УДК 004.9

МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ ЭВОЛЮЦИИ КЛАСТЕРНЫХ ОБРАЗОВАНИЙ

Нуриев Н.К., Аль-Хашеди А.А., Печеный Е.А.

ФГБОУ ВО «Казанский национальный исследовательский технологический университет», Казань, e-mail: alhashedi@mail.ru

В настоящей работе показана возможность использования аппарата анализа временных рядов для отслеживания эволюционных изменений кластерных образований, используемых в задачах распознавания образов. Известно, что кластеры широко и достаточно эффективно применяются для решения задач распознавания на обучающих выборках с небольшой размерностью пространства классифицирующих признаков. Однако, как правило, положение кластеров полагается жестко фиксированным, а размеры неизменными. Здесь предлагается рассматривать их как самоорганизующиеся и саморазвивающиеся динамические структуры, ход эволюции которых можно и нужно контролировать и прогнозировать. Введено понятие кластерной последовательности, как множества состояний кластера, положение, размеры и количество элементов в составе которого достоверно известны в дискретные моменты времени. Сформулированы признаки их успешного развития и деградации. Описан разработанный алгоритм, позволяющий отслеживать текущие изменения и на их основе прогнозировать последующие эволюционные изменения. Рассмотрен и подробно прокомментирован демонстрационный пример, где описано поведение трех различных кластерных последовательностей в пространстве двух классифицирующих признаков. Показана эффективность разработанного алгоритма и методики его использования для любых видов взаимовлияния классифицирующих признаков друг на друга.

Ключевые слова: эволюционная модель, распознавание образов, математическая модель, кластерный анализ, анализ временных рядов, алгоритм, прогнозирование

MATHEMATICAL MODELING OF EVOLUTION OF CLUSTER FORMATIONS

Nuriev N.K., Al-Khashedi A.A., Pechenyy E.A.

Federal State Budget Educational Institution of Higher Education «Kazan National Research Technological University», Kazan, e-mail: alhashedi@mail.ru

In this paper, we show the possibility of using the time series analysis apparatus to track the evolutionary changes in cluster formations studied in pattern recognition problems. It is known that clusters are widely and fairly effectively used to solve recognition problems on training samples with a small dimension of the space of classifying traits. However, as a rule, the position of clusters is assumed to be rigidly fixed, and the dimensions unchanged. It is proposed to consider them as self-organizing and self-developing dynamic structures, the evolution of which can and should be controlled and predicted. The notion of a cluster sequence as sets of cluster states, position, sizes and number of elements in the composition of which are reliably known at discrete instants of time is introduced. The signs of their successful development and degradation are formulated. A developed algorithm is described that allows to track current changes and on their basis to predict subsequent evolutionary changes. A demonstration example is described and explained in detail, where the behavior of three different cluster sequences in the space of two classifying features is described. The efficiency of the developed algorithm and the methodology of its use for any kinds of interference of the classifying characteristics against each other are shown.

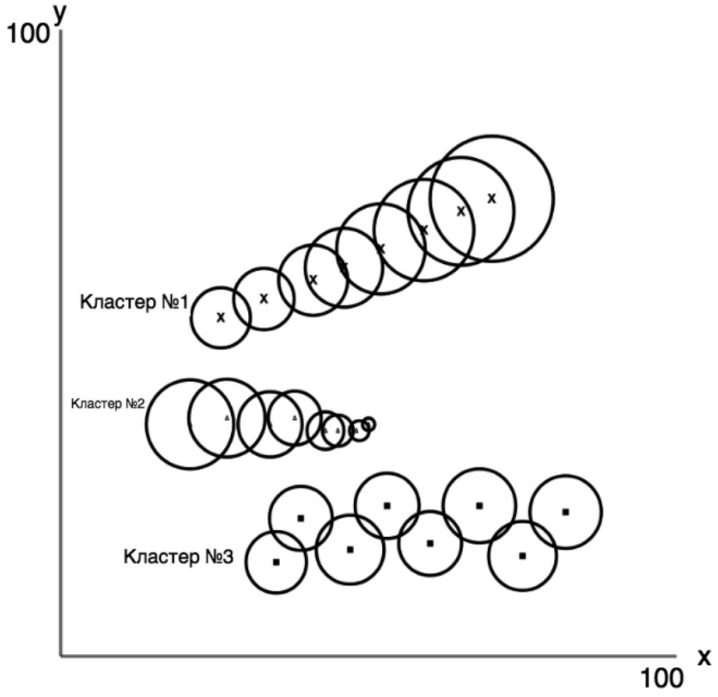
Keywords: evolutionary model, pattern recognition, mathematical model, cluster analysis, time series analysis, algorithm, forecasting

Кластерный анализ и отдельные его элементы достаточно часто используются при решении задач распознавания образов [1–3]. С помощью этого аппарата удается, в рамках принятой метрики, сформировать более или менее плотные группировки в составе обучающей выборки, что позволяет заметно ускорить процедуру классификации. Однако в современных реалиях гипотеза о неизменности положения и границ построенных кластеров представляется малоправдоподобной. Действительно: изменения интенсивности и состава потоков, пополняющих обучающую выборку, миграция объектов между кластерами и т.д. могут стать причиной изменения как отдельных кластерных образований, так и структуры кластерного множества в целом.

В работе [4] был рассмотрен и проиллюстрирован на конкретном числовом примере саморазвивающийся алгоритм распознавания, позволяющий в пространстве классифицирующих признаков формировать кластеры и осуществлять корректировку их положения и размеров сообразно состоянию обучающей выборки. При этом, однако, давалось описание лишь результата действия изменений обучающей выборки и никак не исследовалось влияние их основной движущей силы – времени.

Цель статьи: разработка математической модели эволюционных изменений кластерных образований, используемых в задачах распознавания образов.

Методом исследования является аппарат анализа временных рядов.



Пример расположения кластерных последовательностей

Результаты исследования и их обсуждение

Пусть в фазовом пространстве классифицирующих признаков сформированы несколько последовательностей кластеров, заданных координатами центров своих кластерных сфероидов. Каждому кластеру поставлен в соответствие определенный момент времени $t_i\ i=\overline{1,n}$, на который вполне установлено положение его центра в соответствии с алгоритмом, подробно описанном в работе [4]. На рисунке приведен пример, где таких последовательностей три, а положения кластерных центров обозначены символами x, Δ, ■.

Таким образом, фактически, описанные выше последовательности представляют собой ряды динамики, уровни которых задаются не отдельными числами, а числовыми множествами, чьи мощности равны размерности пространства классифицирующих признаков. И если эта размерность равна *m*, то для построения модели поведения каждой кластерной последовательности необходимо получить *m* уравнений вида $x_j = p_j(t)\ j=\overline{1,m}$, где x_j – элементы множества признаков. Естественно, что чем больше число *m*, тем больше проблем может возникнуть при идентификации параметров модели и сложностей с интерпретацией результатов моделирования. Следовательно, эволюционные изменения каждой из выявленных кластерных последовательностей полностью характеризуются *m* временными рядами по количеству классифицирующих признаков исследуемой обучающей выборки.

Как известно [5], значения уровней временных рядов формируются как результат совокупного действия трех составляющих: трендовой *T*, циклической *S* и случайной *E*. Поэтому анализ временных рядов рекомендуется начинать с оценки независимого вклада каждой из них. Это осуществляется путем построения коррелограмм или таблиц коэффициентов автокорреляции. Если будет выявлено, что преобладающее влияние на формирование уровней ряда имеет трендовая составляющая, то соответствующий классифицирующий признак признается значимо влияющим на ход эволюционных изменений кластерной последовательности. После этого устанавливается тип тренда и строится регрессионная модель, которая после должной статистической проверки может быть использована для прогнозирования дальнейших этапов эволюции. При наличии преобладающих трендовых составляющих у двух и более признаков следует проверить гипотезу коинтегрированности, то есть установить, являются ли выявленные тренды взаимообусловленными или нет. Для решения этой задачи могут быть использованы, например, метод отклонения от тренда [5], кри-

терий Ингла – Грэнджера [6–8] и ряд других. Если гипотеза коинтегрированности нашла свое подтверждение, то совокупное влияние трендов соответствующих признаков должно быть обязательно учтено при прогнозировании эволюционного поведения кластерной последовательности. В противном случае сохранение трендовых тенденций в моменты времени, следующие за наблюдаемыми, не гарантировано и, следовательно, использование их в прогнозировании неправомочно.

В случае преобладания, или заметного присутствия во временном ряде какого либо признака с циклической составляющей, ее влияние должно быть элиминировано одним из известных методов [5]. В сглаженном таким образом временном ряду оценивается трендовая составляющая, неразличимая на фоне циклических изменений или обосновывается ее незначимость на ход эволюционного процесса. Если окажется, что влияние тренда существенно, его вклад учитывается в соответствии с описанной выше процедурой, иначе его следует рассматривать как шум – случайную составляющую с предположительно нормальным законом распределения. При этом действие, оказываемое соответствующим признаком на эволюцию, признается ничтожным.

Весьма важным показателем, характеризующим эволюционные преобразования кластерных последовательностей, является фактор, не принадлежащий пространству классифицирующих признаков: мощность множества элементов, входящих в состав кластера. Эта величина с течением времени может и должна меняться, поскольку количество элементов, принадлежащих кластеру, может как возрастать, так и убывать. Рост будет наблюдаться либо при захвате объектов из других кластеров, либо вследствие пополнения обучающей выборки в целом. Убыль же возможна по причине перехода части элементов в другие кластеры, а также в случае естественной или принудительной ликвидации каких-либо объектов. Если мощность множества элементов членов кластерной последовательности сохраняется на постоянном уровне, или имеет тенденцию к росту, это является свидетельством положительной эволюционной динамики, а кластер имеет ясные перспективы дальнейшего существования. Обратная тенденция говорит о деградации кластера вплоть до возможности его исчезновения, поэтому прогноз его эволюционных изменений неопределён и бесполезен.

Проиллюстрируем возможности описанного выше механизма и особенности его применения на демонстрационном примере. Рассмотрим ход эволюционных изменений трех кластерных сфероидов, заданных в двумерном пространстве классифицирующих признаков x и y (рисунок). Значения признаков, определяющих положение центров кластеров, в равноотстоящие друг от друга моменты времени предполагаются известными, так же как количество элементов, имеющихся в составе кластеров. Область изменения признаков приведена к размеру 100x100 условных единиц. Координаты кластерных центров, зафиксированные в моменты времени $t_i\ i=\overline{1,n}$, а также мощности множеств элементов, входящих в состав кластеров, представлены в табл. 1.

Анализ эволюционного поведения начнем с вычисления коэффициентов автокорреляции первого и второго порядка для рядов обоих классифицирующих признаков во всех трех кластерных последовательностях по формуле

$$r_z^k = \frac{\sum_{t=k+1}^{n}(z_t-\overline{z}_t)(z_{t-k}-\overline{z}_{t-k})}{\sqrt{\sum_{t=k+1}^{n}(z_t-\overline{z}_t)^2 . \sum_{t=k+1}^{n}(z_{t-k}-\overline{z}_{t-k})^2}}, \quad (1)$$

где k – порядок коэффициента автокорреляции, z – текущее значение классифицирующего признака, $\overline{z}$ – среднее значение классифицирующего признака и оно равняется общему количеству значений признаков, деленному на число наблюдений. Результаты вычислений сведены в табл. 2.

Таблица 1

Эволюционные изменения кластеров

время	Кластер № 1			Кластер № 2			Кластер № 3		
t	Признак x	Признак y	Мощность N	Признак x	Признак y	Мощность N	Признак x	Признак y	Мощность N
1	26	54	440	21	37	310	35	15	390
2	33	57	457	24	38	292	42	26	415
3	41	60	460	26	37	267	47	17	403
4	46	62	474	27	38	231	53	31	394
5	52	65	470	29	36	206	58	15	410
6	59	68	490	30	36	181	62	29	398
7	65	71	492	32	36	168	67	16	406
8	70	73	505	35	37	153	74	28	412

Таблица 2

Коэффициенты автокорреляции динамических рядов

Кластер № 1				Кластер № 2				Кластер № 3			
r_x^1	r_x^2	r_y^1	r_y^2	r_x^1	r_x^2	r_y^1	r_y^2	r_x^1	r_x^2	r_y^1	r_y^2
0,998	0,995	0,997	0,991	0,976	0,951	0,176	0,091	0,991	0,990	–0,793	0,931

Таблица 3

Расчет остатков регрессионных моделей

Время	Наблюдаемые признаки		Вычисленные признаки		Остатки	
t	x	y	$\hat{x}$	$\hat{y}$	$\Delta x = x - \hat{x}$	$\Delta y = y - \hat{y}$
1	26	54	27	54,17	–1	–0,17
2	33	57	33,29	56,91	–0,29	0,09
3	41	60	39,58	59,65	1,42	0,35
4	46	62	45,87	62,39	0,13	–0,39
5	52	65	52,16	65,13	–0,16	–0,16
6	59	68	58,45	67,87	0,55	0,13
7	65	71	64,74	70,61	0,26	0,39
8	70	73	71,03	73,35	–1,03	–0,35

Из материалов этой таблицы видно, что в кластерной последовательности № 1 на ход эволюции преобладающее влияние по обоим признакам оказывает возрастающая трендовая составляющая. Однонаправленность трендов совершенно очевидна, и потому вопрос о том, является ли это проявлением сущностной связи между признаками или просто случайностью, представляется вполне актуальным. Для отыскания ответа на этот вопрос используем хорошо известный метод отклонения от тренда. Для рядов обоих признаков кластерной последовательности № 1 построим линейные регрессионные зависимости от времени. Используя процедуру метода наименьших квадратов, получим для признака x

$$\hat{x} = 20,71 + 6,29t, \qquad (2)$$

а для признака y

$$\hat{y} = 51,43 + 2,74t. \qquad (3)$$

Проверка по критерию Фишера показала адекватность обоих уравнений регрессии, что является естественным для столь высоких значений коэффициентов автокорреляции первых двух порядков по каждому признаку первой последовательности кластеров. Далее рассчитываем значения признаков $\hat{x}$ и $\hat{y}$ в точках $t_i\, i = \overline{1,8}$ с помощью уравнений (2) и (3) и отыскиваем остатки. Результаты вычислений приведены в табл. 3.

Находя значения коэффициентов автокорреляции первого порядка для остатков Δx и Δy, видим, что их величины $r_{\Delta x}^1 = 0,08$ и $r_{\Delta y}^1 = 0,304$ значительно меньше, нежели аналогичные коэффициенты r_x^1 и r_y^1 (табл. 2,кластер № 1). Это позволяет утверждать, что влияние трендов классифицирующих признаков x и y, ясно проявляющееся в кластерной последовательности № 1, полностью устранено. Вместе с тем для коэффициентов парной линейной корреляции r_{xy} и $r_{\Delta x \Delta y}$ уровень значений сохраняется $r_{xy} = 0,996$; $r_{\Delta x \Delta y} = 0,768$. Некоторое снижение, конечно, есть, однако проверка по критерию Стьюдента

$$t = \frac{r_{\Delta x \Delta y}}{\sqrt{1 - r_{\Delta x \Delta y}^2}} \sqrt{n-2} \qquad (4)$$

показывает, что величина коэффициента $r_{\Delta x \Delta y}$ сохраняет статистическую значимость $t = 2,45 > t_{kp} = 2,26$. Отсюда следует, что совпадение трендовых составляющих признаков x и y, описывающих эволюционные изменения центров первой кластерной последовательности, является не случайным, но есть проявление некоторой сущностной связи между ними. Это, в свою очередь, дает веские основания предполагать, что наблюдаемая связь между трендами может быть экстраполирована для прогнозирования последующих этапов эволюции кластера № 1. В целом, с учетом устойчивого возрастания мощности множества элементов, наполняющих этот кластер, следует оценить перспективы его дальнейшего развития как благоприятные.

Эволюционное поведение кластера № 2 практически полностью определяется влиянием ярко выраженной трендовой составляющей признака x. Другой классифицирующий признак на изменение поло-

жения центров кластерной последовательности не оказывает практически никакого воздействия, ибо в его ряду преобладает случайная составляющая (табл. 2). Величины коэффициентов автокорреляции r_y^1 и r_y^2 во второй кластерной последовательности очень малы, что позволяет считать влияние признака y на эволюционные изменения кластера № 2 ничтожными и воспринимать производимый им эффект как случайный шумовой фон. Следует отметить монотонное и довольно значительное падение мощности кластерообразующих множеств в этой последовательности. Это указывает на то, что кластер № 2 деградирует, перспективы его дальнейшего существования весьма неопределенны, а какие бы то ни было прогнозные предсказания сомнительны.

Эволюционное поведение кластера № 3, как видно из рисунка и табл. 2, определяется совокупным влиянием сильного возрастающего тренда признака x и отчетливо выраженной циклической составляющей динамического ряда признака y с периодом, равным двум лаговым единицам. Подобные явления достаточно типичны и могут наблюдаться, например, при изучении циклов солнечной активности, сезонных колебаний цен на некоторые виды товаров и т.п. Как отмечалось выше, анализ и прогнозирование подобных процессов требует применения специальных приемов, смысл которых состоит в оценке влияния циклической составляющей по каждому элементу цикла в отдельности. Для достижения этой цели осуществляется сглаживание уровней временного ряда с помощью скользящих средних по формуле

$$\tilde{y}_{i+\sum_{i=k}^{k+L} i/L} = \frac{\sum_{i=k}^{k+L} y_i}{L}, k = \overline{1, n-L}, \qquad (5)$$

где L – длина периода циклической составляющей, выраженная в единицах временного лага.

Если число элементов периода L четно, то полученным значениям скользящих средних нельзя поставить в соответствие имеющиеся временные метки. Для устранения этого затруднения вычисляются центрированные скользящие средние $\tilde{y}^*$, как средние арифметические соседних значений скользящих средних. L элементов в результате этой операции теряются, однако впоследствии они будут восстановлены.

Принимая гипотезу об аддитивной структуре ряда признака y, находим оценку циклической составляющей для всех оставшихся уровней.

$$S_i = y_i - \tilde{y}_i^*. \qquad (6)$$

Поскольку в рассматриваемом примере длина цикла составляет две единицы временного лага, находим средние значения циклической составляющей для четных и для нечетных моментов времени $\overline{S}_{\text{чет}} = 6{,}417$, $\overline{S}_{\text{нечет}} = -6{,}5$. Величина поправки K для аддитивной модели ряда будет

$$K = \frac{\overline{S}_{\text{чет}} + \overline{S}_{\text{нечет}}}{2} = -0{,}0415. \qquad (7)$$

Таким образом, окончательные величины циклических компонент для четных и нечетных моментов времени будут $S_{\text{чет}} = \overline{S}_{\text{чет}} - K = 6{,}459$, $S_{\text{нечет}} = \overline{S}_{\text{нечет}} - K = -6{,}459$ согласно гипотезы аддитивности

$$y = T + S + E, \qquad (8)$$

поэтому ряд признака y с элиминированной циклической составляющей y^* для третьей кластерной последовательности определится из соотношения

$$y^* = y - S. \qquad (9)$$

Таблица 4

Расчет сглаженного ряда признака y для кластера № 3

Время t	y_i	Скользящее среднее $\tilde{y}_i$	Центрированное скользящее среднее $\tilde{y}_i^*$	Циклическая составляющая S_i	$y_i^* = y_i - S_i$
1	15	20,5 21,5			21,459
2	26	24	21	5	19,541
3	17	23	22,75	–5,75	23,459
4	31	22	23,5	7,5	24,541
5	15	22,5	22,5	–7,5	21,459
6	29	22	22,25	6,75	22,541
7	16		22,25	–6,25	22,459
8	28				21,541

Результаты вычислений представлены в табл. 4.

Величина коэффициента автокорреляции первого порядка, вычисленная для ряда y^*, $r^1_{y^*} = -0{,}033$, указывает на полное отсутствие в этом ряду, а значит, и в ряду признака y трендовой составляющей. Следовательно, в ход эволюционных процессов кластера № 3 классифицирующий признак y вносит только регулярные циклические изменения. В целом, при сохранении условий внешней среды, перспективы дальнейшего существования кластера № 3 можно считать благоприятными. Мощность множества элементов, входящих в его состав, сохраняется на постоянном уровне с незначительными нерегулярными колебаниями, что говорит об отсутствии признаков деградации.

Выводы

1. Построена математическая модель эволюционных изменений саморазвивающихся кластерных образований.

2. С помощью аппарата динамических рядов показана возможность прогнозирования хода их развития.

3. Некоторые возможности предлагаемой модели продемонстрированы на иллюстративном примере.

Список литературы

1. Игнатьев Н.А. Кластерный анализ данных и выбор объектов-эталонов в задачах распознавания с учителем / Н.А. Игнатьев // Вычислительные технологии. – 2015. – Т. 20, № 6. – С. 36–45.

2. Ahmad T. et al. Clinical Implications of Chronic Heart Failure Phenotypes Defined by Cluster Analysis. J. Am. Coll. Cardiol. 2014, 64, P. 1765–1774.

3. Ximing Lv. Research on P2P Network Loan Risk Evaluation Based on Generalized DEA Model and R-Type Clustering Analysis under the Background of Big Data / Lv. Ximing, Lan Zhou, Xiaona Guo // Financial Risk Management, 2017, Vol.6, No. 2, P. 163–190.

4. Печеный Е.А. Математическая модель динамической кластеризации в задачах распознавания образов / Е.А. Печеный, А.А. Аль-Хашеди, Н.К. Нуриев // Современные наукоемкие технологии. – 2018. – № 5. – С. 124–130.

5. Елисеева И.И. Эконометрика: учебник. – М.: Финансы и статистика, 2001. – 344 с.

6. Валеев Н.Н. Анализ временных рядов и прогнозирование: учебное пособие / Н.Н. Валеев, А.В. Аксянова, Г.А. Гадельшина. – Казань: Изд-во Казан. гос. технол. ун-та, 2010. – 160 с.

7. Ярушкина Н.Г. Интеллектуальный анализ временных рядов: учебное пособие / Н.Г. Ярушкина, Т.В. Афанасьева, И.Г. Перфильева. – Ульяновск: УлГТУ, 2010. – 320 с.

8. Ruey S. Tsay. Multivariate Time Series Analysis: With R and Financial Applications. John Wiley & Sons, 2013, 520 p.