УДК 519.237.8
DOI

МЕТОД РАСЧЕТА ЧИСЛА КЛАСТЕРОВ ДЛЯ АЛГОРИТМА K-MEANS

CLUSTERS NUMBER CALCULATING METHOD FOR THE K-MEANS ALGORITHM

В.В. Фролов[1], С.Е. Слипченко[2], О.Ю. Приходько[3]
V.V. Frolov[1], S.E. Slipchenko[2], O.Yu. Prikhodko[3]

[1)] Харьковский национальный университет имени В.Н. Каразина,
Украина, 61022, Харьков, площадь Свободы 4
[2)] НТУ «ХПИ», Украина, 61002, Харьков, ул. Кирпичева, 2
[3)] БГТУ им. В.Г. Шухова, Россия, 308012, г. Белгород, ул. Костюкова 46,

[1)] V.N. Karazin Kharkiv National University, 4 Svobody Sq, Kharkiv, 61022, Ukraine
[2)] NTU «KhPI», 2 Kyrpychova St, Kharkiv, 61002, Ukraine
[3)] BSTU named after V.G. Shukhov, 46 Kostyukova St, Belgorod, 308012, Russian Federation

E-mail: vvicfrol@rambler.ru, serg.slip@gmail.com, prihodko.o.u@gmail.com

Аннотация
В статье предложен метод оценки оптимального числа кластеров для алгоритма k-средних. Метод обеспечивает расчет оптимального количества кластеров для разделения исходного множества на основе анализа нескольких критериев оценки. Основным критерием является динамика перераспределения объектов в кластерах при переходе от одного разбиения к другому. Оценка динамики проводится при расчете нормы матрицы перехода. В качестве дополнительного критерия используется оценка изменения потенциальной энергии объектов внутри кластеров одного и того же разбиения. Вспомогательный критерий определяет количество кластеров в соответствии с характерными точками графиков основного и дополнительного критериев. Суть метода заключается в наборе правил использования основных, дополнительных и вспомогательных критериев. Последовательность выполнения правил реализована в виде функции системы Matlab. Сравнительный анализ показывает, что метод комплексной оценки позволяет повысить точность определения оптимального количества кластеров на 40 %.

Abstract
In article the method of estimate of optimum number of clusters for an algorithm k-means is offered. The method provides calculation of optimum number of clusters for partitioning an source set on the basis of the analysis of several evaluation criteria. The main criterion is dynamics of redistribution of objects in clusters upon transition from one partitioning towards another. Assessment of dynamics is carried out at calculation of norm of matrix of transition. As an additional criterion, an estimate of the change in the potential energy of objects inside clusters of the same partition is used. The auxiliary criterion determines number of clusters according to characteristic points of plots of the main and additional criteria. The essence of a method consists in a rules set of use of the main, additional and auxiliary criteria. The sequence of execution of rules is implemented by way of function of the Matlab system. Contrastive analysis shows that the method of integrated assessment allows to increase the accuracy of determination of optimum number of clusters by 40 %.

Ключевые слова: кластерный анализ, кластер, устойчивость кластеризации, разбиение множества, критерий качества разбиения, k-means, центр кластера, центроид.
Keywords: cluster analysis**,** cluster, clustering stability, partition of a set, partition quality criterion, k-means, cluster center, centroid.

Введение

Методы кластерного анализа находят широкое применение в различных областях знаний, преимущественно для классификации неупорядоченных и разрозненных данных.

В этих задачах при анализе учитывается большое количество признаков, совокупно определяющих на множестве объектов разбиение, анализ которого помогает исследователю или инженеру в принятии решений. Одной из основных проблем применения методов кластерного анализа в практической деятельности является наличие неопределенности при предварительной оценке количества кластеров, на которые разбивается исследуемая область.

Существующие подходы позволяют решить эту задачу с разной степенью эффективности на основе вычисления критериев качества разбиения, которые можно условно разбить на две группы: критерии, где вычисляются метрические характеристики разбиения (метрические); критерии, использующие для оценки разбиения физические аналоги (физические). Характерный пример «физического критерия» приводится в работе [Кольцов, 2017], здесь оптимальное количество кластеров находится при анализе фазовых состояний системы с наименьшей флуктуацией. Литвиненко В.И. [Литвиненко, 2009] использует для автоматической кластеризации данных свойства самоорганизации иммунной системы. Д.С. Шалымов [Шалымов, 2017] оценивает «метрические критерии» по устойчивости кластеризации, с этой точки зрения он выделяет пять критериев, которые основаны на вычислении межкластерных и внутрикластерных расстояний: Calinski – Harabasz; Krzanowski and Lai; Hartigan; Silhouette; GAP.

В работе [Елизаров, Куприянов, 2009] авторы классифицируют методы кластерного анализа на центроидные и нецентроидные, где для центроидных рассматривают следующие критерии качества разбиения: коэффициент разбиения, основанный на матрице принадлежности, максимум которого показывает наилучшее разбиение; модифицированный коэффициент разбиения за вычетом мощности множества кластеров; энтропия разбиения и модифицированная энтропия; критерий эффективности разбиения на основе вычисления внутрикластерных и межкластерных отличий.

А. Ложкинс и В.М. Буре [Ложкинс, Буре, 2016] предлагают метод, который позволяет определить оптимальное количество кластеров с учетом ошибок в данных на основе оценки устойчивости кластеризации. Выбирают такое количество кластеров, которое дает меньше ошибок и выше устойчивость кластеризации.

В работе [Бондарев и др., 2007] авторы оценивают проблему определения оптимального числа кластеров как основную в кластерном анализе для иерархических и итеративных методов кластеризации.

Шокина М.О. [Шокина, 2017] определяет оптимальное количество кластеров, описывающих разбиение множества хромосом, на основе анализа графиков silhouette coefficient. В результате оптимальным будет такое разбиение на кластеры, у которого нет отрицательных значений этого коэффициента.

Яцкив И., Гусарова Л. [Яцкив, Гусарова, 2003] разбивают критерии на две большие группы: глобальные и локальные правила остановки процесса кластеризации. В этой работе авторы относят критерий Calinski-Harabasz к правилам глобальной остановки и отмечают, что он не работает, когда необходимо оценить принципиальную возможность разбиения множества объектов на кластеры.

Миркин Б.Г. [Миркин, 2011] указывает, что количество подходов к оценке необходимого числа кластеров доходит до нескольких десятков и приводит их таксономию по различным признакам.

Все это доказывает, что универсального критерия на все случаи нет. Каждый критерий, применяемый на практике, показывает хорошую результативность при определении количества кластеров в определенных границах, обусловленных предметной областью и применяемым алгоритмом кластеризации. Особенности предметной области выражаются в конкретных параметрах процесса кластеризации и свойствах кластеров, таких как: форма, размеры кластера, расстояние между соседними кластерами, расстояния внутри кластера.

Следовательно, задача разработки метода оценки оптимального количества кластеров для решения определенного круга задач в конкретной предметной области остается актуальной, при условии, что универсальные критерии не позволяют получить

приемлемое решение, т. е. не определяют наличия возможности разбиения на множестве объектов, которое позволит упростить решение конкретных практических задач.

Для подтверждения этого тезиса рассмотрим пример из области биологии. Возьмем DATASET «fisheriris» [MathWorks, Inc., 2019], который содержит измерения длины и ширины от чашелистиков и лепестков трех видов цветков ириса. В этом наборе данных заранее известно количество кластеров. Используем для разбиения на кластеры алгоритм к-средних и наиболее универсальные критерии оценки в системе Матлаб: Calinski-Harabasz (*CH*) [Calinski, Harabasz, 1974]; Davies-Bouldin (*DB*) [Davies, Bouldin, 1974]; Gap (*G*) [Tibshirani et al., 2001]; Silhouette (*SLH*) [Rouseeuw, 1987]. В результате, используя функцию системы Matlab – evalclusters [MathWorks, Inc., 2019], получаем следующие числа кластеров по критериям: *CH*=3; *DB*=2; *G*=5; *SLH*=2. Расчеты показывают, что правильно определяет количество кластеров в наборе только CH критерий. Такая ситуация часто встречается на реальных наборах данных, где анализ предметной области не дает возможности предварительно оценить, хотя бы приблизительно, на какое количество кластеров можно произвести разбиение, и может возникнуть ситуация принятия решения в условиях неопределенности, когда ни один критерий не дает приемлемых результатов.

В основе большинства критериев оценки лежит вычисление внутрикластерного и межкластерного расстояний в рамках одного разбиения. В работе [Московкин, Казимиру, 2017] авторы оценивают качество разбиения по следующим критериям: сумма средних внутрикластерных расстояний должна стремиться к минимуму, а сумма межкластерных расстояний должна стремиться к максимуму. Функции от этих параметров не имеют ярко выраженных экстремумов, и это не позволяет получить единичное значение числа кластеров. Поэтому на практике конструируют оценочную функцию, описывающую различные соотношения этих параметров, так, чтобы получить экстремум.

Зачастую приходится выбирать из допустимого диапазона значений кластеров, характерным примером такого подхода служит применение «метода локтя» в работе [Селуков, Шилов, 2016] или похожий подход в работе [Шалымов, 2009], где авторы аппроксимируют гладкую кривую двумя пересекающимися прямыми и в точке их пересечения определяют количество кластеров.

Целью данной работы является разработка метода оценки оптимального количества кластеров для алгоритма к-средних, обеспечивающего автоматическую классификацию множества проектных решений расчета размерных цепей без предварительного анализа их особенностей с точки зрения эффективности инженерной конструкции. Здесь последовательно решаются следующие задачи: анализ особенностей процесса кластеризации заданного множества решений; разработка метода предварительной оценки разбиения на кластеры с учетом особенностей процесса кластеризации в рамках одного разбиения и динамики изменения параметров процесса при переходе от одного разбиения к другому; проверка эффективности метода для множеств объектов с похожими характеристиками и заранее определенным количеством кластеров.

Разработка метода оценки количества кластеров на множестве решений

Имеем результаты имитационного моделирования расчетов размерных цепей на основе предложенного в работе [Фролов, 2019] метода, которые представлены множеством хромосом $H = \{h : h \in H\}, |H| = 100$, описывающих параметры допусков и отклонений значений каждого звена размерной цепи. Хромосома – вектор-строка следующего вида: $h = (9, H, 11, H, 7, g, 7, g, 7, H, 0, L, 13, h)$. Здесь цифра – номер квалитета, а буква – отклонение. Тогда, указанная хромосома описывает размерную цепь, состоящую из 7 звеньев, где на каждое звено выделяется два гена. Нечетный ген относится к квалитету, а четный к отклонению. Допустимое множество квалитетов в каждом гене:

$$IT = \{it : \exists it \in \{0, 5, 6 \ldots 16\} \wedge it \in IT\}, |IT| = 13.$$

Допустимое множество отклонений для гена:

$$E = \{e : \exists e \in \{a, b, c, d, e, f, g, h, js, m, n, p, r, s, u, x, A, B, C, D, E, F, G, H, L\} \wedge e \in E\}.$$

Во множестве хромосом H выделяем те, у которых погрешность расчета размерной цепи равна нулю, таких в этом множестве 96. Анализ этого множества показывает, что есть стабильное решение по квалитетам, так как они не меняются для всех хромосом множества, а сочетание отклонений в каждой хромосоме разное. Понизим размерность задачи удалением генов, относящихся к квалитетам, поскольку они неразличимы, и последних двух генов отклонений, которые не изменяются в хромосомах множества по условиям проектирования, указанным в работе [Фролов, 2019]. Тогда хромосома будет $h = (H, H, g, g, H)$. Заменим буквы на их номера в опорном множестве гена, таких номеров всего 25 и получим: $h = (a_1, a_2, \ldots a_5), \exists a_i \in \{1,2,\ldots 25\}, i = \overline{1,5}$. Исходным для кластерного анализа будет множество целочисленных векторов $H^R = \{h : h \in H^R \wedge h = (a_1, \ldots a_5)\}, |H^R| = 96$. Необходимо разбить данное множество на кластеры, которые опишут близкие решения, чтобы инженер-проектировщик выбрал по одному решению из каждого кластера.

Решать задачи кластерного анализа будем в системе Matlab с помощью алгоритма к-средних. Порядок решения:

1. Зададим диапазон изменения чисел кластеров, на котором будем искать оптимальное число. Поскольку мощность исходного множества равна 96, примем диапазон $m = \overline{1,96}$ кластеров. Максимальное значение объясняется тем, что минимально кластер может содержать один объект.

2. Вычислим для исходного набора хромосом стандартизованную оценку (z-оценку), используя функцию Matlab – zscore.

3. Для проверки наличия разбиения зададим диапазон чисел кластеров от 1 до 96. Результат на рис. 1 (*CH*=57, *DB*=57, *G*=57, *SLH*=54) показывает, что в одном кластере может быть в среднем до двух хромосом (среднее значение 1,7), количество кластеров принимаем по критерию *CH*=57.

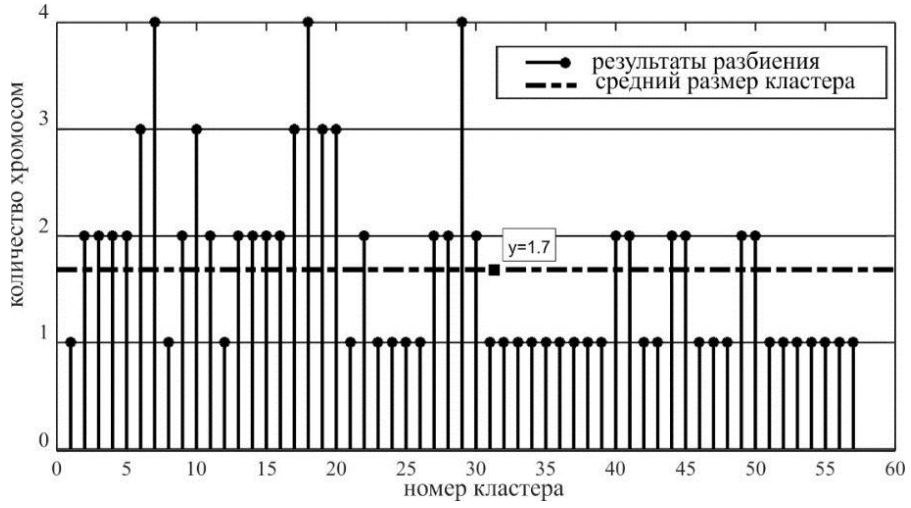


Рис. 1. Результаты разбиения на кластеры множества H^R по критерию CH=57

Fig. 1. Cluster results of a set H^R by criterion CH=57

В данной задаче объектов всего 96, поэтому можно проверить, сколько реально кластеров получилось. Проверка вручную показывает, что данное множество решений можно предельно разбить на 54 подмножества. Согласно этому, правильно оценку оптимального количества кластеров выполнил только критерий SLH=54.

В результате кластеризации было получено предельно возможное разбиение множества на 54 кластера, где собраны неразличимые с точки зрения метрики, применяемой в данном методе, объекты. Такое разбиение (см. рис. 1) не позволяет снизить размерность задачи, чтобы упростить работу проектировщика.

Для вычисления критерия CH в системе Matlab [MathWorks, Inc., 2019] используется следующая формула:

$$VRC_k = \frac{SS_B}{SS_W} \times \frac{N-k}{k-1}, \quad (1)$$

где SS_B – общая дисперсия между кластерами; SS_W – общая дисперсия внутри кластера; N – число объектов, для которых выполняется разбиение; k – количество кластеров.

Общая дисперсия между кластерами рассчитывается

$$SS_B = \sum_{i=1}^{k} n_i \|m_i - m\|^2, \quad (2)$$

где n_i – количество объектов в текущем кластере; m_i – центр текущего кластера; m – общее среднее значение данных выборки.

Общая дисперсия внутри кластера рассчитывается

$$SS_W = \sum_{i=1}^{k} \sum_{x \in c_i} \|x - m_i\|^2, \quad (3)$$

где x – вектор, характеризующий объект, который входит кластер c_i, т. е. это суммарное расстояние между объектами и центром кластера.

В критерии CH есть один недостаток – значение внутрикластерных сумм SS_W здесь находится в знаменателе и в случае, когда существуют кластеры с неразличимыми объектами, как в нашей задаче, оно обращается в нуль, а при делении на нуль будет ошибка при вычислении критерия. Последнее нарушает весь процесс поиска оптимального разбиения.

Выполним анализ изменения SS_B и SS_W с целью определения возможности получения более равномерного разбиения с меньшим числом кластеров при заданном выше диапазоне. На рис. 2, совмещаем два графика, построенные по формулам (2) и (3), где по вертикальной оси откладываем значения SS_B и SS_W, а по горизонтальной оси – текущее количество кластеров k.

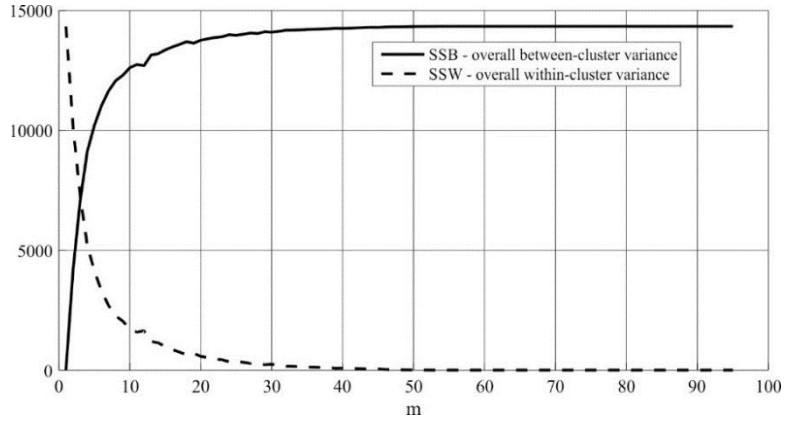


Рис. 2. Изменение общих дисперсий между кластерами SS_B и внутри кластеров SS_W

Fig. 2. Change in main variances between clusters SS_B and within clusters SS_W

Стабилизация результатов этих характеристик наступает после 50 кластеров, что соответствует полученным по критериям значениям. Причем SS_B симметрична SS_W относительно их точки пересечения. С другой стороны, в диапазоне от 10 до 50 кластеров наблюдаются небольшие изменения характеристик по сравнению с диапазоном от 1 до 10,

а это говорит о том, что полученные разбиения вполне можно использовать для практических целей.

Сформулируем задачу кластеризации в терминах, на основе которых будем разрабатывать метод и проводить численные эксперименты. Итеративные методы кластеризации выполняют анализ разбиений объектов по кластерам на основе вычисления расстояний от центров кластеров до объектов множества $O = \{o : o = (pr_1, pr_2, \dots pr_k), o \in O\}$, где k – размерность пространства признаков, pr – отдельный признак кортежа или в матричном виде $O = [O_i]_{|H^R|}$.

В результате кластеризации с помощью функции Matlab имеем следующую информацию о разбиении:

– центры кластера (центроиды) в виде матрицы $C = [c_{i,j}]_{m \times k}$ или $C = [C_i]_{m \times 1}$, где m – количество строк, которое определяется числом кластеров результирующего разбиения, а количество столбцов k – размерностью пространства признаков, C_i – вектор-строка, характеризующая центр кластера, $c_{i,j}$ – значение отдельного признака для центра кластера;

– матрица $D = [d_{i,j}]_{n \times m}, n = |O|$ расстояний всех объектов множества O до центров кластеров;

– внутрикластерные суммы расстояний между центроидом и объектами, входящими в кластер в виде вектора $S = [s_{i,j}]_{1 \times m}$, размерность которого определяется количеством кластеров;

– вектор $IDX = [idx_{i,j}]_{n \times 1}, \forall i : idx_{i,1} \in \{1, \dots m\}$, определяющий принадлежность объектов кластерам.

На основе анализа этих данных необходимо провести оценку необходимого количества кластеров.

Определим понятия, которые составляют основу предлагаемого метода:

1. Расчет оптимального числа кластеров для разбиения исходного множества выполняется на основе анализа нескольких критериев оценки;

2. Комплексная оценка количества кластеров заключается в расчете динамики изменения связей между объектами как в рамках одного разбиения, так и между разбиениями, а также в анализе возможных характерных точек на графиках, отражающих эти изменения.

3. Основной критерий – оценка динамики перераспределения объектов внутри кластеров при переходе от одного разбиения к другому, выраженная в связях между кластерами этих разбиений в виде двудольного графа.

4. Центр кластера является точкой равновесия для всех объектов этого кластера. Тогда положение объектов относительно центра кластера можно оценивать по величине линейной восстанавливающей силы, подчиняющейся закону Гука при постоянном коэффициенте жесткости.

5. Согласно п. 4, степень притяжения объекта к центру кластера будем оценивать величиной работы линейной восстанавливающей силы, чем больше работа, тем дальше от центра кластера находится объект. Следовательно, для расчета степени притяжения объектов к центру кластера можно использовать потенциальную энергию силы упругости пружины с коэффициентом жесткости, равным единице. Потенциальную энергию силы упругости пружины рассчитываем по формулам работы [Никитин, 1990, с. 348–349]:

$$П = -U,\ U = -\frac{c \cdot r^2}{2},\ r^2 = x^2 + y^2 + z^2, \quad (4)$$

где $П$ – потенциальная энергия; U – силовая функция линейной силы упругости; c – коэффициент жесткости пружины, для нашего случая $c = 1$; r – расстояние от точки

равновесия до рассматриваемой точки, x, y, z – координаты точки в трехмерном пространстве.

6. Дополнительный критерий – динамика изменения потенциальной энергии объектов внутри кластеров в рамках разбиения.

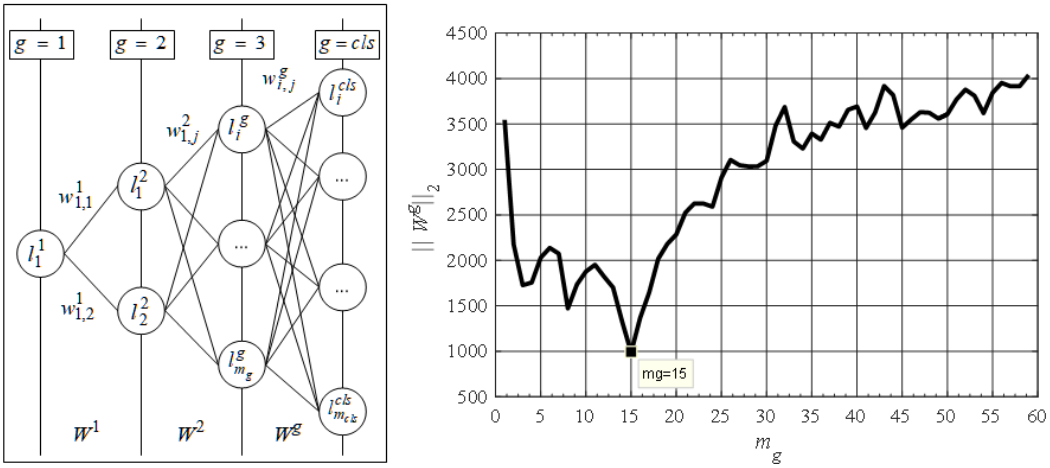
7. Вспомогательный критерий – характерные точки графиков основного и дополнительного критериев.

При расчете основного критерия, согласно п. 3, процесс разбиения на кластеры можно представить в виде нейроподобного элемента на рис. 3а. Каждый слой g этого элемента содержит количество кластеров (нейронов), равное номеру слоя, связи между слоями реализуются по принципу все ко всем в виде двудольного графа. Матрица синаптических коэффициентов W^g отражает расстояние между кластерами разных слоев в количестве составляющих их объектов. Тогда появляется возможность оценить степень близости между всеми кластерами соседних слоев на основе Евклидовой метрики. Множество матриц W^g, назовем их матрицами переходов от одного слоя к другому, будет характеризовать весь процесс разбиения, который можно отразить на графике, рис. 3б,

$$W^g = \left[w_{i,j}^g\right]_{m_g \times m_{g+1}}, i = \overline{1, m_g},\ j = \overline{1, m_{g+1}},\ m_g = \overline{1, g},\ g = \overline{1, cls-1},\ w_{i,j}^g = d_E\left(l_i^g, l_j^{g+1}\right), \quad (5)$$

где g – номер текущего разбиения; m_g – максимальное количество кластеров для текущего разбиения; i, j – номера соседних разбиений; cls – максимальное количество кластеров для оценки разбиения; l_i^g – количество объектов в кластере.

Для комплексной оценки связей между слоями (см. рис. 3а) используем Евклидову норму матрицы переходов $\left\|W^g\right\|_2$, поскольку она показывает максимальный коэффициент растяжения этой матрицей вектора, характеризующего текущее разбиение.



а) б)

Рис. 3. Модель динамики процесса разбиения

Fig. 3. Model of the decomposition process dynamics

Евклидову норму вычисляем так

$$\left\|W^g\right\|_2 = \lambda_1, \quad (6)$$

где λ_1 – сингулярное число матрицы. Тогда на DATASET s1 [Fränti, Virmajoki, 2006] определяем оптимальное количество кластеров, равное 15 (см. рис. 3б). Для нашего случая минимум нормы будет при 10 кластерах.

Опишем расчет дополнительного критерия. С учетом обозначений, принятых в данной работе, формула потенциальной энергии (4) для одного объекта перепишется следующим образом:

$$П_{p,i} = \frac{1}{2} \cdot d_E\left(O_i, C_p\right)^2, \tag{5}$$

где d_E – Евклидово расстояние между p-им центром кластера и i-им объектом, C_p – вектор-строка, описывающая p-й центр кластера.

Рассчитываем потенциальные энергии для всех объектов кластера по (5) и выбираем для характеристики кластера энергию объекта $П_p^*$, который расположен наиболее близко к его центру

$$П_p^* = \min_{i \in I}\left\{П_{p,i}\right\}, I = \overline{1, l_p}, \tag{6}$$

где l_p – количество объектов в кластере. Для разбиения с учетом (6) получаем множество таких энергий $П^g = \left\{П_p : П_p = П_p^*, p = \overline{1, m_g}\right\}$. Поэтому для корректной характеристики всего разбиения используем статистические характеристики: среднее значение энергии $\overline{П}^g$ и среднеквадратическое отклонение значений энергии $\sigma_П^g$

$$\overline{П}^g = \frac{1}{m_g} \cdot \sum_{p=1}^{m_g} П_p, \sigma_П^g = \sqrt{\frac{1}{m_g - 1} \cdot \sum_{p=1}^{m_g}\left(П_p - \overline{П}^g\right)^2}. \tag{7}$$

На основе (7) для характеристики всего разбиения можно будет записать такой критерий:

$$KST_g = \overline{П}^g + \sigma_П^g. \tag{8}$$

Оптимальное количество кластеров должно соответствовать минимуму критерия (8), поскольку в нем будет наблюдаться минимум потенциальной энергии разбиения

$$KST^* = \min_{g \in G}\left\{KST_g\right\}, G = \overline{1, cls}. \tag{9}$$

Проведем эксперимент с вычислением критерия по формулам (5)–(8) на DATASET s1 [Fränti,Virmajoki, 2006], где известно оптимальное количество кластеров $m^* = 15$. На рис. 4 изображен график изменения критерия, он имеет минимум в точке, где количество кластеров равно 15. Поскольку график строится со второго разбиения, для определения оптимального количества кластеров по (9) необходимо к значению на графике добавить единицу.

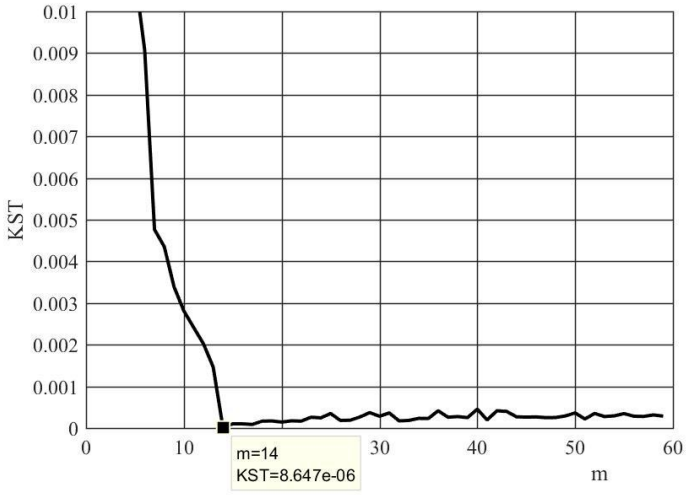


Рис. 4. Изменение потенциальной энергии (фрагмент)
Fig. 4. Potential energy change (fragment)

Результаты и обсуждение

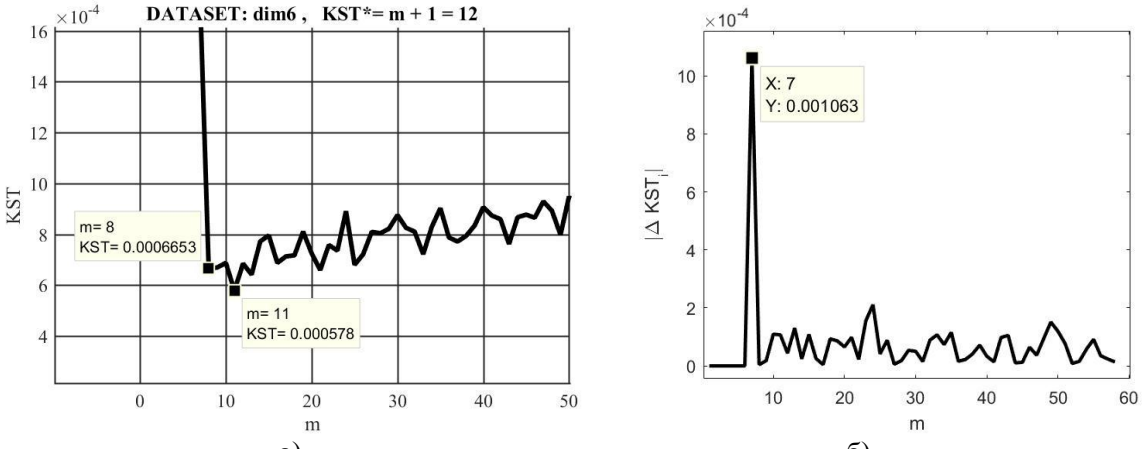
Проведем вычислительные эксперименты на 14 DATASET-ах, чтобы оценить возможности каждого критерия оценки в предлагаемом методе. Результаты сведем в табл. 1, здесь в скобках отражен локальный минимум критерия при наличии спорных результатов.

Таблица 1
Table 1

Результаты численных экспериментов для оценки количества кластеров
The results of numerical experiments to estimate the number of clusters

№	Имя	Источник	Кол-во класт. базовое	KST^*	$\left\|W^g\right\|_2$	$\left\|\Delta KST_i\right\|$	CPC
1	dim2	[Kärkkäinen, Fränti, 2007]	9	10	9	9	9
2	dim4	[Kärkkäinen, Fränti, 2007]	9	15	9	9	9
3	dim 6	[Kärkkäinen, Fränti, 2007]	9	12	(9) – 50	9	9
4	dim 12	[Kärkkäinen, Fränti, 2007]	9	12	(9) – 52	9	9
5	s1	[Fränti,Virmajoki, 2006]	15	15	15	15	15
6	s2	[Fränti,Virmajoki, 2006]	15	15	15	5	15
7	a1	[Kärkkäinen, Fränti, 2002]	20	20	21	7	21
8	a2	[Kärkkäinen, Fränti, 2002]	35	38	(8) – 38	4	38
9	dim032	[Fränti et al., 2006]	16	16	16	16	16
10	dim064	[Fränti et al., 2006]	16	16	16	16	16
11	meas	[MathWorks, Inc., 2019]	3	55	3	20	3
12	m1	–	–	57	10	57	10
13	dim15	[Kärkkäinen, Fränti, 2007]	9	10	9	9	9
14	unbalance	[Rezaei, Fränti, 2016]	8	10	5	9	5

Основной критерий $\left\|W^g\right\|_2$ в таблице 1 дает противоречивые результаты для 3, 4 и 8-го набора данных, поскольку не имеет явно выраженного минимума, такого как на рис. 4 (5-й набор данных в таблице). В этом случае необходимо оценить KST^* потенциальную энергию кластеров внутри разбиения для набора № 8. Она имеет минимум на 38 кластерах, что совпадает с одним из минимумов основного критерия, поэтому принимаем 38 кластеров. В наборах 3 и 4 минимум KST^* находится между локальными экстремумами основного критерия, поскольку эти экстремумы не совпадают с характерной точкой, где происходит резкое изменение потенциальной энергии, что отображено на рис. 5а для набора 3.



а) б)

Рис. 5. Характерные точки на графике потенциальной энергии кластеров
Fig. 5. Characteristic points on the graph of potential energy of clusters

Здесь в точке $m=8$ наблюдается резкое изменение потенциальной энергии, а минимальное значение находится дальше. В этом случае для выявления характерной точки на графике используем конечную разность первого порядка (см. рис. 5б), взятую по модулю. Характерная точка находится в максимуме. Поскольку конечная разность сдвигает значения на единицу, к полученной цифре необходимо добавить 2, чтобы получить оптимальное количество кластеров (см. рис. 5б).

Обозначим комплексный критерий разбиения *CPC* (complex partitioning criterion) и опишем алгоритм его вычисления:

1. Формируем кластеры для каждого уровня g с помощью алгоритма к-средних;
2. Рассчитываем матрицы переходов W^g между уровнями g;
3. Рассчитываем нормы матрицы переходов $\left\|W^g\right\|_2$;
4. Проверяем количество минимумов на графике основного критерия;
5. Существует ярко выраженный глобальный минимум, возвращаем оптимальное количество кластеров, в противном случае переходим к пункту 6.
6. Вычисляем потенциальную энергию кластеров.
7. Если минимум потенциальной энергии совпадает с одним из локальных минимумов п. 4, возвращаем оптимальное количество кластеров по совпадению, в противном случае переходим в п. 8;
8. Определяем характерные точки графика потенциальной энергии по модулю конечных разностей первого порядка.
9. Выбираем максимум из п. 8 и возвращаем оптимальное количество кластеров, в противном случае возвращаем 0.

Данный алгоритм реализован в виде функции системы Matlab, чтобы обеспечить автоматизированный расчет оптимального количества кластеров. Для сравнительного анализа были рассчитаны значения количества кластеров по критериям, используемым в Matlab, для заданных наборов данных в табл. 2.

Таблица 2
Table 2

Результаты численных экспериментов для оценки количества кластеров в Matlab
Results of numerical experiments to estimate the number of clusters in Matlab

Имя	Критерии			
	CH	DB	Gap	SH
dim2	10	10	10	10
dim4	10	9	10	10
dim6	10	10	12	10
dim12	10	9	10	9
s1	15	15	15	15
s2	15	15	15	15
a1	21	17	20	19
a2	38	34	37	34
dim032	16	16	16	16
dim064	16	16	19	16
meas	2	2	17	2
dim15	9	9	12	10
unbalance	25	6	25	6
Правильно, %	38	54	30	38

Предлагаемый в данной статье метод дает 76 % правильных результатов. В сравнении с реализованными в Matlab критериями (см. табл. 2), это лучший результат. Причем основной критерий метода дает 61 % правильных ответов на наборах данных, согласно табл. 1.

Заключение

Метод комплексной оценки количества кластеров для алгоритма к-средних позволяет повысить точность определения оптимального количества кластеров на 40 %, поскольку в отличие от других выполняет анализ изменения параметров разбиения как в рамках одного разбиения, так и в динамике, при переходе от одного разбиения к другому. Причем основным критерием является оценка динамики изменения параметров при переходе от одного разбиения к другому на основе норм матриц переходов $\left\|W^g\right\|_2$, поскольку практика применения показывает, что это дает наиболее достоверный результат при анализе данных (порядка 60 % правильных ответов). Остальные критерии используются для более подробного анализа процесса разбиения при возникновении спорных результатов по основному критерию, что позволяет повысить точность определения числа кластеров.

Список литературы

1. Бондарев В.А., Лисица А.В., Меньшутина Н.В. 2007. Применение правил остановки кластерного анализа в случае слабой и сильной иерархии кластеров на примере белковых структур. Успехи в химии и химической технологии. Т. 21. 1 (69): 105–109. URL: https://cyberleninka.ru/article/n/primenenie-pravil-ostanovki-klasternogo-analiza-v-sluchae-slaboy-i-silnoy-ierarhii-klasterov-na-primere-belkovyh-struktur (дата обращения: 23.10.2019).

2. Елизаров С.И., Куприянов М.С. 2009. Проблема определения количества кластеров при использовании методов разбиения. Изв. вузов. Приборостроение. 52 (12): 3–8. URL: https://cyberleninka.ru/article/n/problema-opredeleniya-kolichestva-klasterov-pri-ispolzovanii-metodov-razbieniya (дата обращения: 22.10.2019).

3. Кольцов С.Н. 2017. Термодинамический подход к проблеме определения числа кластеров на основе тематического моделирования. Письма в журнал технической физики. 43 (12): 90–95. URL: https://elibrary.ru/item.asp?id=29359329 (дата обращения: 22.10.2019).

4. Литвиненко В.И. 2009. Кластерный анализ данных на основе модифицированной иммунной сети. УСиМ. (1): 54–61. URL: http://usim.irtc.org.ua/arch/2009/1/8.pdf (дата обращения: 30.11.2019).

5. Ложкинс А., Буре В.М., 2016. Вероятностный подход к определению локально-оптимального числа кластеров. Вестник СПбГУ. Серия 10. Прикладная математика. Информатика. Процессы управления. (1): 28–37. URL: https://cyberleninka.ru /article/n/veroyatnostnyy-podhod-k-opredeleniyu-lokalno-optimalnogo-chisla-klasterov (дата обращения: 22.10.2019).

6. Московкин В.М., Казимиру Эринелту. 2017. Матричная кластеризация как кластеризация матриц одинаковой размерности. Научные ведомости БелГУ. Серия: Экономика. Информатика. 23 (272): 123–127. URL: https://elibrary.ru/item.asp?id=32265026 (дата обращения: 22.10.2019).

7. Миркин Б.Г. 2011. Методы кластер-анализа для поддержки принятия решений: обзор. М., Изд. дом Национального исследовательского университета «Высшая школа экономики», 88. URL: https://www.hse.ru/data/2011/05/19/1213868030/WP7_2011_03f.pdf (дата обращения: 25.10.2019).

8. Никитин Н.Н. 1990. Курс теоретической механики. 5-е изд., перераб. и доп. М., Высшая школа, 607.

9. Селуков Д.А., Шилов В.С. 2016. Нахождение оптимального числа кластеров «методом локтя». Инновационные технологии: теория, инструменты, практика. 1: 107–111. URL: https://elibrary.ru/item.asp?id=28990633 (дата обращения: 25.10.2019).

10. Фролов В.В. 2019. Проектный расчет размерных цепей на основе имитационного моделирования. Вестник витебского государственного технологического университета. 2 (37): 76–88. URL: https://elibrary.ru/item.asp?id=41653699 (дата обращения: 16.01.2020).

11. Шалымов Д.С. 2009. Рандомизированный метод определения количества кластеров на множестве данных. Научно-технический вестник Санкт-Петербургского государственного университета информационных технологий, механики и оптики. 5 (63): 111–116. URL: https://cyberleninka.ru/article/n/randomizirovannyy-metod-opredeleniya-kolichestva-klasterov-na-mnozhestve-dannyh. (дата обращения: 22.10.2019).

12. Шокина М.О. 2017. Применение алгоритма k-means++ для кластеризации последовательностей с неизвестным количеством кластеров. Новые информационные технологии в автоматизированных системах. (20). URL: https://cyberleninka.ru /article/n/primenenie-algoritma-k-means-dlya-klasterizatsii-posledovatelnostey-s-neizvestnym-kolichestvom-klasterov (дата обращения: 22.10.2019).

13. Яцкив И., Гусарова Л. 2003. Методы определения количества кластеров при классификации без обучения. Transport and Telecommunication. 4 (1): 23–28. URL: http://www.tsi.lv/sites/default/files/editor/science/Research_journals/Tr_Tel/2003/V1/yatskiv_gousarova.pdf (дата обращения: 25.10.2019).

14. MathWorks. 2019. Calinski-Harabasz criterion clustering evaluation object. URL: https://www.mathworks.com/help/stats/clustering.evaluation.calinskiharabaszevaluation-class.html (accessed 25 October 2019).

15. Fränti P., Virmajoki O. 2006. Iterative shrinking method for clustering problems. Pattern Recognition. 39 (5): 761–765.

16. Fränti P., Virmajoki O., Hautamäki V. 2006. Fast agglomerative clustering using a k-nearest neighbor graph. IEEE Trans. on Pattern Analysis and Machine Intelligence. 28 (11): 1875–1881.

17. Rezaei M., Fränti P. 2016. Set-matching measures for external cluster validity. IEEE Trans. on Knowledge and Data Engineering. 28 (8): 2173–2186.

18. Kärkkäinen I., Fränti P. 2007. Gradual model generator for single-pass clustering. Pattern Recognition. 40(3): 784–795.

19. Calinski T., Harabasz J. 1974. A dendrite method for cluster analysis. Communications in Statistics. 3(1): 1–27.

20. Davies D.L., Bouldin D.W. 1979. A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Intelligence. PAMI-1 (2): 224–227.

21. Tibshirani R., Walther G., Hastie T. 2001. Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B. 63 (2): 411–423.

22. Rouseeuw P.J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics. 20 (1): 53–65.

References

1. Bondarev V.A., Lisica A.V., Men'shutina N.V. 2007. Primenenie pravil ostanovki klasternogo analiza v sluchae slaboj i sil'noj ierarhii klasterov na primere belkovyh struktur. [Application of the rules for stopping cluster analysis in the case of a weak and strong hierarchy of clusters using protein structures as an example] Uspehi v himii i himicheskoj tehnologii. [Advances in chemistry and chemical technology] T. 21. 1 (69): 105–109 Available at: https://cyberleninka.ru/article/n/primenenie-pravil-ostanovki-klasternogo-analiza-v-sluchae-slaboy-i-silnoy-ierarhii-klasterov-na-primere-belkovyh-struktur (accessed 23 November 2019).

2. Elizarov S.I., Kuprijanov M.S. 2009. Problema opredelenija kolichestva klasterov pri ispol'zovanii metodov razbienija. [The problem of determining the number of clusters when using partitioning methods.] Izv. vuzov. Priborostroenie. [University News. Instrument making]. 52 (12): 3–8. Available at: https://cyberleninka.ru/article/n/problema-opredeleniya-kolichestva-klasterov-pri-ispolzovanii-metodov- razbieniya (accessed 22 October 2019).

3. Kol'cov S.N. 2017. Termodinamicheskij podhod k probleme opredelenija chisla klasterov na osnove tematicheskogo modelirovanija. [Thermodynamic approach to the problem of determining the number of clusters based on thematic modeling] Pis'ma v zhurnal tehnicheskoj fiziki. [Technical Physics Letters] 43 (12): 90–95. Available at: https://elibrary.ru/item.asp?id=29359329 (accessed 22 October 2019).

4. Litvinenko V.I. 2009. Klasternyj analiz dannyh na osnove modificirovannoj immunnoj seti. [Cluster data analysis based on a modified immune network.] USiM [Control systems and computers]. (1): 54–61. Available at: http://usim.irtc.org.ua/arch/2009/1/8.pdf (accessed 30 November 2019).

5. Lozhkins A., Bure V.M., 2016. Verojatnostnyj podhod k opredeleniju lokal'no-optimal'nogo chisla klasterov. [The probabilistic method of finding the local-optimum of clustering] Vestnik SPbGU. Serija 10. Prikladnaja matematika. Informatika. Processy upravlenija. [Bulletin of St. Petersburg State University. Applied Mathematics. Computer science. Management processes] (1): 28–37. Available at: https://cyberleninka.ru /article/n/veroyatnostnyy-podhod-k-opredeleniyu-lokalno-optimalnogo-chisla-klasterov (accessed 22 October 2019)

6. Moskovkin V.M., Kazimiru Jerineltu. 2017. Matrichnaja klasterizacija kak klasterizacija matric odinakovoj razmernosti. [Matrix clustering as a clustering matrices of the same dimension] Nauchnye vedomosti BelGU. Serija: Jekonomika. Informatika. [Belgorod State University Scientific Bulletin. Economics. Information technologies.] 23 (272): 123–127. Available at: https://elibrary.ru/item.asp?id= 32265026 (accessed 22 October 2019).

7. Mirkin B.G. 2011. Metody klaster-analiza dlja podderzhki prinjatija reshenij: obzor [Cluster Analysis Methods for Decision Support: An Overview] M., Izd. dom Nacional'nogo issledovatel'skogo

universiteta «Vysshaja shkola jekonomiki», 88. Available at: https://www.hse.ru/data/2011/05/19/1213868030/WP7_2011_03f.pdf (accessed 25 October 2019).

8. Nikitin N.N. 1990. Kurs teoreticheskoj mehaniki. [Theoretical Mechanics Course] 5-e izd., pererab. i dop. M., Vysshaja shkola, 607

9. Selukov D.A., Shilov V.S. 2016. Nahozhdenie optimal'nogo chisla klasterov "metodom loktja". [Finding the optimal number of clusters by method «elbow»] Innovacionnye tehnologii: teorija, instrumenty, praktika. [Innovative technologies: theory, tools, practice] 1: 107–111. Available at: https://elibrary.ru/item.asp?id=28990633 (accessed 25 October 2019).

10. Frolov V.V. 2019. Proektnyj raschet razmernyh cepej na osnove imitacionnogo modelirovanija. [Design Calculation of Dimensional Chains on the Basis of Simulation Modeling] Vestnik of Vitebsk State Technological University. 2 (37): 76–88. Available at: https://elibrary.ru/ item.asp?id=41653699 (accessed 16 January 2020).

11. Shalymov D.S. 2009. Randomizirovannyj metod opredelenija kolichestva klasterov na mnozhestve dannyh. [A randomized method for determining the number of clusters on a data set.] Nauchno-tehnicheskij vestnik Sankt-Peterburgskogo gosudarstvennogo universiteta informacionnyh tehnologij, mehaniki i optiki. [Scientific and Technical Bulletin of the St. Petersburg State University of Information Technologies, Mechanics and Optics.] 5 (63): 111–116. Available at: https://cyberleninka.ru/ article/n/randomizirovannyy-metod-opredeleniya-kolichestva-klasterov-na-mnozhestve-dannyh (accessed 22 October 2019).

12. Shokina M.O. 2017. Primenenie algoritma k-means++ dlja klasterizacii posledovatel'nostej s neizvestnym kolichestvom klasterov. [Application of the k-means ++ algorithm for clustering sequences with an unknown number of clusters.] Novye informacionnye tehnologii v avtomatizirovannyh sistemah. [New information technologies in automated systems.] (20). Available at: https://cyberleninka.ru /article/n/primenenie-algoritma-k-means-dlya-klasterizatsii-posledovatelnostey-s-neizvestnym-kolichestvom-klasterov (accessed 22 October 2019).

13. Jackiv I., Gusarova L. 2003. Metody opredelenija kolichestva klasterov pri klassifikacii bez obuchenija. [Methods for determining the number of clusters in the classification without training.] Transport and Telecommunication 4 (1):23–28. Available at: http://www.tsi.lv/sites/default/files/editor/science/Research_journals/Tr_Tel/2003/V1/yatskiv_gousarova.pdf (accessed 25 October 2019).

14. MathWorks. 2019. Calinski-Harabasz criterion clustering evaluation object. Available at: https://www.mathworks.com/help/stats/clustering.evaluation.calinskiharabaszevaluation-class.html (accessed 25 October 2019).

15. Fränti P., Virmajoki O. 2006 Iterative shrinking method for clustering problems. Pattern Recognition. 39 (5): 761–765.

16. Fränti P., Virmajoki O., Hautamäki V. 2006. Fast agglomerative clustering using a k-nearest neighbor graph. IEEE Trans. on Pattern Analysis and Machine Intelligence. 28 (11): 1875–1881.

17. Rezaei M., Fränti P. 2016. Set-matching measures for external cluster validity. IEEE Trans. on Knowledge and Data Engineering. 28 (8): 2173–2186.

18. Kärkkäinen I., Fränti P. 2007. Gradual model generator for single-pass clustering. Pattern Recognition. 40 (3): 784–795.

19. Calinski T., Harabasz J. 1974. A dendrite method for cluster analysis. Communications in Statistics. 3 (1): 1–27.

20. Davies D.L., Bouldin D.W. 1979. A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Intelligence. PAMI-1(2): 224–227.

21. Tibshirani R., Walther G., Hastie T. 2001. Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B. 63 (2): 411–423.

22. Rouseeuw P.J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics. 20 (1): 53–65.

Ссылка для цитирования статьи
For citation

Фролов В.В., Слипченко С.Е., Приходько О.Ю. 2020. Метод расчета числа кластеров для алгоритма k-means. Экономика. Информатика. 47 (1): 213–225. DOI:

Frolov V.V., Slipchenko S.E., Prikhodko O.Yu. 2020. Clusters number calculating method for the k-means algorithm. Economics. Information technologies. 47 (1): 213–225 (in Russian). DOI: