

## Оптимизация кластерных разбиений с привлечением техники латентного анализа классов

С.В. Дронов

Алтайский государственный университет (Барнаул, Россия)

## Cluster Partitions Optimization Using the Latent Class Analysis Technique

S.V. Dronov

Altai State University (Barnaul, Russia)

Если каждый из изучаемых объектов отождествить с вектором, координаты которого являются значениями факторов, характеризующих этот объект, то построение кластерного разбиения превращается в формирование системы пучков подобных векторов. В работе изучены закономерности изменения суммарной тесноты этих пучков при перемещении объекта из одного кластера в другой. Исходя из полученных результатов, предлагается алгоритм, действуя согласно которому, можно понизить суммарную внутрикластерную изменчивость уже имевшегося начального кластерного разбиения. В основу предлагаемого алгоритма положена идея современной методики анализа латентных классов. Она состоит в том, что внутри каждого из построенных кластеров формирующие показатели объектов должны максимально возможным образом коррелировать между собой. Это требование заменяется на максимально возможную близость к среднему вектору соответствующего пучка. Степень такой близости и называется теснотой пучка векторов кластера. Для построенной с помощью предлагаемого алгоритма неулучшаемой кластеризации предложен новый метод квантификации ее кластеров. Рассмотрен практический пример применения алгоритма к медицинским данным. Обсуждаются причины зависимости результата от выбора начального кластерного разбиения.

**Ключевые слова:** кластерное разбиение, внутрикластерная вариация, латентные классы, пошаговая оптимизация разбиений.

DOI: 10.14258/izvasu(2023)1-14

### 1. Постановка задачи

Задача разбиения заданного конечного множества объектов на непересекающиеся части так, чтобы объекты каждой из частей были более похожи друг на друга, чем объекты, относящиеся к разным частям, известна под названием задачи кластеризации. Построенная система частей исходного множе-

The paper explores the formation of a cluster partition using a system of bundles of vectors. Each object is represented by a vector with coordinates corresponding to its characteristic factors. The study focuses on the principles of changes in the total density of these bundles when objects move from one cluster to another. Based on the findings, an algorithm is proposed to reduce the total intracluster variance of the initial cluster partition.

The proposed algorithm is based on a modern technique for analyzing latent classes. This technique emphasizes that the forming factors of objects within each cluster should be highly correlated. Since the number of forming factors is arbitrary, the algorithm replaces this requirement with the maximum possible proximity of object factor vectors to the average vector within each cluster. The degree of this closeness is referred to as the tightness of the cluster vector bundle.

The paper also introduces a new method for quantifying clusters in the improvable clustering constructed using the proposed algorithm. A practical example of the algorithm's application to process medical data is presented. The study discusses the reasons for the dependence of the algorithm's outcome on the choice of initial cluster partition.

**Key words:** cluster partition, intracluster variance, latent classes, step-by-step optimization of partitions.

ства называется его кластерным разбиением, а сами части кластерами.

В связи с востребованностью построения кластерных разбиений на сегодня разработано огромное количество разнообразных алгоритмов кластеризации, см., например, обзоры [1–2] и обширную библиографию в них. Часто разные ме-

тоды из этого арсенала, примененные к одному и тому же множеству объектов, приводят к различным результатам. Поэтому возникает вопрос: можно ли удовлетвориться одной из получающихся кластеризаций или стоит поискать ей какую-то альтернативу? Чаще всего решение этого вопроса в каждом конкретном исследовании основывается просто на интуиции или личном предпочтении автора этого исследования. Проблема улучшения имеющегося разбиения, не связанная с частными или интуитивными представлениями исследователей, видимо, впервые была поставлена в рамках аппарата латентного анализа классов [3–5], который также иногда называют латентным кластерным анализом.

В любых прикладных статистических задачах изучаемые объекты принято отождествлять с наборами числовых значений измеряемых у них показателей. Таким образом, любая кластеризация обязательно должна быть связана с этими показателями, которые поэтому будем называть формирующими. Согласно методике латентного анализа классов, кластеры в идеале должны быть организованы так, чтобы у объектов, лежащих в каждом из них, формирующие показатели имели в своем составе весомую общую часть. Она как бы формализует процесс формирования кластеров, который должен в максимальной степени вобрать в себя то общее из формирующих показателей, что обеспечивает близость объектов. Если формулировать точнее, то можно рассмотреть некую искусственную переменную, новый показатель, принимающий близкие значения на всех объектах одного и того же кластера, который аккумулирует эти существенные составляющие всех формирующих показателей. Такую переменную, следуя [6], будем называть кластерной переменной. Кластерное разбиение, следовательно, должно быть таким, чтобы условные распределения всех формирующих показателей относительно кластерной переменной оказались бы практически независимыми, — ничего общего в формирующих показателях после ее фиксации не остается.

Если подобную переменную удастся построить, то можно считать, что имеющееся кластерное разбиение в некотором смысле оптимально. Если же независимость показателей в кластерах соблюдается неидеально, можно попытаться переформатировать кластеры, перемещая некоторые из элементов между ними, добываясь лучшего качества результирующей кластерной переменной. Тот вариант кластерного разбиения, который невозможно улучшить за счет перемещения элементов между кластерами, назовем наилучшим.

Основная цель настоящей работы — разработка алгоритма переформатирования имеющегося кластерного разбиения в наилучший его вариант.

К сожалению, разрабатываемую ниже идею невозможно применить для построения наилучшего кластерного разбиения на множестве каких-то «сырых» данных — нужно иметь начальное, стартовое разбиение. К тому же необходимо запретить новому разбиению содержать кластеров больше, чем было в стартовом разбиении. Без такого требования, как это станет ясно ниже, в качестве наилучшего всегда будет получаться разбиение, в котором каждый элемент исходного множества представляет собой отдельный кластер. Понятно, что такое решение имеет нулевую практическую ценность. Требование же фиксированного количества кластеров естественно и является обычной практикой статистического исследования (см., например, [7–8]). Модифицировать требование неувеличения количества кластеров до требования его постоянства оказывается несложно, как будет продемонстрировано в разделе 3.

## 2. Построение тесных пучков из векторов-объектов

Предположим, что у нас имеется конечное множество  $U$  различных объектов, которое необходимо разбить на кластеры, и у каждого из этих объектов известны значения  $p$  формирующих показателей. Будем отождествлять каждый из объектов  $U$  с вектором, координаты которого представляют собой значения этих показателей. Условимся говорить, что все такие векторы-объекты, отнесенные в один и тот же кластер, составляют пучок. Все объекты пучка достаточно близки между собой, если каждый из них близок к центральному вектору пучка. Качеством имеющегося кластерного разбиения объявим сумму средних квадратов отклонений векторов от центрального вектора содержащего их пучка и будем решать задачу ее минимизации путем перемещения отдельных векторов из пучка в пучок. Как уже было сказано, дополнительно предположим, что общее число имевшихся пучков не может увеличиваться. Сформулируем задачу точнее.

Для пучка векторов  $\mathbf{A}$  через  $n_{\mathbf{A}}$  обозначим количество его элементов, через  $\bar{\mathbf{a}}(\mathbf{A})$  — его средний вектор, а

$$S^2(\mathbf{A}) = \frac{1}{n_{\mathbf{A}}} \sum_{i=1}^{n_{\mathbf{A}}} \left\| \vec{X}_i - \bar{\mathbf{a}}(\mathbf{A}) \right\|^2$$

будем называть рассеиванием пучка. Задача состоит в минимизации величины

$$S^2 = \sum_{\mathbf{A}} S^2(\mathbf{A}), \quad (1)$$

где сумма берется по всем пучкам  $\mathbf{A}$  путем перемещения отдельных векторов между пучками. Общее количество имевшихся пучков при этом

увеличиваться не должно, т.е. при допустимых изменениях пучков число слагаемых в сумме (1) может лишь уменьшаться.

Заметим в завершение постановки задачи, что, поскольку результат считается тем лучшим, чем меньше величина  $S^2$ , а она равна 0 тогда и только тогда, когда каждый из пучков одноэлементен, то условие ограниченности числа пучков отсекает тривиальное решение.

Перейдем к детальному изучению процесса перереформатирования пучков. Для  $n$  векторов в  $R^p$ , составляющих пучок, определим

$$\bar{a} = \frac{1}{n} \sum_{i=1}^n \vec{X}_i, \quad Q_n = \sum_{i=1}^n \left\| \vec{X}_i - \bar{a} \right\|^2.$$

Добавим к пучку новый вектор  $\vec{X}$ . Ясно, что отклонения добавляемого вектора от нового  $\bar{a}$  и от старого среднего  $\bar{a}$  будут связаны условием

$$\bar{a} = \frac{n\bar{a} + \vec{X}}{n+1} \implies \vec{X} - \bar{a} = \frac{n}{n+1} (\vec{X} - \bar{a}). \quad (2)$$

Кроме этого,

$$\begin{aligned} \|\bar{a}\|^2 - \|\bar{a}\|^2 &= \left\langle \frac{\bar{a} - \vec{X}}{n+1}, \bar{a} + \frac{n\bar{a} + \vec{X}}{n+1} \right\rangle = \\ &= \left\langle \frac{\bar{a} - \vec{X}}{n+1}, \frac{(2n+1)\bar{a} + \vec{X}}{n+1} \right\rangle = \frac{\langle \bar{a} - \vec{X}, (2n+1)\bar{a} + \vec{X} \rangle}{(n+1)^2}. \end{aligned} \quad (3)$$

Здесь  $\langle \cdot, \cdot \rangle$  — скалярное произведение векторов. Далее

$$\begin{aligned} Q_{n+1} - Q_n &= \left\| \vec{X} - \bar{a} \right\|^2 + \\ &+ \sum_{i=1}^n \left( \left\| \vec{X}_i - \bar{a} \right\|^2 - \left\| \vec{X}_i - \bar{a} \right\|^2 \right) = I_1 + I_2. \end{aligned} \quad (4)$$

Преобразуем второе слагаемое в (4).

$$\begin{aligned} I_2 &= \sum_{i=1}^n \langle \bar{a} - \bar{a}, 2\vec{X}_i - (\bar{a} + \bar{a}) \rangle = \\ &= \left\langle \frac{\bar{a} - \vec{X}}{n+1}, 2n\bar{a} \right\rangle - n (\|\bar{a}\|^2 - \|\bar{a}\|^2). \end{aligned}$$

Теперь, привлекая (3), выводим

$$I_2 = \frac{n}{(n+1)^2} \|\bar{a} - \vec{X}\|^2.$$

Формула (2) позволяет преобразовать  $I_1$ , и из равенства (4) вытекает

$$Q_{n+1} - Q_n = \frac{n}{n+1} \|\vec{X} - \bar{a}\|^2. \quad (5)$$

Для более удобного дальнейшего использования сменим обозначения. Средний вектор пучка из произвольного числа  $k$  векторов будем обозначать  $\bar{a}_k$ , а рассеивание этого пучка пусть равно  $S_k^2$ . Тогда, используя (5), приходим к

$$S_{n+1}^2 = \frac{Q_{n+1}}{n+1} = \frac{nS_n^2}{n+1} + \frac{n}{(n+1)^2} \|\vec{X} - \bar{a}_n\|^2. \quad (6)$$

Для описания ситуации, когда из пучка, содержащего  $n$  векторов, напротив, удаляется вектор  $\vec{X}$ , достаточно в (6) заменить  $n$  на  $n-1$  и выразить из получившегося равенства величину изменчивости нового пучка. Применяя (2), получаем

$$S_{n-1}^2 = \frac{nS_n^2}{n-1} - \frac{n-1}{n^2} \|\vec{X} - \bar{a}_{n-1}\|^2. \quad (7)$$

Итак, рассеивание пучка при добавлении или выносе  $\vec{X}$  будет, соответственно, изменяться на

$$\begin{aligned} \Delta^+(\mathbf{A}) &= \frac{n_{\mathbf{A}}}{(n_{\mathbf{A}}+1)^2} \left\| \vec{X} - \bar{a}(\mathbf{A}) \right\|^2 - \frac{S^2(\mathbf{A})}{n_{\mathbf{A}}+1}, \\ \Delta^-(\mathbf{A}) &= \frac{S^2(\mathbf{A})}{n_{\mathbf{A}}-1} - \frac{n_{\mathbf{A}}}{(n_{\mathbf{A}}-1)^2} \left\| \vec{X} - \bar{a}(\mathbf{A}) \right\|^2. \end{aligned}$$

Последняя из этих двух формул работает только при условии  $n_{\mathbf{A}} \geq 2$ . Если пучок состоял из единственного вектора, то его рассеивание было равным 0, и это же значение можно сохранить и для пустого пучка, поэтому в этом случае условимся полагать  $\Delta^-(\mathbf{A}) = 0$ . Собирая (6) и (7), приходим к справедливости теоремы.

**Теорема.** Если вектор  $\vec{X}$  из пучка  $\mathbf{A}$  перенести в пучок  $\mathbf{B}$ , то суммарное рассеивание пучков изменится на величину

$$\Delta(\mathbf{A}; \mathbf{B}) = \Delta^-(\mathbf{A}) + \Delta^+(\mathbf{B}). \quad (8)$$

Отметим, что величина рассеивания уменьшается при добавлении вектора в пучок тем сильнее, чем ближе оказывается он к среднему вектору имевшегося пучка. При исключении вектора рассеивание пучка уменьшается сильнее при удалении вектора, наиболее удаленного от имевшегося среднего.

**Следствие.** При добавлении  $\vec{X}$  в пучок  $\mathbf{A}$  его рассеивание уменьшится тогда и только тогда, когда

$$\|\vec{X} - \bar{a}_{n_{\mathbf{A}}}\| < S(\mathbf{A}) \cdot \sqrt{\frac{n_{\mathbf{A}} + 1}{n_{\mathbf{A}}}},$$

а при исключении  $\vec{X}$  из пучка в том и только том случае, если

$$\|\vec{X} - \bar{a}_{n_{\mathbf{A}}}\| > S(\mathbf{A}) \cdot \sqrt{\frac{n_{\mathbf{A}} - 1}{n_{\mathbf{A}}}}.$$

Утверждение теоремы позволяет указать путь решения задачи уменьшения суммарного рассеивания некоторого набора пучков. Следует, коротко говоря, переносить элементы из пучка в пучок до тех пор, пока среди чисел  $\Delta(\mathbf{A}; \mathbf{B})$  остаются отрицательные, причем начинать перенос нужно с тех элементов, для которых модуль соответствующего отрицательного числа окажется наибольшим.

При выполнении описанных перенесений общее число пучков может уменьшиться, если удаляется вектор из одноэлементного пучка. Если, как

в кластерном алгоритме  $k$ -средних [7–8], дополнительно выставить условие, чтобы оно и не уменьшалось, придется слегка изменить алгоритм, запретив перенесения таких векторов. Поскольку понятно, что уменьшить нулевое рассеивание одноэлементного пучка не удастся, то в этом варианте алгоритма одноэлементные пучки автоматически исключаются из рассмотрения в качестве потенциальных «доноров».

**3. Алгоритм переформатирования пучков и пример его работы**

На входе в алгоритм имеется множество векторов, разбитое на пучки (стартовое разбиение). Организуем поэлементное перемещение векторов из пучка в пучок так, чтоб минимизировать суммарное рассеивание пучков (1).

Шаг 1. Для каждого из пучков  $\mathbf{A}$  определим число его элементов  $n_{\mathbf{A}}$ , вычислим центральный вектор  $\vec{a}(\mathbf{A})$  и рассеивание  $S^2(\mathbf{A})$ . Эти вычисления можно делать только для тех пучков, которые изменились на предыдущей итерации.

Шаг 2. Для каждого вектора каждого из пучков вычислим его отклонения от всех центральных векторов пучков, а затем для пучка  $\mathbf{A}$ , в котором содержался этот вектор, и каждого из пучков  $\mathbf{B} \neq \mathbf{A}$  найдем потенциальное изменение  $\Delta(\mathbf{A}; \mathbf{B})$  величины  $S^2$  по формулам доказанной выше теоремы. При этом полагаем  $\Delta(\mathbf{A}; \mathbf{A}) = 0$ .

Шаг 3. Есть ли среди чисел  $\Delta(\mathbf{A}; \mathbf{B})$  отрицательные? Если нет, то неуплучшаемое разбиение построено — выход из алгоритма. Если нет — к шагу 4.

Шаг 4. Находим такой элемент  $\vec{X}$  и такой пучок  $\mathbf{B}$ , что  $\vec{X} \in \mathbf{A}$ ,  $\Delta(\mathbf{A}; \mathbf{B})$  отрицательно и максимально по модулю среди всех отрицательных значений. Переносим вектор  $\vec{X}$  из  $\mathbf{A}$  в  $\mathbf{B}$ . Начиная новую итерацию, возвращаемся к шагу 1.

Для превращения алгоритма в такой его вариант, при котором сохраняется первоначальное число пучков, не уменьшаясь, на шаге 3 следует исключить из изучаемых те  $\Delta(\mathbf{A}; \mathbf{B})$ , где пучок  $\mathbf{A}$  состоит из единственного вектора.

В качестве иллюстрации рассмотрим работу алгоритма на примере реальных данных анализа лимфоцитарного профиля периферической крови 15 пациентов по 4 показателям, выполненного в Алтайском краевом диагностическом центре. За стартовое разбиение был взят результат применения алгоритма 5-средних, полученный с помощью статистического компьютерного пакета IBM SPSS 23. Пять пучков-кластеров этого разбиения имеют следующие составы:  $\mathbf{A} = \{1; 5; 6; 15\}$ ,  $\mathbf{B} = \{3; 12\}$ ,  $\mathbf{C} = \{2; 4; 7; 8; 10; 13; 14\}$ ,  $\mathbf{D} = \{9\}$ ,  $\mathbf{E} = \{11\}$ . Характеристики всех пучков приведены в таблице 1.

Изменения кластеров в процессе работы алгоритма описываются таблицей 2. Во втором столбце указан номер перемещаемого объекта-вектора,

Таблица 1

Стартовое разбиение 1

Пучок	$n_{\mathbf{A}}$	$\vec{a}(\mathbf{A})$				$S^2(\mathbf{A})$
<b>A</b>	4	15,66	8,09	30,95	14,26	1,50
<b>B</b>	2	15,89	8,39	29,27	13,09	0,83
<b>C</b>	7	15,62	8,04	31,58	12,92	1,22
<b>D</b>	1	16,47	8,53	34,41	15,59	0,00
<b>E</b>	1	16,96	8,77	33,04	13,45	0,00
		$S^2$				3,55

в третьем — маршрут его перемещения, в четвертом — достигнутая после этого перемещения суммарная величина рассеивания.

Таблица 2

Изменения пучков

Итерация	Объект	Перемещение	$S^2$
1	1	<b>A</b> $\rightsquigarrow$ <b>B</b>	3,137
2	5	<b>A</b> $\rightsquigarrow$ <b>C</b>	3,066
3	6	<b>A</b> $\rightsquigarrow$ <b>C</b>	2,702
4	12	<b>B</b> $\rightsquigarrow$ <b>C</b>	2,578
5	3	<b>B</b> $\rightsquigarrow$ <b>C</b>	2,070

После пятой итерации отрицательных чисел среди  $\Delta(\mathbf{A}; \mathbf{B})$  не осталось. Полученное таким образом неуплучшаемое разбиение состоит из четырех одноэлементных кластеров  $\{1\}$ ,  $\{9\}$ ,  $\{11\}$ ,  $\{15\}$ , а остальные 11 элементов образуют пятый кластер. Достигнутая величина рассеивания равна 2,07.

Далее на том же множестве объектов было построено разбиение на 5 пучков-кластеров с помощью агломеративного иерархического кластерного алгоритма. Получены кластеры  $\mathbf{A} = \{1; 3; 12\}$ ,  $\mathbf{B} = \{2; 4; 14\}$ ,  $\mathbf{D} = \{9\}$ ,  $\mathbf{E} = \{11\}$ , а пучок  $\mathbf{C}$  содержит остальные 7 объектов. В таблице 3 приведены количества элементов, средние векторы и рассеивания каждого из пучков.

Таблица 3

Стартовое разбиение 2

Пучок	$n_{\mathbf{A}}$	$\vec{a}(\mathbf{A})$				$S^2(\mathbf{A})$
<b>A</b>	3	16,12	8,30	29,35	13,65	1,30
<b>B</b>	3	15,38	7,99	31,26	12,03	0,41
<b>C</b>	7	15,61	8,08	31,65	13,81	0,73
<b>D</b>	1	16,47	8,53	34,41	15,59	0,00
<b>E</b>	1	16,96	8,77	33,04	13,45	0,00
		$S^2$				2,44

Это разбиение оказалось неуплучшаемым — среди чисел  $\Delta$  нет отрицательных.

В традиционном кластерном анализе характеристика  $S^2$  носит название суммарной внутрикластерной изменчивости, и оптимизируя разбиение множества объектов на кластеры, наряду с уменьшением этой характеристики стараются также максимизировать межкластерную изменчивость. Во введенных выше обозначениях эта величина равна

$$M = \frac{1}{k} \sum_{\mathbf{A}} \|\vec{a}(\mathbf{A}) - \bar{a}\|^2, \quad \bar{a} = \frac{1}{k} \sum_{\mathbf{A}} \vec{a}(\mathbf{A}),$$

если всего имелось  $k$  пучков. Поскольку уменьшение величины  $S^2$  чаще всего сопровождается уменьшением  $M$ , то обычным компромиссом является минимизация  $Z = S^2/M$  (см, к примеру, [9]). В рассмотренных выше примерах численные значения: для улучшенного разбиения  $M = 3,91; Z = 0,5288$ , для разбиения иерархическим методом  $M = 4,61; Z = 0,5292$ . Таким образом, хотя межкластерная изменчивость в модифицированном разбиении оказалась немного хуже, но величина  $Z$  все же улучшилась, хотя и незначительно.

#### 4. О квантификации кластерной переменной

После построения пучков, имеющих минимальное суммарное рассеивание, можно предложить подход к присвоению числовых меток пучкам-кластерам. Заметим, что после работы алгоритма переформатирования пучков все векторы каждого пучка  $\mathbf{A}$  оказались достаточно близкими к его среднему вектору  $\vec{a}(\mathbf{A})$ . Поэтому этот средний вектор может выполнять роль многомерной метки пучка. Если же размерность полученных меток покажется слишком высокой или потребуются построение именно числовых меток, то к набору многомерных меток можно применить любой из методов сокращения размерностей (см., например, [10–11]). Обычно метки центрируют и нормируют, вычитая из них среднее значение и деля на сумму их квадратов.

В частности, если ставить целью как можно больший разброс получающихся меток, то для первого примера предыдущего раздела имеем в качестве меток нормированные значения четвертых координат векторов средних. Метки пуч-

ков можно принять за значения кластерной переменной на каждом из объектов в пределах соответствующего кластера. Это 0,04; -0,66; -0,02; 0,74 и -0,10 соответственно. В частности, это приводит к заключению о том, что естественный порядок кластеров таков: **В; Е; С; А; D**.

#### 5. Обсуждение результатов и краткие выводы

Предложенный несложный алгоритм позволяет уменьшать суммарное рассеивание имеющихся не старте пучков-кластеров произвольных объектов путем их перенесения из одного пучка в другой.

То, что в результате использования разных стартовых разбиений примера алгоритм привел к разным результатам, хотя и примерно одинакового качества, видимо, показывает, что абсолютный минимум суммарного рассеивания не всегда достигается путем последовательного перемещения отдельных элементов из кластера в кластер. Перебор же всех возможных кластерных разбиений данного множества с целью найти этот минимум вряд ли может быть рекомендован к использованию. Поэтому алгоритм, на взгляд автора, служит разумной альтернативой этому перебору.

Отметим также, что использованный подход повышения тесноты пучка фактически подменяет сильную коррелированность показателей внутри пучка этого пучка, которая использовалась в работе [12]. Но там изучался случай наличия лишь двух формирующих показателей. Использовать напрямую идею сильной коррелированности, как предлагает латентный анализ классов, в случае, когда показателей больше двух, весьма затруднительно, поскольку коэффициент корреляции является парной характеристикой. Правда, если иметь в виду геометрическую интерпретацию коэффициента корреляции через косинусы углов между векторами соответствующих показателей, можно при переформатировании пучков поставить цель помещения векторов внутри многомерного конуса с возможно меньшим телесным углом при его вершине. Построению и изучению подобного алгоритма будет посвящена одна из следующих работ автора.

### Библиографический список

1. Xu R., Wunsch D. II. Survey of Clustering Algorithms // IEEE Transactions on Neural Networks, 2005. Vol. 16. № 3. DOI: 10.1109/TNN.2005.845141.
2. Xu D., Tian Y. A. Comprehensive Survey of Clustering Algorithms // Ann. Data. Sci. 2015. Vol. 2. DOI: 10.1007/s40745-015-0040-1.
3. Rindskopf D. Latent Class Analysis. In: The SAGE Handbook of Quantitative Methods in Psychology, N.Y.: Sage, 2009.
4. Hagenaaars J.A., McCutcheon A.L. Applied Latent Class Analyses Models // Canadian Journal of Sociology, 2003. Vol. 28 (3). DOI: 10.2307/3341848.

5. Obersky D.L., Hagenaars J.A., Saris W. The Latent Class Multitrait-Multimethod Model 5 // *Psychological Methods*, 2014. Vol. 20 (4). DOI: 0.1037/a0039783.
6. Dronov S.V., Sazonova A.S. Two approaches to cluster variable quantification // *Model Assisted Statistics and Applications*, 2015. Vol. 10. DOI: 10.3233/mas-140314.
7. Coates A., Ng A.Y. Learning feature representations with k-means // In: Montavon, G.; Orr, G. B.; Muller, K.-R. (eds.). *Neural Networks: Tricks of the Trade (Lecture Notes in Computer Science, volume 7700)*. Springer, 2012.
8. Vinnikov A., Shalev-Shwartz S. K-means Recovers ICA Filters when Independent Components are Sparse // *Proceedings of the International Conference on Machine Learning*, 2014. Vol. 32 (II).
9. Halkidi M., Batistakis Y. Vazirgiannis M. On clustering validation techniques // *Journal of intelligent information systems*, 2001. Vol. 17 (2-3). DOI: 10.1023/A:1012801612483.
10. Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. *Классификация и снижение размерности*. М., 1989.
11. Zahorian S.A., Hu H. Nonlinear Dimensionality Reduction Methods for Use with Automatic Speech Recognition. In: *Speech Technologies*. Intech Open, 2011. DOI: 10.5772/16863.
12. Дронов С.В., Шеларь А.Ю. Новый алгоритм выявления и квантификации латентных классов // *Известия Алт. гос. ун-та*. 2020. Вып. 4 (114). DOI: 10.14258/izvasu(2020)4-12 .