УДК 519.237.8:519.233.5

С.Н. ЛАПАЧ[*], С.Г. РАДЧЕНКО[*]

РЕГРЕССИОННЫЙ АНАЛИЗ В УСЛОВИЯХ НЕОДНОРОДНОСТИ ФАКТОРНОГО ПРОСТРАНСТВА

[*]Национальный технический университет Украины «Киевский политехнический институт», Киев, Украина

Анотація. *Досліджується застосування нечіткого кластерного аналізу для виділення однорідних підобластей факторного простору при побудові регресійних моделей. Викладено застосування нечіткого кластерного аналізу. Проведений обчислювальний експеримент показав, що необхідно зробити аналіз результатів по суті задачі і перевірку різних варіантів розбиття на кластері для отримання правильного розв'язку. Виконано аналіз моделювання болтового з'єднання композиційних матеріалів в авіабудуванні. Приведено діаграми розподілу експериментів по 4-х кластерах. Дано рекомендації щодо формалізації процесу підбору методів і засобів з метою розбиття на однорідні підобласті факторного простору при апріорі невідомих формі і кількості кластерів.*
Ключові слова: *кластерний аналіз, регресійний аналіз, нечіткий кластерний аналіз, розбиття факторного простору на однорідні підобласті.*

Аннотация. *Исследуется применение кластерного анализа для выделения однородных подобластей факторного пространства при построении регрессионных моделей. Изложено применение нечеткого кластерного анализа. Проведение вычислительного эксперимента показало, что необходим смысловой анализ результатов и проверки различных вариантов разбиения на кластеры для получения правильного решения. Проанализировано моделирование болтового соединения композиционных материалов в авиастроении. Приведены диаграммы распределения экспериментов по 4-м кластерам. Даны рекомендации по формализации процесса подбора методов и средств с целью разделения на однородные подобласти факторного пространства при заранее не известных форме и количестве кластеров.*
Ключевые слова: *кластерный анализ, регрессионный анализ, нечеткий кластерный анализ, разбиение факторного пространства на однородные подобласти.*

Abstract. *The application of cluster analysis for selection of homogeneous subfields of the factor space under the building of regression models is investigated. The use of fuzzy cluster analysis was outlined. The computational experiment has shown that it is necessary to make semantic analysis of the results and tests of the different options of partitioning on clusters to obtain the correct solution. Simulation analysis of bolted connection of composite materials in aircraft construction was done. The charts of distribution of experiments in 4 clusters were given. Recommendations for formalization of processes of selection of methods and tools in order to separate into homogeneous subfields of the factor space with an a priori unknown form and number of clusters.*
Keywords: *cluster analysis, regression analysis, fuzzy cluster analysis, partitioning on the factor space into homogeneous subfields.*

1. Введение

Качественную и надежную регрессионную модель невозможно построить в случае, если факторное пространство неоднородно или разрывно. Необходимо определить неразрывные (однородные) подобласти и построить в каждой модель отдельно. Эта проблема в настоящее время не получила разрешения [1, 2]. Использование традиционного кластерного анализа не приносит гарантированного успеха даже в достаточно простых случаях. Успешность его применения зависит от значений параметров, которые нужно подбирать в соответствии с формой кластеров. В общем случае уверенное определение возможно для кластеров простой вытянутой формы, расстояние между которыми больше, чем расстояние между элементами в кластере [3–5]. Для широкого практического применения это малопригодно, поскольку, во-первых, форма кластеров и их расположение a priori неизвестно, а

ISSN 1028-9763. Математичні машини і системи, 2016, № 3

во-вторых, кластеры часто имеют сложную форму и частично связаны друг с другом, а иногда и частично перекрываются.

Цель статьи – исследовать возможность применения нечеткого кластерного анализа для выделения однородных подобластей факторного пространства при построении регрессионных моделей.

Постановка вопроса: определение однородных подобластей факторного пространства.

2. Нечеткий кластерный анализ

При нечеткой кластеризации методом k-средних предполагается [6], что некоторые точки могут принадлежать нескольким кластерам одновременно. Принадлежность элементов выборки к определенному кластеру описывается матрицей $\mathbf{U}=\left[\mu_{ij}\right]$, $\mu_{ij}\in 0,1$, $i=\overline{1,N}$, $j=\overline{1,k}$. Строка i содержит значение, соответствующее степени принадлежности объекта i к кластеру j. При этом $\sum_{j=1}^{k}\mu_{ij}=1,\ 0<\sum_{i=1}^{N}\mu_{ij}<N$, здесь N – количество объектов (в нашем случае число опытов), k – число кластеров.

Кластеризация выполняется следующим образом:

1. Выбираются параметры: k, g_e – экспоненциальный вес, ε – значение критерия остановки.

2. Случайным образом генерируется начальная матрица нечеткого разбиения на кластеры **U**.

3. Рассчитываются центры кластеров по формуле

$$V_j=\frac{\sum_{i=1}^{N}(\mu_{ij})^{ge}X_i}{\sum_{i=1}^{N}(\mu_{ij})^{ge}},\ j=\overline{1,k},$$

где V_j – вектор длиной M;

X_i – матрица размера $N\times M$;

M – число факторов (при разбиении на кластеры отклик входит в число факторов).

4. Рассчитываются расстояния между объектами и центрами кластеров $d_{ij}=\sqrt{\|X_i-V_i\|^2}$.

5. Пересчитываются элементы матрицы нечеткого разбиения:

$$\mu_{ij}=\begin{cases}\dfrac{1}{\left(d_{ji}^2\sum\limits_{l=1}^{k}\dfrac{1}{d_{li}^2}\right)^{1/(g_e-1)}},\ d_{ij>0}\\ \begin{cases}1, d_{ij}=0, i=j\\ 0, d_{ij}=0, i\neq j\end{cases}\end{cases}.$$

6. Проверяется условие остановки. Если $\left\|U-U^*\right\|<\varepsilon$, то процедура заканчивается, иначе – выполняется переход в п.3.

Для оценки качества разбиения используют параметры рассеивания. Общее рассеивание $S=\sum_{i=1}^{N} d_{i\overline{X}}^2$, где $\overline{X}=\frac{1}{N}\sum_{i=1}^{N} X_i$ – общий центр веса, межгрупповое рассеивание (между центрами кластеров) $B=\sum_{i=1}^{k}\sum_{j=1}^{k} d^2(\overline{X}_i,\overline{X}_j)$, внутригрупповое рассеивание $Q=\sum_{i=1}^{k}\sum_{j=1}^{n_i} d^2(X_j,\overline{X}_i)$. Здесь n_i – количество элементов в кластере.

Для нечеткого кластерного анализа в этих формулах добавляется множитель μ_{ij}. Статистика T, которая показывает долю общего рассеивания, поясняемую межгрупповым рассеиванием, определяется как $T=1-Q/S$. Кластером считается множество точек, для которых выполняется условие $Q_i/N<S/N$, а сгущением – множество точек при выполнении условия $\max\ d_i^2\ <S/N$.

Был проведен вычислительный эксперимент со специально сконструированными кластерами разной формы и с разным расстоянием друг от друга. Анализ их результатов показал, что, с одной стороны, нечеткий кластерный анализ может определять кластеры произвольной формы и с различными расстояниями друг от друга. С другой стороны, без смыслового анализа результатов и проверки различных вариантов разбиения на кластеры правильное решение невозможно.

Для определения разделения на кластеры с использованием нечеткого кластерного анализа предлагается выполнять следующие действия:

1. Выдвижение гипотезы о количестве кластеров в выборке.
2. Выполнение разбиения на кластеры.
3. Анализ результатов разбиения и выдвижение уточненной гипотезы о количестве кластеров. Здесь выполняется как анализ качества разбиения с точки зрения кластерного анализа в сравнении с другими вариантами, так и смысловой анализ полученного разбиения на кластеры с точки зрения знаний предметной области.
4. Проверка уточненной гипотезы.
5. П.п. 2–4 могут повторяться несколько раз до получения удовлетворительных результатов.
6. Выбор наилучшего разбиения и использование его для регрессионного анализа.

3. Моделирование болтового соединения композиционных материалов

В [7] подробно описана задача по моделированию болтового соединения композиционных материалов в авиастроении. Факторы $X_{\text{факт1}}\ldots X_{\text{факт10}}$, с которыми была построена математическая модель, приведены в табл. 1.

В работе была получена регрессионная модель разрушающей удельной нагрузки от описанных выше факторов. Модель имеет отличные информационные свойства, хорошую вычислительную стойкость. К недостаткам относится формальная неадекватность модели по критерию Фишера. Кроме того, описывающие свойства можно назвать только удовлетворительными, что контрастирует с высокой информативностью.

Неадекватность и недостаточно хорошие описывающие свойства предположительно связаны с неоднородностью факторного пространства. Разработанный алгоритм разделения на кластеры с помощью нечеткого кластерного анализа был апробирован в этой задаче, как раз требующей именно такого решения, поскольку описывающие свойства модели не позволяют использовать её в системах автоматизированного проектирования.

Таблица 1. Описание факторов

Фактор	Название	Условное обозначение	Обозначение уровня в матрице плана эксперимента	Натуральное значение уровня
$X_{\text{факт1}}$	Толщина пластины болтового соединения с усилением, мм	δ_c	$n_1 = -1; 0; 1$	$\delta_c = d_m + 2n_1 + d_m n_1 / 6$
$X_{\text{факт2}}$	Диаметр болта номинальный, мм	d_m	6; 8; 10; 12	6; 8; 10; 12
$X_{\text{факт3}}$	Величина перемычки вдоль действия силы (от центра отверстия до края пластины), мм	a	2; 2,5; 3; 4	ad_m
$X_{\text{факт4}}$	Величина перемычки поперек действия силы (от центра отверстия до края пластины), мм	b	2; 2,5; 3; 3,5	bd_m
$X_{\text{факт5}}$	Относительная величина усиления толщины пластины	k_y	0,2; 0,4; 0,6; 1	0,2; 0,4; 0,6; 1
$X_{\text{факт6}}$	Угол направления волокон в усиливающих слоях относительно направления действия силы, град	φ	0; 30; 60; 90	0; 30; 60; 90
$X_{\text{факт7}}$	Кол-во прослоек усиления	n_2	0; 1; 2; 3	$2 + (0{,}5d_m - 2)n_2$
$X_{\text{факт8}}$	Характер посадки болта в отверстии пластины	П	0; 1; 2; 3	H9/h6; H9/h6 $d_m + 0{,}1$; H9 ГОСТ131042-79 H9/h6 ВК-9
$X_{\text{факт9}}$	Кол-во болтов и их шаг в соединении, мм	m	0; 1; 2; 3; 4; 5; 6	1; $3\times3{,}5d_m$; $3\times4d_m$; $3\times5d_m$; $5\times3{,}5d_m$; $5\times4d_m$; $5\times5d_m$
$X_{\text{факт10}}$	Разбиение плана на ортогональные блоки	Б	0; 1; 2; 3; 4; 5; 6; 7	0; 1; 2; 3; 4; 5; 6; 7

Первоначально было выдвинуто предположение о разделении пространства на два кластера. Явное разделение на два кластера подтверждено кластерным анализом. В каждом кластере была получена регрессионная модель. Их характеристики приведены в табл. 2.

Как видно из таблицы, полученные модели $\hat{Y}_1$ и $\hat{Y}_2$ имеют близкие характеристики к модели $\hat{Y}$, полученной по всей выборке, при этом их описывающие свойства значительно лучше. Это показывает анализ последних четырех строк табл. 2. Вместе с тем эти характеристики все еще неудовлетворительны с точки зрения требований предметной области: погрешности слишком велики для практического использования в САПР. В связи с этим проведен анализ принадлежности объектов (экспериментов) к отдельным кластерам, то есть значения элементов массива U (рис. 1). Из рисунка видно, что прослеживается предположительное разбиение выборки на 4 кластера.

Таблица 2. Статистические характеристики моделей для всей выборки и при разбиении на два кластера

Параметры статистического анализа		Условные обозначения	$\hat{Y}$	$\hat{Y}_1$	$\hat{Y}_2$
Проверка результатов опытов на однородность	Дисперсия воспроизводимости	$s^2_{\text{восп}}$	0,608812	0,927009	0,29016
	Среднеквадратическое отклонение	$s_{\text{восп}}$	0,780264	0,962813	0,539088
	Число степеней свободы для дисперсии воспроизводимости	$f_{\text{восп}}$	64	32	32
	Экспериментальное значение *G*-критерия	$G^{\text{эксп}}$	0,131403	0,172598	0,225948
	Критическое значение *G*-критерия	$G^{\text{крит}}$	0,165178	0,289486	0,289486
	Уровень значимости	α	0,05	0,05	0,05
	Однородность дисперсий		Однородные	Однородные	Однородные
Число обусловленности		cond($\mathbf{X}^1\mathbf{X}$)	1,91327	1,11724	2,02779
Проверка гипотезы об адекватности	Дисперсия адекватности	$s^2_{\text{ад}}$	7,45993	4,45534	1,94985
	Экспериментальное значение *F*-критерия	$F^{\text{эксп}}$	12,2533	4,80616	6,70929
	Критическое значение *F*-критерия для адекватности	$F^{\text{крит}}$	1,54931	2,31	2,31
	Число степеней свободы для адекватности	$f_{\text{ад}}$	49	11	11
	Уровень значимости	α	0,05	0,05	0,05
	Адекватность модели		Неадекватн.	Неадекватн.	Неадекватн.
Анализ полученной модели на информативность	Коэффициент множественной корреляции	R	0,958281	0,980104	0,989976
	Число степеней свободы для коэффициентов модели	$f_{k'}$	14	11	11
	Число степеней свободы для остаточной суммы квадратов	$f_{\text{ост}R}$	113	52	52
	Экспериментальное значение *F*-критерия	$F^{\text{эксп}}$	39,3405	44,3326	89,3332
	Критическое значение *F*-критерия для информативности	$F^{\text{крит}}$	1,78025	1,97821	1,97821
	Уровень значимости	α	0,05	0,05	0,05
	Значение параметра для критерия Бокса и Веца	γ	4	3	5
	Информативность модели		Высокая	Хорошая	Высокая
Описывающие свойства	Среднее абсолютных величин погрешностей аппроксимации	$\lvert\bar{e}_u\rvert$	3,0754	1,36096	0,942767
	Доля рассеивания, объясняемая моделью	$Q_{\hat{y}}$	0,918302	0,960603	0,980053
	Средняя погрешность аппроксимации, %	$\varepsilon_{\text{ср}}$	22,5927	9,9847	8,27
	Максимальная погрешность аппроксимации, %	$\varepsilon_{\max}$	77,89	42,06	17,66

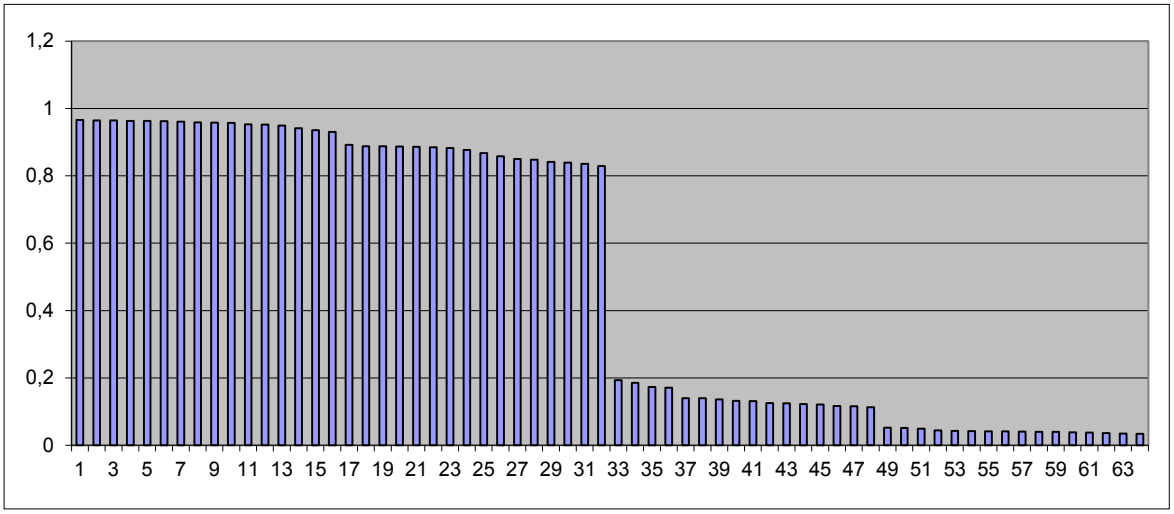

Рис. 1. Степень принадлежности (по вертикали) экспериментов (по горизонтали) к разным кластерам

Проверка разбиения на четыре кластера показала, что это разбиение предпочтительней с точки зрения кластерного анализа (табл. 3, рис. 2).

Таблица 3. Качество разбиения при разном количестве кластеров

Характеристика разбиения	Количество кластеров	
	4	2
Параметр экспоненциального веса	2	2
Критерий остановки	0,00001	0,00001
Внутригрупповое рассеивание	3716,965	17876,97
Межгрупповое рассеивание	36212,92	7684,063
Качество разделения	0,906914	0,300324

Внутригрупповое рассеивание для случая четырех кластеров на порядок меньше межгруппового в отличие от сравнимых характеристик этих величин для разбиения на два кластера. Это позволяет отдавать предпочтение четырем кластерам [5].

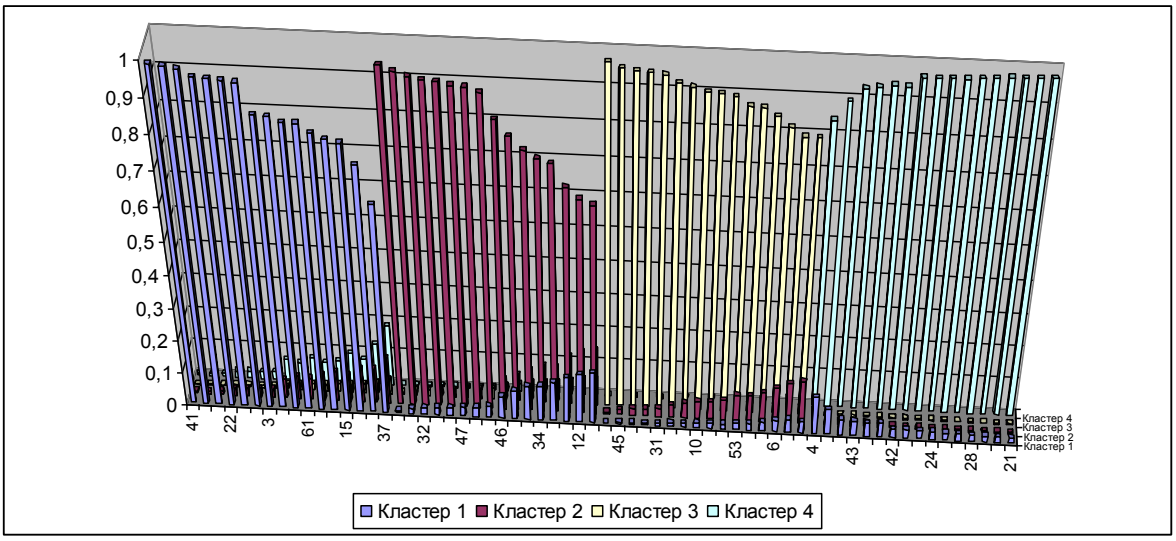

Рис. 2. Диаграмма распределения экспериментов по 4-м кластерам

Смысловой анализ разбиения показал, что каждому кластеру соответствует свой угол ориентации волокон композита (значение фактора $X_{\text{факт6}}$ в табл. 1). То есть в кластерах собраны эксперименты, в которых значение угла ориентации волокон составляет 0° для первого кластера, 30° для второго и соответственно 60° и 90° для третьего и четвертого. Это позволяет сделать физическую интерпретацию причин разбиения именно на такие кластеры.

В каждом кластере была построена регрессионная модель. Характеристики моделей $\hat{Y}_0$, $\hat{Y}_{30}$, $\hat{Y}_{60}$, $\hat{Y}_{90}$ приведены в табл. 4. Анализ табл. 4 показал, что описательные свойства этих моделей значительно лучше, чем модели $\hat{Y}$, построенной на всей выборке, и моделей $\hat{Y}_1$ и $\hat{Y}_2$, полученных для двух кластеров (табл. 3), и они могут быть использованы на практике.

Таблица 4. Статистические характеристики моделей для всей выборки и при разделении на четыре кластера

Параметры статистического анализа		Условные обозначения	$\hat{Y}$	$\hat{Y}_0$	$\hat{Y}_{30}$	$\hat{Y}_{60}$,	$\hat{Y}_{90}$
1	2	3	4	5	6	7	8
Проверка результатов опытов на однородность	Дисперсия воспроизводимости	$s^2_{\text{восп}}$	0,608812	0,233894	0,347337	1,04414	0,809881
	Среднеквадратическое отклонение	$s_{\text{восп}}$	0,780264	0,483625	0,589353	1,02183	0,899934
	Число степеней свободы для дисперсии воспроизводимости	$f_{\text{восп}}$	64	16	16	16	16
	Экспериментальное значение *G*-критерия	$G^{\text{эксп}}$	0,131403	0,561486	0,166408	0,267569	0,395120
	Критическое значение *G*-критерия	$G^{\text{крит}}$	0,165178	0,451677	0,451677	0,451677	0,451677
	Уровень значимости	α	0,05	0,05	0,05	0,05	0,05
	Однородность дисперсий		Однородные	Неоднородные	Однородные	Однородные	Однородные
Число обусловленности		$\text{cond}(\mathbf{X}^{\mathrm{T}}\mathbf{X})$	1,91327	1,27991	1,4872	2,20828	2,01531
Проверка гипотезы об адекватности	Дисперсия адекватности	$s^2_{\text{ад}}$	7,45993	0,387069	1,16761	1,25423	6,04992
	Экспериментальное значение *F*-критерия	$F^{\text{эксп}}$	12,2533	1,6549	3,3615	1,20123	7,47013
	Критическое значение *F*-критерия для адекватности	$F^{\text{крит}}$	1,54931	2,6572	2,6572	2,34194	2,49351
	Число степеней свободы для адекватности	$f_{\text{ад}}$	49	8	8	9	5
	Уровень значимости	α	0,05	0,05	0,05	0,05	0,05

Продолж. табл. 4

	Адекватность модели		Неадекватн.	Адекватн.	Неадекватн.	Адекватн.	Неадекватн.
1	2	3	4	5	6	7	8
Анализ полученной модели на информативность	Коэффициент множественной корреляции	R	0,958281	0,996182	0,995247	0,996684	0,983331
	Число степеней свободы для коэффициентов модели	$f_{k'}$	14	8	8	9	5
	Число степеней свободы для остаточной суммы квадратов	$f_{\text{ост}R}$	113	23	23	22	26
	Экспериментальное значение F-критерия	$F^{\text{эксп}}$	39,3405	113,929	91,4166	100,021	25,7505
	Критическое значение F-критерия для информативности	$F^{\text{крит}}$	1,78025	2,37481	2,37481	2,34194	2,58679
	Уровень значимости	α	0,05	0,05	0,05	0,05	0,05
	Значение параметра для критерия Бокса и Веца	γ	4	5	4	4	2
	Информативность модели		Высокая	Высокая	Высокая	Высокая	Хорошая
Среднее абсолютных величин погрешностей аппроксимации		$\lvert\bar{e}_u\rvert$	3,0754	0,310894	0,637529	0,548391	1,55774
Доля рассеивания, объясняемая моделью		$Q_{\hat{y}}$	0,918302	0,992387	0,990515	0,993379	0,928007
Средняя погрешность аппроксимации, %		$\varepsilon_{\text{ср}}$	22,5927	4,48608	4,31869	3,61086	13,659
Максимальная погрешность аппроксимации, %		$\varepsilon_{\max}$	77,89	14,3384	15,0079	16,1326	49,6679

Степень улучшения по сравнению с моделью для всей выборки $\hat{Y}$ показана в табл. 5. Как видно, имеют место существенные улучшения характеристик – это улучшения в разы.

Таблица 5. Улучшение описывающих характеристик моделей $\hat{Y}_0$, $\hat{Y}_{30}$, $\hat{Y}_{60}$, $\hat{Y}_{90}$

Характеристика	Обозначение	Улучшение (разы)
Среднее абсолютных величин погрешностей аппроксимации	$\lvert\bar{e}_u\rvert$	≈2…10
Средняя погрешность аппроксимации, %	$\varepsilon_{\text{ср}}$	1,65…6,26
Максимальная погрешность аппроксимации, %	$\varepsilon_{\max}$	1,57…5,43

4. Выводы и рекомендации

Проведенные исследования по использованию нечеткого кластерного анализа для выделения однородных (неразрывных) подобластей факторного пространства позволили установить, что его можно использовать для уверенного определения областей типа сгущений.

Алгоритм успешно апробирован на реальных задачах. Вместе с тем общий алгоритм не является полностью формализованным и требует выдвижения предположений о количестве кластеров, сравнения результатов разбиения с точки зрения кластерного анализа, характеристик полученных регрессионных моделей и смыслового анализа полученного разбиения.

Раньше было установлено [4, 5], что для кластеров, имеющих протяженные формы, успешно применяются классические алгоритмы кластерного анализа с подбором параметров.

Поскольку для реальных задач форма кластеров a priori неизвестна, то в настоящее время требуется экспериментирование с подбором методов и смысловым анализом. Направлением дальнейшей работы могут быть формализация и программное обеспечение процессов подбора методов или параметров кластерного анализа с целью разделения на однородные подобласти факторного пространства при заранее не известных форме и количестве кластеров.

СПИСОК ЛИТЕРАТУРЫ

1. Котюков В.И. Многофакторные кусочно-линейные модели / Котюков В.И. – М.: Финансы и статистика, 1984. – 216 с.
2. Лапач С.Н. Основные проблемы построения регрессионных моделей / С.Н. Лапач, С.Г. Радченко // Математичні машини і системи. – 2012. – № 4. – С. 125 – 133.
3. Лапач С.Н. Статистические методы в фармакологии и маркетинге фармацевтического рынка / Лапач С.Н., Пасечник М.Ф., Чубенко А.В. – К.: ЗАТ «Укрспецмонтаж», 1999. – 312 с.
4. Лапач С.М. Кластерний аналіз при визначенні однорідних областей факторного простору в регресійному аналізі / С.М. Лапач // П’ятнадцята міжнар. конф. ім. акад. Михайла Кравчука, (м. Київ, 15–17 травня 2014 р.). – Т. 3: Теорія ймовірностей та математична статистика. – К.: НТУУ «КПІ», 2014. – С. 82 – 84.
5. Лапач С.М. Визначення оптимальної кількості кластерів / С.М. Лапач // Математичні машини і системи. – 2015. – № 3. – С. 53 – 56.
6. Штовба С.Д. Проектирование нечетких систем средствами MATLAB / Штовба С.Д. – М.: Горячая линия – Телеком, 2007. – 288 с.
7. Математическое моделирование прочности болтовых соединений композиционных материалов типа углепластиков / С.Г. Радченко, С.Н. Лапач, А.З. Двейрин [и др.] // Открытые информационные и компьютерные интегрированные технологии: сб. научных трудов. – Харьков: «ХАИ», 2014. – Вып. 63. – С. 61 – 71.

Стаття надійшла до редакції 17.03.2016