ТЕХНИЧЕСКИЕ НАУКИ

TECHNICAL SCIENCE

УДК 519.22
ББК 22.172.6
Л 86

Луценко Е.В.
Доктор экономических наук, кандидат технических наук, профессор кафедры автоматизированных систем обработки информации и управления физического факультета Адыгейского государственного университета, тел. (8772) 59-39-11
Коржаков В.Е.
Кандидат технических наук, доцент, зав. кафедрой автоматизированных систем обработки информации и управления физического факультета Адыгейского государственного университета, тел. (8772) 59-39-11, e-mail: korvie@yandex.ru

Некоторые проблемы классического кластерного анализа

(Рецензирована)

Аннотация

В статье рассматриваются основные проблемы кластерного анализа, связанные с тем, что: а) параметры обобщенного образа кластера вычисляются как средние от исходных объектов (классов) или центры тяжести; б) в качестве критерия сходства используется евклидово расстояние или его варианты, некорректные в неортонормированных пространствах, которые обычно и встречаются на практике; в) кластерный анализ проводится на основе исходных переменных или матрицы сопряженности, зависящих от единиц измерения по осям, для формализации которых используются шкалы различных типов. По этим причинам результаты кластеризации часто не понятны специалистам и не поддаются содержательной интерпретации, не согласующиеся с оценками экспертов, их опытом и интуитивными ожиданиями. Предлагается идея решения перечисленных проблем.

Ключевые слова: *автоматизированный системно-когнитивный анализ, интеллектуальная система «эйдос», когнитивное пространство, агломеративная кластеризация.*

Lutsenko E.V.
Doctor of Economics, Candidate of Technical Sciences, Professor of Department of Automated Systems of Processing Information and Control at Physical Faculty of Adyghe State University, ph. (8772) 59-39-11
Korzhakov V.E.
Candidate of Technical Sciences, Associate Professor, Head of Department of Automated Systems of Processing Information and Control at Physical Faculty of Adyghe State University, ph. (8772) 59-39-11, e-mail: korvie@yandex.ru

Some problematic aspects of the classical cluster analysis

Abstract

The paper discusses the basic problems related to the cluster analysis, namely: (a) parameters of the cluster generalized image are calculated as the average of initial objects (classes) or the gravity centers; (b) the Euclidean distance or its variants, incorrect in neortho normalized spaces, which usually occurs in practice, is used as a criterion for similarity; (c) the cluster analysis is carried out on the basis of initial variables or a matrix of conjugacy, depending on units of measure on axes for formalization of which scales of different types are used. For these reasons results of clusterization are often unclear to experts and cannot be substantially interpreted, they do not agree with experts' estimations, their experience and intuitive expectations. The idea of the solution of the listed problems is suggested.

Key words: *the automated system-cognitive analysis, an intellectual system «Eidos», cognitive space, agglomerative clusterization.*

«Мышление – это обобщение, абстрагирование, сравнение, и классификация»
Патанджали[1], II в. до н.э.

«Истинное знание – это знание причин»
Френсис Бэкон (1561–1626 гг.)

Кластерный анализ[2] (англ. *Data clustering*) – это задача разбиения заданной выборки *объектов* (ситуаций) на подмножества, называемые кластерами, так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались. Кластерный анализ очень широко применяется как в науке, так и в различных направлениях практической деятельности. Значение кластерного анализа невозможно переоценить, оно широко известно[3] и нет необходимости его специально обосновывать.

Существует *большое количество* различных методов кластерного анализа, хорошо описанных в многочисленной специальной литературе [1] и прекрасных обзорных статьях [2-5]. Поэтому в данной статье мы не ставим перед собой задачу дать еще одно подобное описание, а обратим основное внимание на ***проблемы***, существующие в кластерном анализе и вариант их решения, предлагаемый в автоматизированном системно-когнитивном анализе (АСК-анализ). Эти проблемы, *в основном*, хорошо известны специалистам, и поэтому наш краткий обзор будет практически полностью основан на уже упомянутых работах [2-5]. Необходимо специально отметить, что специалисты небезуспешно работают над решением этих проблем, предлагая все новые и новые варианты, которые и являются различными вариантами кластерного анализа. Мы в данной статье также предложим еще один ранее не описанный в специальной литературе (т.е. новый, авторский) теоретически обоснованный и программно-реализованный вариант решения некоторых из этих проблем, а также проиллюстрируем его на простом численном примере.

Почему же разработано так много различных методов кластерного анализа, почему это было необходимо? Кажутся почти очевидными мысли о том, что различные методы кластерного анализа дают результаты *различного качества*, т.е. одни методы *в определенном смысле* «лучше», а другие «хуже», и это действительно так [6], и, следовательно, по-видимому, *должен* существовать только один-единственный метод кластеризации, *всегда* (т.е. на любых данных) дающий «правильные» результаты, тогда как все остальные методы являются «неправильными». Однако если задать аналогичный вопрос по поводу, например, автомобиля или одежды, то становится ясным, что нет просто наилучшего автомобиля, а есть лучшие по определенным критериям-требованиям или лучшие для определенных *целей*. При этом сами критерии также должны быть обоснованы и не просто могут быть различными, но и должны быть различными при различных целях, чтобы отражать цель и соответствовать ей. Так авто-

[1] http://ru.wikipedia.org/wiki/Патанджали
[2] http://ru.wikipedia.org/wiki/Кластерный%20анализ
[3] http://yandex.ru/yandsearch?text=кластерный%20анализ

мобиль, лучший для семейного отдыха, не являются лучшим для гонок Формулы-1 или для представительских целей. Аналогично можно обоснованно утверждать, что одни методы кластерного анализа являются более подходящими для кластеризации данных определенной структуры, а другие – другой, т.е. не существует одного наилучшего во всех случаях *универсального метода кластеризации*, но существуют методы более универсальные и методы менее универсальные. Но все же многообразие разработанных методов кластерного анализа на наш взгляд указывает не только на это, но и на то, что *их можно рассматривать как различные более или менее успешные варианты решения или попытки решения тех или иных проблем, существующих в области кластерного анализа*.

Для структурирования дальнейшего изложения сформулируем требования к исходным данным в кластерном анализе и фундаментальные вопросы, которые решают разработчики различных методов кластерного анализа.

Считается[4], что кластерный анализ предъявляет следующие *требования к исходным данным*:

1. Показатели не должны коррелировать между собой.
2. Показатели должны быть безразмерными.
3. Распределение показателей должно быть близко к нормальному.
4. Показатели должны отвечать требованию «устойчивости», под которой понимается отсутствие влияния на их значения случайных факторов.
5. Выборка должна быть однородна, не содержать «выбросов».

Даже поверхностный анализ сформулированных требований к исходным данным сразу позволяет утверждать, что *на практике они в полной мере никогда не выполняются*, а приведение исходных данных к виду, удовлетворяющему этим требованиям, или очень сложно, т.е. представляет собой ***проблему***, и не одну, или *даже теоретически невозможно* в полной мере. В любом случае пытаться это делать можно *различными способами*, хотя *чаще всего на практике этого не делается вообще* или потому, что необходимость этого плохо осознается исследователем, или чаще потому, что в его распоряжении нет соответствующих инструментов, реализующих необходимые методы[5]. Конечно, в последнем случае не приходится удивляться тому, что результаты кластерного анализа получаются мягко сказать «несколько странными», а если они соответствуют здравому смыслу и точке зрения экспертов, то можно сказать, что это получилось случайно или потому, что «просто повезло».

Остановимся подробнее на анализе перечисленных требований к исходным данным, а также проблем, возникающих при попытке их выполнения и решения.

Первое требование связано с использованием в большинстве методов кластеризации *евклидова расстояния* или различных его вариантов в качестве меры близости объектов и кластеров. Другими словами это требование означает, что описательные шкалы, рассматриваемые как оси семантического пространства, должны быть *ортонормированны*, т.к. *в противном случае применение евклидова расстояния и большинства других метрик* (таблица 1) *(кроме расстояния Махалонобиса) теоретически необоснованно и некорректно*.

[4] http://ru.wikipedia.org/wiki/Кластерный%20анализ

[5] Справедливости ради отметим, что подобных инструментов вообще *мало* и они практически недоступны исследователям.

Таблица 1

Основные типы метрик при кластер-анализе[6]

№	Наименование метрики	Тип признаков	Формула для оценки меры близости (метрики)
1	Эвклидово расстояние	Количественные	$d_{ik} = \left(\sum_{j=1}^{N} (x_{ij} - x_{kj})^2 \right)^{\frac{1}{2}}$
2	Мера сходства Хэмминга	Номинальные (качественные)	$\mu_{ij}^{H} = \frac{n_{ik}}{N}$, где n_{ik} – число совпадающих признаков у образцов X_i и X_k
3	Мера сходства Роджерса-Танимото	Номинальные шкалы	$\mu_{ij}^{R-T} = n''_{ik}(n'_i + n'_k - n''_{ik})$, где n''_{ik} – число совпадающих единичных признаков у образцов X_i и X_k; n'_i, n'_k – общее число единичных признаков у образцов X_i и X_k соответственно
4	Манхэттенская метрика	Количественные	$d_{ik}^{(1)} = \sum_{j=1}^{N} \lvert x_{ij} - x_{kj} \rvert$
5	Расстояние Махаланобиса	Количественные	$d_{ik}^{M} = (x_{ij} - x_{kj})^T W^{-1} (x_{ij} - x_{kj})$, где **W** – ковариационная матрица выборки $\mathbb{X} = (X_1, X_2, \ldots, X_n)$
6	Расстояние Журавлева	Смешанные	$d_{ik} = \sum_{j=1}^{N} I_{ik}^{j}$, где $I_{ik}^{j} = \begin{cases} 1, \text{если } \lvert x_{ij} - x_{kj} \rvert < \varepsilon \\ 0, \text{иначе} \end{cases}$

Существуют и другие метрики, в частности: квадрат евклидова расстояния, расстояние городских кварталов (манхэттенское расстояние), расстояние Чебышева, степенное расстояние, процент несогласия, метрики Рао, Хемминга, Роджерса-Танимото, Жаккара, Гауэра, Воронина, Миркина, Брея-Кертиса, Канберровская и многие другие [2, 4]. Когда *корреляции между переменными равны нулю*, расстояние Махаланобиса эквивалентно квадрату евклидового расстояния [2]. Это означает, что метрику Махала-

[6] Источник: проф. Зайченко Ю.П. http://www.masters.donntu.edu.ua/2005/kita/kapustina/library/cluster.htm

нобиса можно считать обобщением евклидовой метрики для неортонормированных пространств[7].

Но на практике это требование *никогда* в полной мере не выполняется, а для его выполнения необходимо выполнить операцию ортонормирования семантического пространства, при которой из модели тем или иным методом[8] (реализованным в программной системе, в которой проводится кластерный анализ) *исключаются* все шкалы, коррелирующие между собой.

Таким образом, первое требование к исходным данным порождает две проблемы:

Проблема 1.1 выбора метрики, корректной для неортонормированных пространств.

Проблема 1.2 ортонормирования пространства.

Второе требование (безразмерности показателей) вытекает из того, что *выбор единиц измерения по осям существенно влияет на результаты кластеризации.* выбор единиц измерения, по сути, произволен (определяется исследователем), вследствие чего и результаты кластеризации, вместо того чтобы объективно отражать структуру данных и описываемой ими объективной реальности, также становятся произвольными и зависящими не только от самой исследуемой реальности, но и от произвола исследователя (причем неизвестно от чего больше: от реальности или исследователя). По сути, *автоматизированная система кластеризации превращается в этих условиях из инструмента исследования структуры объективной реальности в автоматизированный инструмент рисования таких дендрограмм, какие больше нравятся пользователю*. Непонятно также, какой содержательный смысл могут иметь, например корни квадратные из сумм квадратов разностей координат объектов, классов или кластеров, *измеряемых в различных единицах измерения. Разве корректно складывать величины даже одного рода, измеряемые в различных единицах измерения, а тем более разного рода?* Даже если сложить величины одного рода, но измеренные в разных единицах измерения, например *расстояния* от школы до подъезда дома 1,2 (километра), и от подъезда дома до квартиры 25 (метров), то получится 26,2 *непонятно чего*. Если же сложить разнородные по смыслу величины, т.е. *величины различной природы*, такие, например, как квадрат разности веса студентов с квадратом разности их роста, возраста, успеваемости и т.д., а потом еще извлечь из этой суммы квадратный корень, то получится просто *бессмысленная величина*, которая в традиционном кластерном анализе почему-то называется «Евклидово расстояние». В школе на уроке физики в 8-м классе за подобные действия сразу бы поставили «неуд»[9]. Однако, как это ни удивительно, то, что «не прошло бы» на уроке физики в средней школе является вполне устоявшейся практикой в статистике и ее научных применениях.

---

[7] http://matlab.exponenta.ru/fuzzylogic/book1/12_1_3.php http://d3lpirt.narod.ru/dm/dm.htm

[8] Например, для ортонормирования семантического пространства может быть применен метод главных компонент: http://ru.wikipedia.org/wiki/Метод%20главных%20компонент

[9] Конечно, есть случаи, когда производят определенные математические операции над величинами различной природы, измеряемыми в различных единицах измерения, и это вполне корректно, правда это не операция сложения. Например, в физике так производятся вычисления *по формулам*. Но эти формулы теоретически обоснованы в соответствующих физических теориях. Если математические операции производятся так, что это не соответствует обоснованным формулам, то в результате получаются бессмысленные величины неизвестных науке размерностей. В этом случае говорят о проверке размерностей и нарушении размерностей. Такое впечатление, что в статистике подобные нарушения размерностей просто стали нормой.

В подтверждение тому, что подобная практика действительно существует, авторы не могут удержаться от искушения и не привести пространную цитату из работы [4]: «Заметим, что *евклидово расстояние* (и его квадрат) вычисляется по исходным, а не по стандартизованным данным. *Это обычный способ его вычисления*, который имеет определенные преимущества (например, расстояние между двумя объектами не изменяется при введении в анализ нового объекта, который может оказаться выбросом). Тем не менее, *на расстояния могут сильно влиять различия между осями, по координатам которых вычисляются эти расстояния. К примеру, если одна из осей измерена в сантиметрах, а вы потом переведете ее в миллиметры (умножая значения на 10), то окончательное евклидово расстояние (или квадрат евклидова расстояния), вычисляемое по координатам, сильно изменится, и, как следствие, результаты кластерного анализа могут сильно отличаться от предыдущих.*» (выделено нами, авт.)[10]. В работе [4] просто констатируется факт этой ситуации, но ему не дается никакой *оценки*. Наша же оценка этой практике по перечисленным выше причинам *отрицательная*. Приведем еще цитату из той же работы [4]: «***Степенное расстояние.*** Иногда желают (!!!?)[11] прогрессивно увеличить или уменьшить вес, относящийся к размерности, для которой соответствующие объекты сильно отличаются. Это может быть достигнуто с использованием *степенного расстояния*.

Степенное расстояние вычисляется по формуле:

$$\text{расстояние } (x, y) = (\sum_i |x_i - y_i|^p)^{1/r},$$

где r и p – параметры, определяемые пользователем. Несколько примеров вычислений могут показать, как «работает» эта мера. Параметр p ответственен за постепенное взвешивание разностей по отдельным координатам, параметр r ответственен за прогрессивное взвешивание больших расстояний между объектами. Если оба параметра – r и p, равны двум, то это расстояние совпадает с «расстоянием Евклида». Мы считаем, что еще какие-то комментарии здесь излишни.

Таким образом, второе требование к исходным данным порождает следующую проблему 2.1:

Проблема 2.1 сопоставимой обработки описаний объектов, описанных признаками различной природы, измеряемыми в различных единицах измерения (проблема размерностей).

Отметим также, что объекты чаще всего описаны не только признаками, измеряемыми в различных единицах измерения, но как количественными, так и качественными признаками, которые соответственно являются градациями как числовых шкал, так и номинальных (текстовых) шкал. Существует метрика для номинальных шкал: это «Процент несогласия» [4], однако для количественных шкал применяются другие метрики. *Каким образом и с помощью какой комбинации классических метрик вычислять расстояния между объектами, описанными как количественными, так и качественными признаками, а также между кластерами, в которые они входят, вообще не понятно.* Это порождает проблему 2.2:

Проблема 2.2 формализации описаний объектов, имеющих как количественные, так и качественные признаки.

---

[10] Пространные цитаты здесь и далее для удобства читателей приведены мелким шрифтом.

[11] Пометка (!!!?) наша, авт.

Третье требование (нормальности распределения показателей) вытекает из того, что статистическое обоснование корректности вышеперечисленных метрик существенным образом основано на этом предположении, т.е. эти метрики являются параметрическими. На практике это означает, что перед применением кластерного анализа с этими метриками необходимо доказать гипотезу о нормальности исходных данных либо применить процедуру их нормализации. И первое, и второе, весьма ***проблематично*** и на практике не делается, более того, даже вопрос об этом чаще всего не ставится. Процедура нормализации (или взвешивания, ремонта) исходных данных обычно предполагает удаление из исходной выборки тех данных, которые нарушают их нормальность. Ясно, что это непредсказуемым образом может повлиять на результаты кластеризации, которые, скорее всего, существенно изменяться и их уже нельзя будет признать результатами кластеризации исходных данных. Отметим, что на практике исходные данные, не подчиняющиеся нормальному распределению, встречаются достаточно часто, что и делает актуальными методы непараметрической статистики.

Таким образом, 3-е требование к исходным данным порождает проблемы 3.1, 3.2 и 3.3:

Проблема 3.1 доказательства гипотезы о нормальности исходных данных.

Проблема 3.2 нормализации исходных данных.

Проблема 3.3 применения непараметрических методов кластеризации, корректно работающих с ненормализованными данными.

Что можно сказать о четвертом и пятом требованиях?[12] Эти требования взаимосвязаны, т.к. случайные факторы и порождают «выбросы». На практике, строго говоря, эти требования никогда не выполняются и вообще звучат *несколько наивно*, если учесть, что как случайные часто рассматриваются неизвестные факторы, а их влияние даже теоретически, т.е. в принципе, исключить невозможно. С другой стороны эти требования «удобны» тем, что неудачные, неадекватные или не интерпретируемые результаты кластеризации, полученные тем или иным методом кластерного анализа, всегда можно «списать» на эти неизвестные «случайные» факторы или скрытые параметры и порожденные ими выбросы. А поскольку ответственность за обеспечение отсутствия шума и выбросов в исходных данных возложена *этими требованиями* на самого исследователя, то получается, что если что-то получилось не так, то это связано уж не столько с методом кластеризации, сколько с каким-то недоработками самого исследователя. По этим причинам более логично и главное, более *продуктивно* было бы предъявить эти требования не к исходным данным и обеспечивающему их исследователю, а к самому методу кластерного анализа, *который, по мнению авторов, должен корректно работать в случае наличия шума и выбросов в исходных данных*.

Таким образом, четвертое и пятое требования приводят к двум проблемам:

Проблема 4 разработки такого метода кластерного анализа, математическая модель и алгоритм которого органично включали бы фильтр, подавляющий шум в исходных данных, в результате чего данный метод кластеризации корректно работал бы при наличии шума в исходных данных.

Проблема 5 разработки метода кластерного анализа, математическая модель и алгоритм которого обеспечивали бы выявление «выбросов» (артефактов) в исходных

[12] 4. Показатели должны отвечать требованию «устойчивости», под которой понимается отсутствие влияния на их значения случайных факторов. 5. Выборка должна быть однородна, не содержать «выбросов».

данных и позволяли бы либо вообще не показывать их в дендрограммах, либо показывать, но так, чтобы было наглядно видно, что это артефакты.

Далее рассмотрим, как решаются (или не решаются) сформулированные выше проблемы в классических методах кластерного анализа. Для удобства дальнейшего изложения повторим формулировки этих проблем.

Проблема 1.1 выбора метрики, корректной для неортонормированных пространств.

Проблема 1.2 ортонормирования пространства.

Проблема 2.1 сопоставимой обработки описаний объектов, описанных признаками различной природы, измеряемыми в различных единицах измерения (проблема размерностей).

Проблема 2.2 формализации описаний объектов, имеющих как количественные, так и качественные признаки.

Проблема 3.1 доказательства гипотезы о нормальности исходных данных.

Проблема 3.2 нормализации исходных данных.

Проблема 3.3 применения непараметрических методов кластеризации, корректно работающих с ненормализованными данными.

Проблема 4 разработки такого метода кластерного анализа, математическая модель и алгоритм которого органично включали бы фильтр, подавляющий шум в исходных данных, в результате чего данный метод кластеризации корректно работал бы при наличии шума в исходных данных.

Проблема 5 разработки метода кластерного анализа, математическая модель и алгоритм которого обеспечивали бы выявление «выбросов» (артефактов) в исходных данных и позволяли бы либо вообще не показывать их в дендрограммах, либо показывать, но так, чтобы было наглядно видно, что это артефакты.

Сделать это *удобнее всего*, рассматривая какие ответы предлагают классические методы кластерного анализа на сформулированные в работе [2] вопросы:

– как вычислять координаты кластера из двух более объектов;

– как вычислять расстояние до таких «полиобъектных» кластеров от «монокластеров» и между «полиобъектными» кластерами.

Дело в том, что эти вопросы имеют фундаментальное значение для кластерного анализа, т.к. разнообразные комбинации используемых метрик и методов вычисления координат и взаимных расстояний кластеров и порождают все многообразие методов кластерного анализа [2]. Мы бы несколько переформулировали эти вопросы, а также добавили бы еще один:

1. Каким методом вычислять *координаты* кластера, состоящего из одного и более объектов, т.е. каким образом *объединять объекты* в кластеры.

2. Каким методом *сравнивать* кластеры, т.е. как вычислять *расстояния* между кластерами, состоящими из различного количества объектов (одного и более).

3. Каким методом *объединять кластеры*, т.е. формировать *обобщенные* («многообъектные») кластеры.

Вопрос 1-й. Чаше всего ни в теории и математических моделях кластерного анализа, ни на практике между кластером, состоящим из одного объекта («моноообъектным» кластером) и самим объектом не делается *никакого различия*, т.е. считается, что это одно и тоже. «В агломеративно-иерархических методах (aggomerative hierarhical

algorithms) … первоначально все объекты (наблюдения) рассматриваются как отдельные, самостоятельные кластеры, состоящие всего лишь из одного элемента» [2]. В работе [4] также говорится, что древовидная «Диаграмма начинается с каждого объекта в классе (в левой части диаграммы)». Это решение сразу же *порождает* многие из вышеперечисленных проблем (1.1, 1.2, 2.1, 2.2), т.к. объекты могут быть описаны как количественными, так и качественными признаками различной природы, измеряемыми в различных единицах измерения, причем эти признаки *взаимосвязаны* (коррелируют) между собой.

Казалось бы, *проблему размерностей* (2.1) решает кластеризация не исходных переменных, а матриц сопряженности, содержащих *абсолютные частоты* наблюдения признаков по объектам или классам. Однако при таком подходе, например при сравнении моделей автомобилей, *четыре и два* цилиндра у этих моделей, а также *четыре и два болта*, которыми у них прикручен номер, будут давать одинаковый вклад в сходство-различие этих моделей, что едва ли разумно и приемлемо [7]. Тем ни менее матрица сопряженности анализируется в социологических и социометрических исследованиях, а в статистических системах, в разделах справки, посвященных кластерному анализу, приводятся примеры подобного рода.

Другое предложение по решению проблемы размерностей (2.1) основано на четком понимании того, что изменение единиц измерения переменной меняет среднее ее значений и их разброс от этого среднего. Например, переход от сантиметров к миллиметрам увеличивает среднее и среднее отклонение от среднего в 10 раз. Речь идет о методе нормализации или стандартизации исходных данных, когда значения переменных заменяются их стандартизированными значения или z-вкладами [8]. Z-вклад показывает, сколько стандартных отклонений отделяет данное наблюдение от среднего значения:

$$Z_i = \frac{x_i - \bar{x}}{\sigma},$$

где x_i – значение данного наблюдения, $\bar{x}$ – среднее, σ – стандартное отклонение.

Однако этот метод имеет серьезный недостаток, описанный в литературе [2, 4, 8]. Дело в том, что нормализация значений переменных приводит к тому, что независимо от значений их среднего и вариабельности до нормализации (т.е. значимости, измеряемой стандартным отклонением), после нормализации среднее становится равным нулю, а стандартное отклонение 1. Это значит, что *нормализация выравнивает средние и отклонения по всем переменным, снижая, таким образом, вес значимых переменных, оказывающих большое влияние на объект, и завышая роль малозначимых переменных, оказывающих меньшее влияни*е и искажая, таким образом, картину. На взгляд авторов это вряд ли приемлемо. Другой важный недостаток, который в отличие от первого не отмечается в специальной литературе, состоит в том, что стандартизированные значения сложно как-то содержательно интерпретировать, т.е. устранение влияния единиц измерения достигается ценой потери смысла переменных, который как раз и содержался в единицах их измерения. В результате нормализации все переменные становятся как бы «на одно лицо». Это также недопустимо. Таким образом, можно обоснованно сделать вывод о том, *нормализация и стандартизация исходных данных – это весьма радикальное решение проблемы 2.1 «в лоб и в корне», но решение неприемлемо дорогой ценой.*

В классических методах кластерного анализа предлагается два основных варианта *ответов* на 1-й вопрос:

1. Вообще не формировать обобщенных классов или кластеров из объектов, а на всех этапах кластеризации рассматривать только сами первичные объекты.

2. Формировать обобщенные кластеры путем вычисления неких статистических характеристик кластера на основе характеристик входящих в него объектов.

О 1-м варианте ответа в работе [4] говорится: «Диаграмма начинается с каждого объекта в классе (в левой части диаграммы). Теперь представим себе, что постепенно (очень малыми шагами) вы «ослабляете» ваш критерий о том, какие объекты являются уникальными, а какие нет. Другими словами, вы понижаете порог, относящийся к решению об объединении двух или более объектов в один кластер. В результате, вы *связываете* вместе все большее и большее число объектов и агрегируете (*объединяете*) все больше и больше кластеров, состоящих из все сильнее различающихся элементов». Этот подход, когда кластеры реально не формируются, т.к. им не соответствуют какие-либо конструкции математической модели, представляется авторам сомнительным, т.к., во-первых, как было показано выше, это порождает проблемы 1.1, 1.2, 2.1, 2.2, а во-вторых, никак не решает проблемы 3.1, 3.2, 3.3, 4 и 5. *Между тем сам способ формирования кластеров из объектов, по мнению авторов, призван стать средством решения всех этих проблем.*

2-й вариант ответа представляется более обоснованным, однако он сам в свою очередь порождает вопросы о степени корректности и научной обоснованности того или иного метода вычисления обобщенных характеристик кластера и главное о том, *в какой степени этот метод позволяет решить сформулированные выше проблемы*. Описание кластера на основе входящих в него объектов традиционно включает *центр кластера*, в качестве которого обычно используется *среднее* или *центр тяжести* от характеристик входящих в него объектов [2], а также какую-либо количественную оценку степени рассеяния объектов кластера от его центра (как правило, это дисперсия). Ответ на 2-й вопрос является продолжением ответа на 1-й вопрос.

Вопрос 2-й. В работах [2, 3, 4] и других по кластерному анализу описывается большое количество различных мер и методов, которые можно применить как для измерения расстояний между кластерами, так и расстояний от объекта до кластеров. Например, в *невзвешенном центроидном методе* при определении расстояния от объекта до кластера, по сути, определяется расстояние до его центра [4]. В методе *невзвешенного попарного среднего* расстояние между двумя кластерами вычисляется как среднее расстояние между всеми парами объектов в них [4]. При этом, как правило, не решаются перечисленные выше проблемы, т.к. *не устраняются их причины*: а именно средние вычисляются на основе мер расстояния, корректных только для ортонормированных пространств и при этом часто используются размерные или нормализованные формы представления признаков объектов, не формализуется описание объектов, обладающих как количественными, так и качественными признаками. Ответ на 3-й вопрос является продолжением ответа на 2-й вопрос.

Вопрос 3-й. При объединении кластеров характеристики вновь образованного обобщенного кластера обычно пересчитываются тем же методом, каким они рассчитывались для исходных кластеров. Это сохраняет нерешенными и все проблемы, которые были при определении характеристик исходных кластеров и расстояний между этими кластерами.

Основной ***вывод***, который, по мнению авторов можно обоснованно сделать по материалам данной статьи, состоит в том, что, не смотря на существование огромного количества различных методов кластеризации, в этой области существует ряд нерешенных проблем, ждущих своего решения. Анализ этих проблем позволяет высказать ***гипотезу***, что для их решения необходимо выйти за пределы понятийного поля чисто математических рассуждений и привлечь представления из области искусственного интеллекта, в частности основываться на четкой дефиниции содержания таких основополагающих понятий, как данные, информация и знания [7]. Данная статья и содержит описание авторского варианта реализации этой идеи.

Здесь же хотелось бы отметить, что кластеризация классическим методом матрицы знаний, полученной вне статистической системы, реализующий кластерный анализ, не дает желаемых результатов, т.к. только 1-я итерация получается соответствующей предлагаемому подходу, а последующие дают ошибочные результаты, т.к. в статистических системах не реализованы операции обобщения и добавление объекта к кластеру или объединение классов в кластер осуществляется иначе, чем формирование самих классов в исходной матрице знаний.

Материалы данной статьи могут быть использованы при разработке интеллектуальных систем, а также при проведении лабораторных работ по дисциплинам: «Интеллектуальные информационные системы» для специальности: 080801.65 – Прикладная информатика (по областям) и «Представление знаний в информационных системах» для специальности: 230201.65 – Информационные системы и технологии.

Примечания:

1. Мандель И.Д. Кластерный анализ. М.: Финансы и статистика, 1988. 176 с.
2. Леонов В.П. Краткий обзор методов кластерного анализа. URL: http://www.biometrica.tomsk.ru/cluster_2.htm http://www.biometrica.tomsk.ru/cluster_3.htm
3. Леонов В.П. Литература и сайты по кластерному анализу. URL: http://www.biometrica.tomsk.ru/cluster_4.htm
4. Сайт Института Космических Исследований РАН. URL: http://www.iki.rssi.ru/magbase/REFMAN/ STATTEXT/ modules/stcluan.html#general
5. Сайт Internet-сообщества закупщиков. URL: http://zakup.vl.ru/132-metodi_klastern.html
6. Баран О.И., Григорьев Ю.А., Жилина Н.М. Алгоритмы и критерии качества кластеризации // Общественное здоровье и здравоохранение: материалы XLV науч.-практ. конф. с междунар. участием «Гигиена, ор-

References:

1. Mandel I.D. Cluster analysis. M.: Finances and statistics, 1988. 176 p.
2. Leonov V.P. The short review of Cluster analysis methods. URL: http://www.biometrica.tomsk.ru/cluster_2.htm http://www.biometrica.tomsk.ru/cluster_3.htm
3. Leonov V.P. Literature and sites on cluster analysis. URL: http://www.biometrica.tomsk.ru/cluster_4.htm
4. The site of the Space Research Institute of the Russian Academy of Sciences. URL: http://www.iki.rssi.ru/magbase/REFMAN/ STATTEXT/ modules/stcluan.html#general
5. The site of purchasers' Internet-community. URL: http://zakup.vl.ru/132-metodi_klastern.html
6. Baran O.I., Grigorjev Yu.A., Zhilina N.M. The algorithms and criteria of clusterization quality // Public health and public health service: materials of the XLV scient.-pract. conf. with international participation of «Hygiene,

ганизация здравоохранения и профпатология» и семинара «Актуальные вопросы современной профпатологии», Новокузнецк, 17-18 ноября 2010 / под ред. В.В. Захаренкова. Кемерово: Примула, 2010. С. 21-26.

7. Луценко Е.В. Методологические аспекты выявления, представления и использования знаний в АСК-анализе и интеллектуальной системе «Эйдос» // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета. Краснодар: КубГАУ, 2011. № 06(70). С. 233-280. URL: http://ej.kubagro.ru/2011/06/pdf/18.pdf, 3 у.п.л.
8. Близоруков М.Г. Статистические методы анализа рынка: учеб.-метод. пособие. Екатеринбург: Ин-т управления и предпринимательства Урал. гос. ун-та, 2008. 75 с. URL: http://elar.usu.ru/bitstream/1234.56789/1671/6/1334937_schoolbook.pdf

organization of public health services and professional pathology» and of the seminar «The topical questions of professional pathology», Novokuznetsk, November, 17-18th 2010 / ed. by V.V. Zakharenkov. Kemerovo: Primula, 2010. P. 21-26.

7. Lutsenko E.V. The methodological aspects of revealing, representation and use of knowledge in the ASK-analysis and in the intellectual system «Eidos» // Multidisciplinary network electronic scientific journal of the Kuban State Agrarian university. Krasnodar: KubGAU, 2011. No. 06(70). P. 233-280. URL: http://ej.kubagro.ru/2011/06/pdf/18.pdf, 3 у.п.л.
8. Blizorukov M.G. Statistical methods of the market analysis: manual. Ekaterinburg: The institute of management and entrepreneurship of the Ural state university, 2008. 75 p. URL: http://elar.usu.ru/bitstream/1234.56789/1671/6/1334937_schoolbook.pdf