокластеры, критерии валидации*

Введение. Применение методов анализа данных телеметрии бортовой аппаратуры (БА) малых космических аппаратов (МКА) на основе алгоритмов машинного обучения позволяет на новом научно-техническом уровне решить задачу идентификации состояний ее функционирования и тем самым повысить эффективность управленческих и эксплуатационных решений, принимаемых наземным комплексом управления МКА. На основе анализа динамического потока данных телеметрии можно определять различные состояния работы аппаратуры за период функционирования МКА. Использование определенных таким образом состояний, которые характеризуют как штатную, так и нештатную работу БА МКА, позволяет в дальнейшем осуществлять прогноз и самого состояния, и связанных с ним значений показателей надежности БА МКА.

Телеметрические измерения представляют собой поток данных, для анализа которых необходимо использовать специальные алгоритмы, работающие в режиме онлайн и осуществляющие только один проход по данным. Стандартные статические алгоритмы кластеризации, например *k*-средних, не всегда являются в этом случае приемлемыми, так как характеризуются многопроходностью. На некоторых классах множеств для алгоритма *k*-средних сложность по времени сходимости равна $2^{\Omega(\sqrt{n})}$, где n — количество объектов данных. В случае анализа потока данных статические алгоритмы кластеризации не позволяют анализировать динамику изменения кластерной структуры во времени, и, следовательно, невозможно фиксировать моменты времени, в которые происходит переход состояний сложной системы или резкое изменение режима функционирования, что может быть последствием сбоя или отказа БА МКА. Таким образом, авторами было проведено исследование по применению алгоритма потоковой кластеризации для решения вышеуказанной задачи анализа телеметрии на основе реальных данных телеметрии навигационного устройства МКА. Исследуемые алгоритмы применимы как для статической, так и динамической кластеризации данных.

Алгоритмы потоковой кластеризации. В настоящее время актуальным направлением исследований являются разработки в области анализа потоковой информации [1]. Это обусловлено возросшим количеством динамически генерируемых данных, таких как данные

мониторинга компьютерных сетей, интерактивные Web-данные, телекоммуникационные данные, сенсорная информация с различных устройств, данные телеметрии и т.д.

Поток данных может быть формализован как упорядоченная последовательность объектов или временных точек данных $Y = \langle y_1, y_2, y_3, ... \rangle$. Как правило, объектами данных являются многомерные точки или векторы значений некоторых измерений, которые могут содержать как числовые, так и номинальные переменные, а также представлять собой более сложную информацию, как например графы или текст.

К настоящему времени разработан ряд однопроходных алгоритмов кластеризации для потоков данных [2]. Применение этих алгоритмов позволяет решить проблемы масштабируемости традиционных кластерных алгоритмов, однако не позволяет проследить эволюцию данных, т.е. возможность исследования кластеров в различных интервалах времени.

Для анализа динамического поведения кластеров или состояний систем БА МКА предлагается использовать новый подход [3], который основан на проведении двухуровневой кластеризации: онлайн-кластеризации с периодическим сохранением детализированных статистических данных и офлайн-кластеризации, использующей сохраненные статистические данные анализируемого потока. Офлайн-компонент может применяться пользователем для анализа динамических кластерных структур потока данных через предоставление ему интерфейса для ввода различных параметров, как например интервал времени для анализа или число кластеров.

К настоящему времени имеется несколько алгоритмов потоковой динамической кластеризации, где реализован вышеописанный подход: CluStream, ClusTree, DenStream, DStream, DBSTREAM [4]. Далее опишем подробно один из наиболее интересных алгоритмов CluStream, который использует эффективную схему сохранения промежуточных результатов кластеризации в ходе обработки потока данных. В процессе выполнения онлайн-кластеризации в микрокластерах сохраняется статистическая информация о поступающих данных. Микрокластер *M* для набора *n* объектов потока данных $X_{i_1}, X_{i_2}, ..., X_{i_n}$, которые являются *d*-мерными векторами признаков во временных точках $T_{i_1}, T_{i_2}, ..., T_{i_n}$, представляет собой кортеж $(\overline{CF2^x}, \overline{CF1^x}, CF2^t, CF1^t, n)$, где $\overline{CF2^x}$ соответствует *d*-мерному вектору, где хранится сумма квадратов значений объектов данных, *p*-й элемент вектора определяется как $\sum_{i=1}^{n} (x_{i_j}^p)^2$; $\overline{CF1^x}$ соответствует сумме значений объектов данных, где *p*-й элемент *d*-мерного вектора равен $\sum_{i=1}^{n} x_{i_j}^p$; в $CF2^t$ и $CF1^t$ хранится соответственно сумма квадратов и сумма временных точек $T_{i_1}, T_{i_2}, ..., T_{i_n}$; *n* —количество объектов данных в кластере.

Микрокластеры сохраняются согласно оригинальной пирамидальной технологии, которая позволяет избежать сохранения всех микрокластеров в каждый момент времени. Формирование микрокластеров осуществляется в процессе онлайн-кластеризации. Предполагается поддержание заданного количества *q* микрокластеров $M_1, M_2, ..., M_q$ в каждый момент времени. Начальные *q* микрокластеров генерируются с использованием стандартного алгоритма *k*-средних для заданного количества начальных точек потока данных. Далее выполняется онлайн-процесс обновления кластеров, согласно следующим шагам.

Шаг 1. Каждый последующий объект потока либо объединяется с уже имеющимся кластером, либо формирует новый кластер. Для этого рассчитывается $d(M_j, \bar{X}_{i_k})$ — расстояние

от объекта $\bar{X}_{i_k}$ до центроида каждого имеющегося микрокластера $M_1, M_2, ..., M_q$. Центроид каждого микрокластера хранится в его описании $(\overline{CF2^x}, \overline{CF1^x}, CF2^t, CF1^t, n)$.

Шаг 2. Определяется ближайший микрокластер M_p к объекту $\bar{X}_{i_k}$. Так как объект не может быть с определенностью идентифицирован как принадлежащий новому кластеру или являющийся выбросом в данных до поступления большего количества данных потока, то в случае невыполнения критерия принадлежности к ближайшему кластеру образуется новый кластер с оригинальным *id*. Критерием принадлежности к ближайшему кластеру является положение объекта внутри области, определяемой значением максимальной границы микрокластера M_p, в этом случае объект добавляется к кластеру M_p. Значение максимальной границы равно произведению среднеквадратического отклонения объектов кластера M_p и фактора *f*.

Шаг 3. В случае невыполнения критерия формируется новый кластер с новым *id*, который позволяет идентифицировать его в ходе последующего процесса кластеризации потока данных.

Шаг 4. При образовании нового кластера и с целью поддержания определенного количества *q* микрокластеров в каждый момент времени выполняется сокращение количества кластеров. Сокращение происходит либо за счет удаления прежнего кластера, либо объединения двух прежних кластеров. Сначала выполняется проверка на присутствие выбросов среди прежних кластеров, один из которых может быть в последующем удален. Первый способ заключается в удалении микрокластера с наименьшим значением среднего времени поступления *m* объектов в каждый из текущих кластеров, второй способ — в удалении микрокластера с минимальным количеством объектов.

Офлайн-этап кластеризации заключается в формировании макрокластеров на множестве имеющихся микрокластеров. Данный этап позволяет пользователю исследовать кластерную структуру данных в различных временных интервалах. В процессе формирования макрокластеров не используются данные потока как таковые, а используется только компактно сохраненная статистическая информация микрокластеров. Предполагается, что пользователь может задавать временной интервал *h* и количество макрокластеров *k*, которые необходимо сформировать. При задании временного интервала необходимо первоначально найти микрокластеры, специфичные именно для этого интервала.

Описанный двухуровневый подход к динамической кластеризации потока данных позволяет анализировать эволюцию кластеров в определенных временных интервалах.

Предварительная обработка данных. Для анализа состояний процесса функционирования навигационного устройства МКА были использованы временные ряды реальной телеметрии навигационной системы БА МКА. Первоначально была выполнена обработка файлов для формирования общей матрицы входных данных, где в строках содержатся данные отдельных наблюдений в конкретной временной точке, а в столбцах — значения параметров (телеметрических измерений), которые характеризуют наблюдение. Можно сказать, что в столбцах расположены временные ряды последовательности значений отдельных параметров. Сформированная матрица данных содержит 502 118 объектов или временных точек и характеризуется значениями параметров $X_i, i = 1, ..., 10$. Один из параметров представляет собой класс анализируемого объекта данных, где метки класса соответствуют корректной и возможной некорректной работе устройства МКА (0 — корректно, 1 — некорректно).

В данном исследовании применен встроенный алгоритм отбора наиболее информативных признаков векторов телеметрии анализируемой системы [5].

Согласно алгоритму отбора признаков выполняются следующие шаги:

— использование процедуры прямого перебора для выбора очередного подмножества признаков;

— выполнение с использованием алгоритма *k*-средних кластеризации данных, которые заданы значениями признаков из выбранного подмножества;

— оценка результата кластеризации с использованием показателей качества разбиения — отношение суммы квадратов межкластерных расстояний к сумме квадратов расстояний:

$$\sum_{j=1}^{p}\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{i'=1}^{n}d_{i,i'}^{j}-\sum_{i,i'\in C_k}^{n}d_{i,i'}^{j}\right)\Bigg/\sum_{i=1}^{p}\frac{1}{n}\sum_{i=1}^{n}\sum_{i'=1}^{n}d_{i,i'}^{j}.$$

Критерием останова алгоритма является стабилизация показателя качества разбиения на очередной итерации алгоритма.

В результате в качестве информативных были выбраны четыре параметра данных телеметрии STRQ1-STRQ4 — значения кватернионов, которые характеризуют анализируемый набор данных.

Результаты применения алгоритмов динамической кластеризации. Для анализа данных телеметрии навигационного устройства МКА были использованы два алгоритма потоковой кластеризации CluStream и DStream и проведен их сравнительный анализ. Первоначально с использованием алгоритмов кластеризации были выделены состояния процесса функционирования устройства на всем наборе данных векторов телеметрии, состоящих из четырех выбранных параметров.

На первом этапе были выделены микрокластеры, которые, в свою очередь, потом были кластеризованы на предварительно заданное количество макрокластеров $k=4$.

Расположение микро- и макрокластеров после обучения совместно с набором обучающих данных (n=20 000, где n — количество объектов данных) представлено на рис. 1: *а* — алгоритм CluStream, *б* — алгоритм DStream.

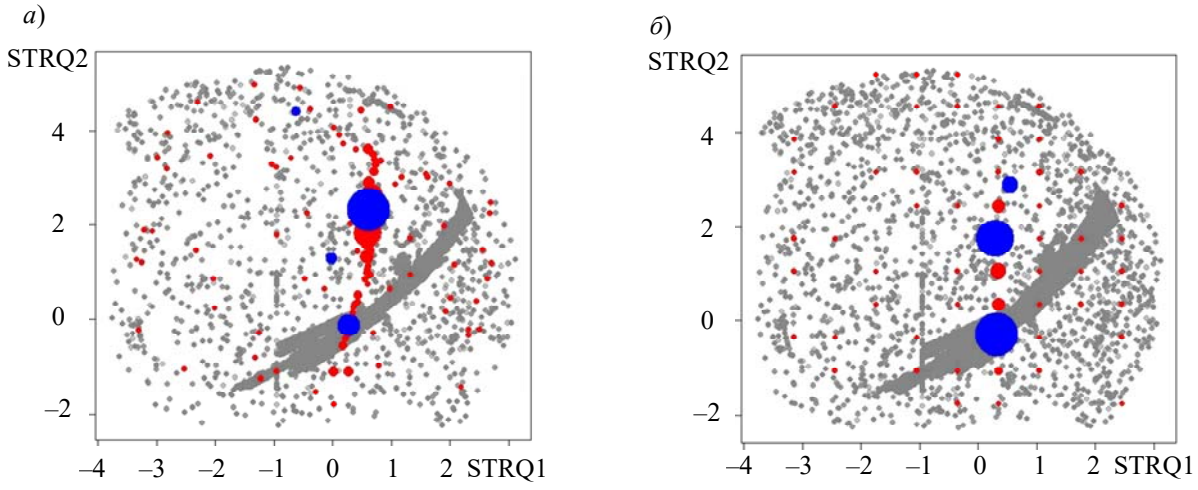

Рис. 1

Количество формируемых микрокластеров изначально равно m=100 для алгоритма CluStream и определяется автоматически (m=93) для алгоритма DStream. Полученные макрокластеры соответствуют четырем состояниям работы навигационного устройства БА МКА.

В таблице представлены результаты оценки качества статической кластеризации с использованием микро- и макрокластеров двух алгоритмов. Оценивались как внутренний критерий качества кластеризации (SSQ — внутрикластерная сумма квадратов отклонений), так и внешние критерии кластеризации с использованием меток класса, задаваемых в данных телеметрии параметром класса (1 — корректная работа устройства, 0 — некорректная). Внешний критерий Purity (чистота кластера) задается для каждого кластера как процентное отно-

шение количества объектов одного из классов, имеющего большинство в кластере, ко всем объектам в кластере. Внешний критерий Precision (точность) определяется с использованием значений TP (true positive), т.е. присвоение двух объектов из одного и того же класса к одному и тому же кластеру; TN (true negative) — присвоение двух объектов из разных классов к двум разным кластерам. Таким же образом определяются значения FP (false positive) и FN (false negative). Далее критерий определяется как $\text{Precision} = \text{TP}/\text{TP} + \text{FP}$.

Критерий Rand определяется на основе построения таблицы сопряженности для заданных классов и полученных результатов кластеризации [6].

Критерий качества кдастеризации	Микрокластеры		Макрокластеры	
	CluStream	DStream	CluStream	DStream
SSQ	2566,7	1236,9	26926,8	23267,8
Purity	0,78	0,81	0,67	0,61
Precision	0,66	0,53	0,59	0,51
Rand	0,65	0,54	0,58	0,52

Согласно таблице критерий качества Purity имеет достаточно большое значение в случае микрокластеров, что говорит об их „чистоте“. Значение внутреннего критерия SSQ меньше для микрокластеров, так как их количество намного превосходит количество макрокластеров.

Графическое представление смены состояний навигационного устройства на основе полученных четырех макрокластеров приведено на рис. 2: *а* — алгоритм CluStream, *б* — алгоритм DStream; здесь ось абсцисс — ось времени (итераций), ось ординат — ось номеров кластеров — состояний: *1* — черный кластер, *2* — красный, 3 — зеленый, 4 — синий. Каждое состояние представляет собой метку кластера, к которому принадлежит объект в данной временной точке (временной ряд размерности 1000).

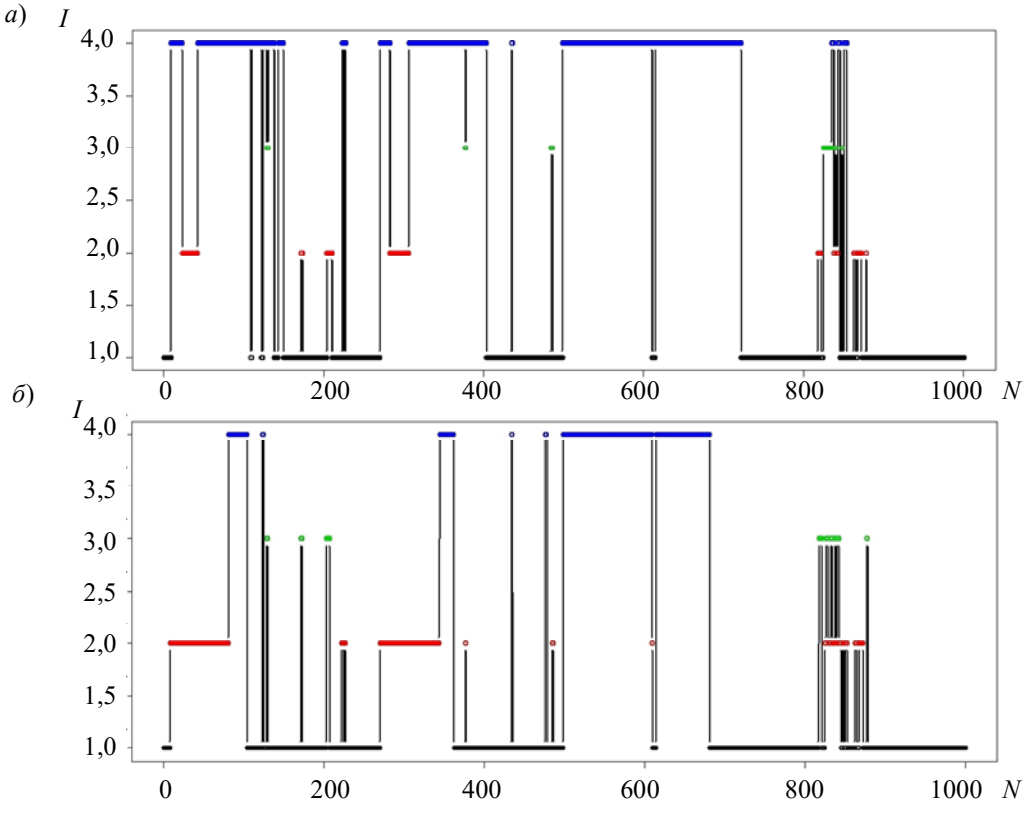

Рис. 2

Четыре кластера в пространстве трех параметров данных телеметрии представлены на рис. 3: *а* — алгоритм CluStream, *б* — алгоритм DStream; здесь номера кластеров — состояний соответствуют: *1* — черному кластеру, *2* — красному, *3* — зеленому, *4* — синему. Исходя из

результатов кластеризации и анализа распределения меток классов по кластерам наиболее чистым является кластер 3 для алгоритма DStream и кластеры 2 и 3 для алгоритма CluStream, в которых содержатся в основном объекты с меткой класса „0“; на рисунке они соответствуют состояниям наименьшей временной продолжительности. Остальные кластеры состоят из объектов обоих классов, причем кластер 4 для CluStream и кластеры 2, 4 для DenStream имеют большее в процентном отношении количество объектов корректной работы. Кроме того, необходимо отметить, что в результате применения алгоритма CluStream получено большее количество „чистых“ кластеров некорректной работы устройства, т.е. кластеров с преобладанием объектов данных с меткой класса „0“.

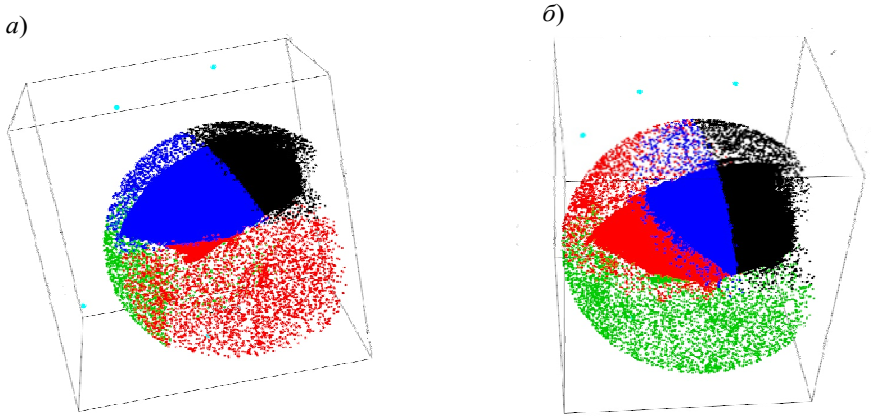

Рис. 3

На втором этапе исследований был проведен анализ набора данных как динамического потока, причем параллельно с обучением кластерной структуры выполнялась динамическая оценка качества кластеризации, которая может быть использована для фиксации резких переходов состояний работы устройства, а также возможных сбоев в работе. Для оценки динамически изменяющейся кластерной структуры весь набор данных был разбит на последовательные интервалы, состоящие из h=100 объектов (интервал оценки). После кластеризации очередного интервала объекты данных последующего интервала были отнесены к ближайшим микрокластерам и для полученного распределения по микрокластерам был рассчитан ряд критериев качества кластеризации [6]. На рис. 4 представлен временной график значений критериев оценки результатов динамической кластеризации (первые 100 000 объектов данных, критерий SSQ): *а* — алгоритм CluStream, *б* — алгоритм DStream. Согласно рисунку имеется несколько резких скачков значений критерия SSQ, их величина для двух алгоритмов различна, что обусловлено результатами кластеризации. Однако временное положение скачков в потоке данных для двух алгоритмов совпадает.

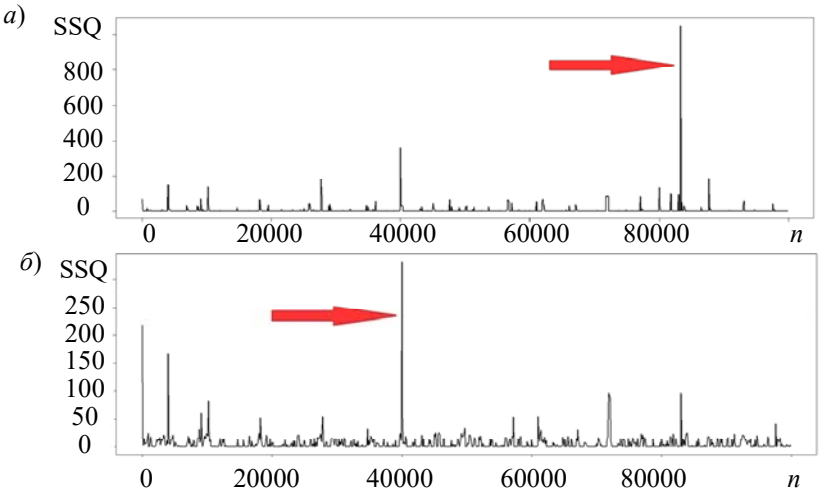

Рис. 4

Изменение значений параметров телеметрии (ТМ) навигационного устройства в эти моменты времени показано на рис. 5: *а* — *n*=40 000, *б* — *n*=83 300. Отсюда следует, что в процессе динамической кластеризации фиксируется переход устройства от достоверной работы к недостоверной. Данный переход характеризуется либо большим разбросом значений параметров телеметрии, либо их резким изменением.

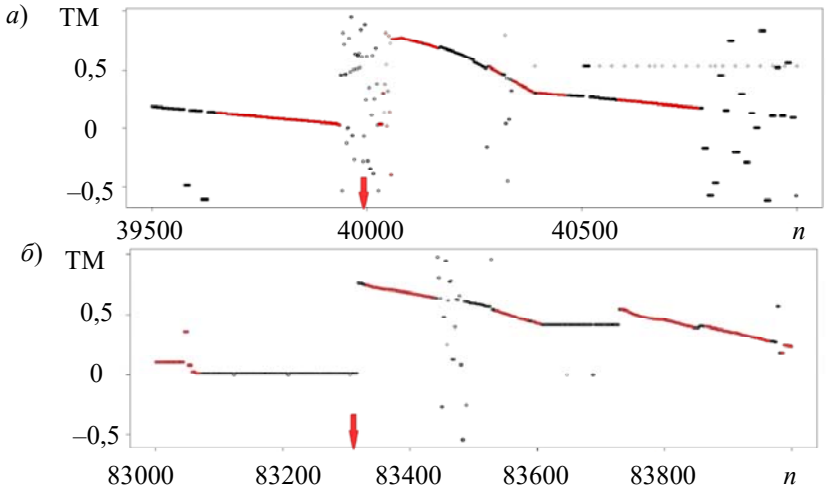

Рис. 5

Резкий скачок значения показателя качества кластеризации означает изменение кластерной структуры данных в указанный момент времени, т.е. в этот момент образуется большое количество новых микрокластеров, и, следовательно, значение критерия оценки SSQ резко увеличивается. На рис. 6 представлена визуализация изменения кластерной структуры данных при поступлении 100 новых объектов данных после того, как обучение уже было произведено до момента времени *n*=40 000: *а* — микрокластеры и новые 100 объектов данных в момент времени *n*=40 000; *б* — микрокластеры в момент времени *n*=40 100.

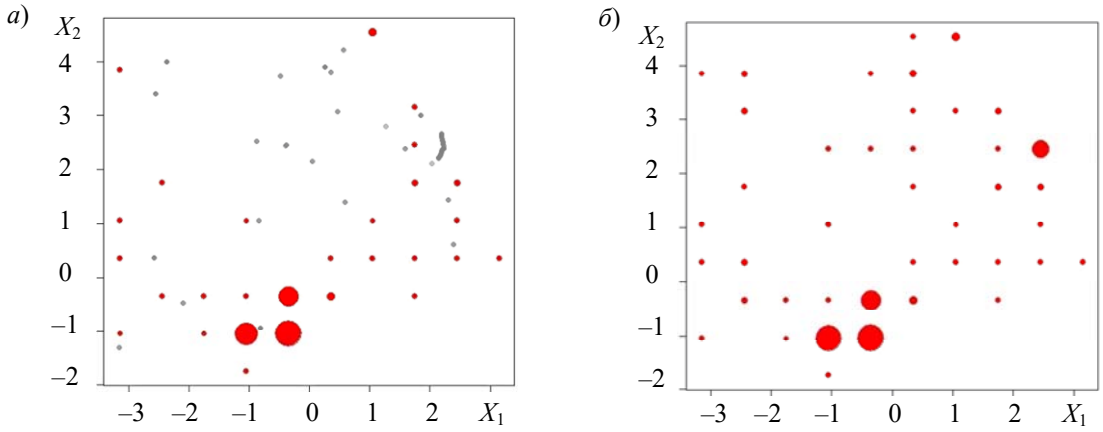

Рис. 6

Как видно из рис. 6, образовалось достаточно большое количество новых микрокластеров, так как значения параметров телеметрии 100 новых объектов значительно отличаются от центров ранее построенных микрокластеров.

Заключение. Приведены результаты исследования по применению алгоритмов динамической кластеризации для идентификации состояний процесса функционирования навигационного устройства БА МКА на основе анализа данных его телеметрии. Обосновано применение данного типа алгоритмов к данным телеметрии, рассматриваемым как непрерывно поступающие динамические потоки данных. На основании проведенного анализа:

— выявлены основные состояния процесса функционирования навигационного устройства, которые соответствуют кластерам, полученным с использованием алгоритмов динамической кластеризации статических данных, что позволило выявить кластерную структуру

данных без необходимости хранения всего набора в оперативной памяти компьютера, а с использованием последовательной процедуры считывания блоков информации из файла;

— выполнена динамическая кластеризация с параллельной пошаговой оценкой критериев качества кластеризации для каждого последующего поступающего блока данных; результаты оценки позволили зафиксировать ряд выбросов значений критерия SSQ, соответствующих резкому изменению поведения навигационного устройства, которое характеризуется разбросом или скачками значений параметров данных телеметрии.

Таким образом, анализ результатов динамического расчета критериев оценки качества кластеризации позволяет фиксировать смену состояний процесса функционирования устройств МКА, а также выявлять различные сбои или отказы в его работе.

Результаты кластерного анализа данных телеметрии, в частности состояния процесса функционирования устройства, полученные с использованием потоковой динамической кластеризации, после предварительной экспертной оценки применимы для построения прогностической модели состояния устройств БА МКА.

Работа выполнена при финансовой поддержке программы Союзного государства „Мониторинг-СГ“.

СПИСОК ЛИТЕРАТУРЫ

1. Data Streams. Models and Algorithms / Ed. *C. Aggarwal.* Springer-Verlag, 2007.
2. *Ailon N., Jaiswal R., Monteleoni C.* Streaming k-means approximation // NIPS: Proc. of the 23rd Annual Conf. 2009. P. 10—18.
3. *Aggarwal C. C.* A framework for diagnosing changes in evolving data streams // Proc. ACM SIGMOD Conf., San Diego, June 9—12, 2003. P. 575—586.
4. *Bifet A., Holmes G., Kirkby R., Pfahringer B.* MOA: Massive online analysis // J. of Machine Learning Research. 2010. N 99. P. 1601—1604.
5. *Liu H., Yu L.* Towards integrating feature selection algorithms for classification and clustering // IEEE Trans. on Knowledge and Data Engineering. 2005. N 17(3). P. 1—12.
6. *Bifet A., de Francisci Morales G., Read J., Holmes G., Pfahringer B.* Efficient online evaluation of big data stream classiffiers // Proc. of the 21st ACM SIGKDD Intern. Conf. on Knowledge Discovery and Data Mining, KDD '15. 2015. P. 59—68.

Сведения об авторах

Вадим Юрьевич Скобцов — канд. техн. наук, доцент; Объединенный институт проблем информатики НАН Беларуси, лаборатория проблем зашиты информации; вед. научный сотрудник; E-mail: vasko_vasko@mail.ru

Наталья Анатольевна Новоселова — канд. техн. наук; Объединенный институт проблем информатики НАН Беларуси, лаборатория биоинформатики; вед. научный сотрудник; E-mail: novos65@mail.ru

Поступила в редакцию
02.10.2020 г.

Ссылка для цитирования: *Скобцов В. Ю., Новоселова Н. А.* Исследование алгоритмов потоковой кластеризации при решении задачи анализа данных телеметрии малых космических аппаратов // Изв. вузов. Приборостроение. 2020. Т. 63, № 11. С. 1003—1011.

INVESTIGATION OF STREAM CLUSTERING ALGORITHMS WHEN SOLVING THE PROBLEM OF SMALL SPACECRAFT TELEMETRY DATA ANALYSIS

V. Yu. Skobtsov, N. A. Novoselova

The Joint Institute for Informatics Problems of the National Academy of Sciences of Belarus, 220012, Minsk, Belarus
E-mail: vasko_vasko@mail.ru

The problem of analysis of telemetry data of small spacecraft onboard equipment aimed at identification of its functioning state, is considered. Algorithms of stream clustering in solving this problem are studied. It is noted that the use of such algorithms makes it possible to single out the cluster data structure, as well as to trace its dynamics together with the automatic detection of abrupt changes associated both with a change in the state of onboard equipment systems functioning, and with the possible appearance of failures in their operation.

Keywords: small satellites onboard equipment, telemetry data, streaming clustering, microclusters, macroclusters, validation criteria

REFERENCES

1. Aggarwal C., ed., *Data Streams. Models and Algorithms*, Springer-Verlag, 2007, 354 p.
2. Ailon N., Jaiswal R., Monteleoni C. *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009*, December 7–10, 2009, Vancouver, British Columbia, Canada, 2009, pp. 10–18.
3. Aggarwal C.C. *Proc. SIGMOD 2003*, June 9–12, 2003, San Diego, pp. 575–586.
4. Bifet A., Holmes G., Kirkby R., Pfahringer B. *Journal of Machine Learning Research*, 2010, no. 99, pp. 1601–1604.
5. Liu H., Yu L. *IEEE Transactions on Knowledge and Data Engineering*, 2005, no. 3(17), pp. 1–12.
6. Bifet A., de Francisci Morales G., Read J., Holmes G., Pfahringer B. *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* KDD '15, 2015, pp. 59–68.

Data on authors

Vadim Yu. Skobtsov — PhD, Associate Professor; The Joint Institute for Informatics Problems of the National Academy of Sciences of Belarus, Laboratory of Information Security Problems; Leading Researcher; E-mail: vasko_vasko@mail.ru

Natalia A. Novoselova — PhD; The Joint Institute for Informatics Problems of the National Academy of Sciences of Belarus, Laboratory of Bioinformatics; Leading Researcher; E-mail: novos65@mail.ru

For citation: Skobtsov V. Yu., Novoselova N. A. Investigation of stream clustering algorithms when solving the problem of small spacecraft telemetry data analysis. *Journal of Instrument Engineering*. 2020. Vol. 63, N 11. P. 1003—1011 (in Russian).

DOI: 10.17586/0021-3454-2020-63-11-1003-1011