8С

УДК 519.711:004.8:681.3

Л.В. Сарычева
Национальный горный университет, Днепропетровск
Sarycheva@prognoz.dp.ua

Пространственно-временной подход в задачах кластеризации

Рассмотрена задача кластерного анализа данных, имеющих пространственную и временную привязку. Предложен новый алгоритм кластеризации GeoTime, учитывающий временное соседство (на основе индуктивного подхода) и топологические свойства месторасположений объектов (на основе геоинформационного подхода). На реальных данных эколого-социально-экономического мониторинга государств Европы проведено экспериментальное сравнение GeoTime с известными алгоритмами по трем критериям качества кластеризации.

Введение

Класс методов и алгоритмов кластерного анализа обширен – алгоритмы иерархической кластеризации, *k*-средних, ISODATA, ОКК и др. [1-4]. В каждом конкретном случае применяются методы, учитывающие особенности исходных данных – объем выборок, число признаков, априорную информацию и т.д.

Для многих практических задач исходные данные носят пространственно-временной характер. Например, в задаче кластерного анализа показателей эколого-социально-экономического (ЭСЭ) мониторинга регионов обычно используются таблицы, содержащие наряду с географическими координатами значения нескольких ЭСЭ-показателей за определенное число месяцев (лет).

Несмотря на существование множества алгоритмов кластеризации, большинство из них не учитывает в явном виде пространственную привязку (свойство географического соседства) и временные свойства исследуемых объектов (результаты кластер-анализа для отдельных временных срезов и всей совокупности данных не должны быть противоречивыми). В известных алгоритмах географические координаты объектов если используются, то равноправно с другими временными свойствами, поэтому с содержательной точки зрения интерпретация результатов кластеризации затруднительна и неоднозначна.

Наглядное представление кластеров в пространственных координатах объектов помогает оценить реальность полученной модели. Например, в задачах районирования территории по данным ЭСЭ-мониторинга анализ геоиконических моделей кластеризации в геоинформационной системе (ГИС) в виде ранговых картограмм, окрашивающих на карте однородные группы объектов в определенный цвет, позволяет избежать появления «маломощных» кластеров, не характерных для определенного региона [5].

Следовательно, актуальной является задача создания алгоритмов кластеризации, оперирующих географическими признаками (координатами объектов) и признаками, отражающими временные свойства, не объединяя их равноправно в одной таблице «объект-свойство». Перспективность таких алгоритмов обусловлена учетом топологических свойств месторасположений объектов и возможностью содержательной интерпретации результатов кластеризации.

Цель работы. изложение нового алгоритма GeoTime-кластеризации, учитывающей географическое соседство объектов (на основе геоинформационного подхода, т.е. рассмотрения топологических свойств месторасположений регионов) и временное соседство их показателей (на основе индуктивного подхода).

Постановка задачи

Содержательная постановка задачи (на примере анализа данных ЭСЭ-мониторинга). Объектами анализа являются *n* регионов территории, которые характеризуются своим месторасположением (определяемым двумя или тремя географическими координатами) и *m* ЭСЭ-показателями (признаками), измеренными в последовательные моменты времени.

Требуется провести районирование территории по совокупности ЭСЭ-показателей мониторинга.

Математическая постановка задачи. Пусть $x_{ij}(t_s)$ – измерения признаков, характеризующих заданное множество объектов $Z=\{Z_1, Z_2,..., Z_n\}$ в момент времени t_s ($i=1, 2,..., n$ – номер наблюдения, n – число наблюдений, $j=1, 2,..., m$ – номер признака, m – число признаков, $s=1, 2,..., L$ – номер момента времени); $Q(Z_1), Q(Z_2),..., Q(Z_n)$ – географические координаты объектов. Исходные данные представляют собой блочную матрицу

$$(\boldsymbol{Q} \mathbb{N} \boldsymbol{X}(t_1) \mathbb{N} \boldsymbol{X}(t_2) \mathbb{N} \ldots \mathbb{N} \boldsymbol{X}(t_L)),$$

где $\boldsymbol{Q}$ – матрица размером $n\times2$ (или $n\times3$) географических координат объектов, $\boldsymbol{X}(t_s)=(\boldsymbol{X}^1(t_s) \mathbb{N} \boldsymbol{X}^2(t_s) \mathbb{N} \ldots \mathbb{N} \boldsymbol{X}^m(t_s))$ – матрица типа «объект-свойство»,
$\boldsymbol{X}^j(t_s)=(x_{1j}(t_s),x_{2j}(t_s),\ldots,x_{nj}(t_s))^T$ – вектор-столбец значений j-го признака для n объектов,
$\boldsymbol{X}_i(t_s)=(x_{i1}(t_s),x_{i2}(t_s),\ldots,x_{im}(t_s))$ – вектор-строка значений m признаков i-го объекта, $j=1,2,\ldots,m;\ i=1,2,\ldots,n;\ s=1,2,\ldots, L$.

Определение 1. Кластеризацией $K=\{K_1, K_2,\ldots,K_k\}$, $1\le k\le n$, множества Z называется семейство непустых, попарно непересекающихся подмножеств (кластеров) K_q, $q=1, 2,.., k$, множества Z, объединение которых совпадает с Z:

$$K_1 \cup K_2 \cup \ldots \cup K_k = Z, \quad K_i \cap K_j = \varnothing,\ i \neq j,\ i,j=1,2,..,k,\ K_q \neq \varnothing,\ q=1,2,..,k.$$

Определение 2. Наилучшей называется кластеризация $K^* \subseteq \Phi$, для которой

$$K^* = \arg\max_{K \subseteq \Phi} J(K) \quad (\text{или } K^* = \arg\min_{K \subseteq \Phi} J(K)),$$

где Φ – множество всех допустимых разбиений (кластеризаций) заданного множества Z; $J(K)$ – критерий качества кластеризации.

Требуется решить задачу поиска наилучшей кластеризации K^*.

Для решения задачи кластеризации необходимо:

а) дать определение кластера – указать свойства, общие для всех объектов отдельного кластера (меру сходства между объектами);

б) задать способ образования кластеров (сортировка, перегруппировка, объединение, разбиение, добавление, поиск);

в) указать критерий J качества кластеризации ($J(K)=J(\alpha(K), \beta(K), \gamma(K))$; $\alpha(K)$, $\beta(K)$, $\gamma(K)$ – критерии: точностной, соседства, непротиворечивости соответственно);

г) организовать движение к максимуму (минимуму) критерия J (при этом определяется и число реально существующих кластеров).

Алгоритм GeoTime кластеризации пространственно-временных данных

Предлагаемый метод и соответствующий алгоритм GeoTime кластеризации пространственно-временных данных основываются на следующих предположениях.

Предположение 1 (учет временного соседства). Объекты, близкие по своим свойствам (входящие в один кластер) в момент времени *t*, могут быть близки и в момент времени *(t+1)* (рис. 1). Это предположение используется для нахождения кластеризации *K(t,t+1)*, оптимизирующей критерий $\gamma(K)$ непротиворечивости (например, $\gamma(K)\rightarrow min$, где $\gamma(K)$ – мера сходства между кластеризациями $K(t_s)$ для отдельных временных срезов $t=t_s$, $s=1, 2,..., L$, и всей совокупности данных [6]).

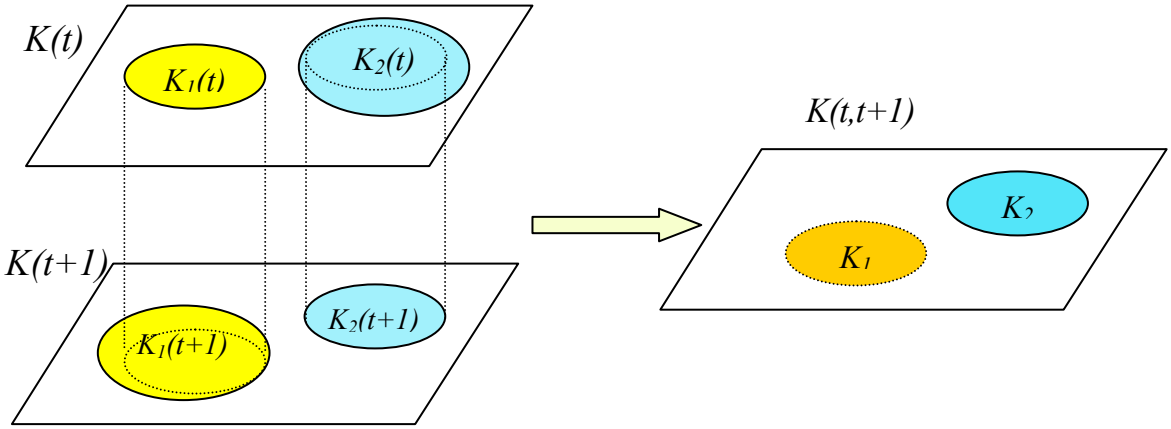


Рисунок 1 – Преемственность кластеризаций *K(t)* и *K(t+1)* в моменты времени *t* и *(t+1)*

Предположение 2 (учет географической смежности). Объекты, соседствующие в географическом пространстве, могут образовать однородные по свойствам группировки, соответствующие связным областям (целостным территориальным зонам). Это предположение используется для нахождения кластеризации, оптимизирующей критерий $\beta(K)$ (например, $\beta(K)\rightarrow max$; $\beta(K)$ – сумма элементов в матрице смежности объектов отдельных кластеров, рассматриваемых в пространстве географических координат).

Предположение 3 (учет близости в пространстве признаков). Объекты, соседствующие в пространстве признаков *X*, могут образовать однородные по свойствам группировки – кластеры. Это распространенное для большинства методов кластерного анализа предположение используется для нахождения кластеризации, оптимизирующей критерий $\alpha(K)$ (например, $\alpha(K)\rightarrow min$; $\alpha(K)$ – сумма внутрикластерных расстояний, где *K* – кластеризация множества *Z* без учета ***Q***).

Предположения 1 – 3 обусловливают, что предлагаемый алгоритм реализует метод разделения заданного множества объектов на группы, состоящие из взаимосвязанных, однородных объектов. Здесь имеется в виду тип внутренней однородности: совокупность считается однородной, если в зафиксированном пространстве признаков закономерность формирования объектов приводит к тому, что объекты близки друг к другу в этом пространстве [7].

Рассмотрим более подробно вопрос о преемственности кластеризаций, соответствующих двум последовательным моментам времени.

Пусть $K(t_s)=\{K_1(t_s), K_2(t_s),...,K_k(t_s)\}$ – кластеризация объектов в момент времени $t=t_s$, $s=1,2,...,L$, где *k* – число кластеров, $1<k<n$ ($k=1$, $k=n$ здесь не рассматриваются):

$\bigcup_{i=1}^{k} K_i(t_s) = X(t_s)$, $K_i(t_s) \neq \varnothing$, $K_i(t_s) \cap K_j(t_s) = \varnothing$, при $i \neq j$; $i, j=1, 2, \ldots, k$.

Кластеризация $K(t_s)$ получена на основе признаков, характеризующих момент времени $t=t_s$, т.е. с использованием матрицы

$$\boldsymbol{X}(t_s)=(\boldsymbol{X_1}(t_s) \boldsymbol{\cup} \boldsymbol{X_2}(t_s) \boldsymbol{\cup} \ldots \boldsymbol{\cup} \boldsymbol{X_m}(t_s))=\begin{pmatrix} x_{11}(t_s) & \ldots & x_{1m}(t_s) \\ \ldots & \ldots & \ldots \\ x_{n1}(t_s) & \ldots & x_{nm}(t_s) \end{pmatrix}.$$

Каждому объекту Z_i (отождествленному с точкой $(x_{i1}(t_s), x_{i2}(t_s), \ldots, x_{im}(t_s))$, $i=1,2,\ldots,n$, евклидова пространства R^m) кластеризация $K(t_s)$ ставит в соответствие номер кластера, к которому он принадлежит в момент времени t_s.

По результатам кластеризаций $K(t_s)$ и $K(t_{s+1})$ строится матрица $\boldsymbol{A}(t_s,t_{s+1})=(a_{ij}(t_s,t_{s+1}))$, $i, j=1, 2, \ldots, k$, где элемент $a_{ij}(t_s,t_{s+1})$ определяет число объектов, одновременно входящих в кластер $K_i(t_s)$ и кластер $K_j(t_{s+1})$ (табл. 1):

Таблица 1

$K(t_s)$	$K(t_{s+1})$					
	$K_1(t_{s+1})$	$K_2(t_{s+1})$	…	$K_j(t_{s+1})$	…	$K_k(t_{s+1})$
$K_1(t_s)$	a_{11}	a_{12}	…	a_{1j}	…	a_{1q}
$K_2(t_s)$	a_{21}	a_{22}	…	a_{2j}	…	a_{2k}
…	…	…	…	…	…	…
$K_i(t_s)$	a_{i1}	a_{i2}	…	a_{ij}	…	a_{ik}
…	…	…	…	…	…	…
$K_k(t_s)$	a_{k1}	a_{k2}	…	a_{k3}	…	a_{kk}

Элементы $a_{ij}(t_s,t_{s+1})$ матрицы $\boldsymbol{A}(t_s, t_{s+1})$ подсчитываются следующим образом:

$$a_{ij}(t_s, t_{s+1}) = \sum_{\substack{e:(X_e \in K_i(t_s)) \wedge \\ \wedge (X_e \in K_j(t_{s+1}))=1 \\ e \in \{1,2,\ldots,n\}}} 1, \quad i, j=1, 2, \ldots, k. \tag{1}$$

Сумма элементов матрицы $\boldsymbol{A}(t_s, t_{s+1})$ равна числу кластеризуемых объектов:

$$\sum_{i=1}^{k} \sum_{j=1}^{k} a_{ij}(t_s, t_{s+1}) = n, \quad \forall s=1,2,\ldots, L-1.$$

Номер кластера является не более чем меткой, то есть можно поменять нумерацию кластеров, сохранив при этом состав входящих в них элементов. Поэтому для определения преемственности кластеризации $K(t_{s+1})$ по отношению к $K(t_s)$ в матрице $\boldsymbol{A}(t_s, t_{s+1})$ производится перестановка столбцов таким образом, чтобы максимальный элемент i-й строки попал на главную диагональ:

$$K_i(t_{s+1}) \leftrightarrow \max_i a_{ii}(t_s,t_{s+1}).$$

Обозначим $\boldsymbol{P}(t_s, t_{s+1})=\frac{1}{n}\boldsymbol{A}(t_s, t_{s+1})=(p_{ij}.(t_s,t_{s+1}))$, $p_{ij}(t_s,t_{s+1})=a_{ij}(t_s,t_{s+1})/n$, $i, j=1,2,\ldots,k$.

Сумма элементов матрицы $\boldsymbol{P}(t_s, t_{s+1})$ равна единице:

$$\sum_{i=1}^{k} \sum_{j=1}^{k} p_{ij}(t_s, t_{s+1}) = 1, \quad \forall s=1,2,\ldots, L-1.$$

Элемент $p_{ij}(t_s,t_{s+1})$ можно рассматривать как вероятность того, что объект, принадлежавший кластеру $K_i(t_s)$, попадет в кластер $K_j(t_{s+1})$.

Заметим, что в построенной таким образом последовательности матриц $\{\boldsymbol{A}(t_s, t_{s+1})\}$, $s=1,2,...,L-1$, выполняются следующие отношения:

$$\forall i \in \{1,2,...,k\} \quad a_{ii}(t_s, t_{s+1}) \neq 0, \; a_{ii}(t_{s-1}, t_s) \geq a_{ii}(t_s, t_{s+1}), \; s=2,...,L-1, \tag{2}$$

$$\{ j_1^L, j_2^L, ..., j_{n_i(L)}^L \} \subset \{ j_1^{L-1}, j_2^{L-1}, ..., j_{n_i(L-1)}^{L-1} \} \subset ... \subset \{ j_1^1, j_2^1, ..., j_{n_i(1)}^1 \}, \tag{3}$$

где $j_1^s, j_2^s, ..., j_{n_i(s)}^s$ – номера объектов, вошедших в кластер $K_i(t_s)$;

$n_i(s) = |K_i(t_s)|$ – число объектов в $K_i(t_s)$, $\sum_{i=1}^{k} n_i(s) = n$, $s=1,2,...,L$.

Отношение (3) показывает, что для всякого $i \in \{1,2,...,k\}$ последовательности номеров $j_1^s, j_2^s, ..., j_{n_i(s)}^s$ объектов кластеров $K_i(t_s)$ $s=1,2,...,L$, являются вложенными.

Теорема. Для всякого $i \in \{1,2,...,k\}$ существует хотя бы одна точка, принадлежащая всем кластерам $K_i(t_s)$, $s=1,2,...,L$, одновременно, то есть:

$$\{ j_1^L, j_2^L, ..., j_{n_i(L)}^L \} \cap \{ j_1^{L-1}, j_2^{L-1}, ..., j_{n_i(L-1)}^{L-1} \} \cap ... \cap \{ j_1^1, j_2^1, ..., j_{n_i(1)}^1 \} \neq \varnothing .$$

Доказательство проводится методом от противного.

Из теоремы следует, что существует k точек $X_{i_1}, X_{i_2}, ..., X_{i_k}$:

$$X_{i_1} \in K_1(t_1) \cap K_1(t_2) \cap ... \cap K_1(t_L),$$
$$X_{i_2} \in K_2(t_1) \cap K_2(t_2) \cap ... \cap K_2(t_L),$$
$$.....$$
$$X_{i_k} \in K_k(t_1) \cap K_k(t_2) \cap ... \cap K_k(t_L).$$

При поиске наилучшей кластеризации в методе GeoTime такие точки принимаются за начальные центры (ядра) кластеров (отождествляя объект Z_i с точкой $X_i=(x_{i1}(t_1), x_{i2}(t_1), ..., x_{im}(t_1), x_{i1}(t_2), x_{i2}(t_2), ..., x_{im}(t_2), ..., x_{i1}(t_L), x_{i2}(t_L), ..., x_{im}(t_L)) \in R^p$, $p=mL$).

Меры близости между двумя объектами, между объектом и кластером, между двумя кластерами, применяемые в GeoTime-кластеризации, представлены в табл. 1 [6], [8]. Для оценки близости между двумя различными кластеризациями K и Q конечного множества объектов используется мера близости

$$d(K,Q) = \frac{\frac{1}{2}\left(\sum_{i=1}^{k_1} |K_i|^2 + \sum_{i=1}^{k_2} |Q_i|^2\right) - \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} |K_i \cap Q_j|^2}{\frac{1}{2}\left(\sum_{i=1}^{k_1} |K_i|^2 + \sum_{i=1}^{k_2} |Q_i|^2\right)}, \tag{4}$$

где k_1, k_2 – число кластеров (подмножеств исходного множества) в кластеризациях K и Q соответственно; $|K_i|, |Q_j|, i = 1,2,...,k_1; \; j = 1,2,...,k_2$ – мощности соответствующих подмножеств, т.е. число элементов в кластерах.

Величина $d(K,Q)$ принимает значения от *0* до *1*:

0 – при полностью совпадающих разбиениях в кластеризациях K и Q,

1 – при полностью несовпадающих, когда $\sum_{i=1}^{k_1} \sum_{j=1}^{k_2} |K_i \cap Q_j|^2 = 0$.

Предположение 2 обусловливает применение геоматического алгоритма, общая схема которого изложена в работе [9].

GeoTime включает в себя основные шаги геоматического алгоритма и усложняет его процедурой выделения (на основе преемственности кластеризаций в последовательные моменты времени) ядер – кластеров, образующих целостный географический район.

Общая схема алгоритма GeoTime следующая.

1. Найти матрицу смежности $G=(g_{ij}),\ i, j=1,2,\ldots,n,$ между объектами $Z_1, Z_2,\ldots, Z_n$ в географическом пространстве; g_{ij} определяется одним из способов:

$$\text{а) } g_{ij}=\begin{cases} 1, \text{ если } \rho_{ij} < c, \\ 0 - \text{в противном случае,} \end{cases}, \qquad \text{б) } g_{ij}=\begin{cases} 1, \text{ если } Z_i \text{ и } Z_j - \text{соседи,} \\ 0 - \text{в противном случае,} \end{cases}$$

где ρ_{ij} – евклидово расстояние между объектами Z_i и Z_j в географическом пространстве, ($\boldsymbol{Q}$ – исходная матрица для его вычисления), c – порог.

2. Вычислить матрицу $D=(d_{ij}),\ i, j=1,2,\ldots,n,$ где $d_{ij}=d(Z_i,Z_j)$ – мера сходства между Z_i и Z_j в пространстве признаков (табл. 1) ($\boldsymbol{X}=(\boldsymbol{X}(t_1)\boldsymbol{\cup}\boldsymbol{X}(t_2)\boldsymbol{\cup}\ldots\boldsymbol{\cup}\boldsymbol{X}(t_L))$ – исходная матрица для вычисления d_{ij}). Для каждого момента времени $t_s,\ s=1,2,\ldots,L,$ вычислить матрицу $D(t_s)=(d_{ij})_{/ts}$, аналогичную D (используя $\boldsymbol{X}(t_s)$).

3. Определить центры кластеров начального разбиения (ядра) $X_{i_1}, X_{i_2},\ldots, X_{i_k}$, используя изложенный алгоритм определения преемственности кластеризаций в последовательные моменты времени. Принять

$$K_1^1=\{X_{i_1}\},\ K_2^1=\{X_{i_2}\},\ldots,\ K_k^1=\{X_{i_k}\}.$$

4. Положить $r=1$.

5. Пусть на шаге $r\in\{1,\ldots,n-k\}$ получены классы $K_1^r,\ldots,K_k^r$.

6. Найти $j\in\{1,\ldots,n\},\ v\in\{1,\ldots,k\}$:

$$d_{ji_v}=\min\{d_{ji_q}\,|\,\exists X_m\in K_q : g_{jm}=1,\ q\in\{1,2,\ldots,k\}\}.$$

7. Положить $\forall i\in\{1,\ldots,k\}\quad K_i^{r+1}=\begin{cases} K_i^r\cup\{X_j\},\ i=v; \\ K_i^r,\ i\neq v. \end{cases}$

(На этом шаге учитывается предположение 2).

8. Положить $d_{ji_v}=\infty$.

9. Если $r=n-k,$ конец, иначе положить $r=r+1$ и перейти к шагу 5.

Таким образом, выделяя в каждом кластере несколько регионов, образующих целостный географический район (пространственный объем), полагают их ядрами. Выделенные ядра кластеров расширяются путем доклассификации оставшихся объектов по приведенному выше алгоритму.

Преимущества предлагаемой кластеризации:

- учет топологических свойств месторасположений объектов для выделения целостных географических зон, образуемых отдельными кластерами;
- учет временных тенденций изменения показателей для выделения начальных ядер кластеров;
- возможность содержательной интерпретации выделенных кластеров.

Экспериментальное сравнение GeoTime с известными алгоритмами проводилось на доступных реальных данных ЭСЭ-мониторинга, опубликованных на сайте [10]. В качестве исходных признаков выступали шестнадцать ЭСЭ-показателей двадцати четырех государств Европы, временной период составил девять лет (1996 – 2004 гг.):

1) экономические показатели – деловые инвестиции, % от ВВП (Business investment, % of GDP); потребительские расходы, индекс по 1995 г. (Consumption expenditure at constant prices, index 1995 = 100); ВВП на душу населения, индекс от EU-25 (GDP per capita in PPS, index EU-25 = 100); общий государственный долг, % от ВВП (General goverment gross debt, % of GDP); уровень инфляции (Inflation rate), %; сетевые инвестиции в общественный сектор, % от ВВП (Net saving: Public sector, % of GDP); общественные инвестиции, % от ВВП (Public investment, % of GDP); темпы роста ВВП (Real GDP growth rate), %; общие инвестиции, % от ВВП (Total investment, % of GDP); прирост стоимости единицы труда (Unit labour cost growth: total economy), %;

2) социальные показатели – уровень занятости (Total employment growth), % ; прирост уровня занятости (Total employment rate), %; потребление электроэнергии, индекс от 1995 г. (Electricity consumption by households, Index 1995 = 100);

3) экологические показатели – выбросы CO_2 на душу населения, т (CO_2 emissions per capita, tonnes); показатель повреждения лесов дефолиацией (Forest trees damaged by defoliation), %; муниципальные загрязнения, кг на душу населения (Municipal waste generated, kg per capita).

В табл. 2 представлены некоторые результаты экспериментального сравнения GeoTime с известными алгоритмами кластеризации (число кластеров $k=8$, признаков $m=16$, объектов $n=24$).

В качестве меры сходства между объектами в 16-мерном признаковом пространстве здесь применялся угол между векторами X_i и X_j:

$$d(X_i, X_j)=arccos((X_i \cdot X_j)/(|X_i| \cdot |X_j|)).$$

Для сравнения алгоритмов в табл. 2 приведены J_1, J_2, J_3 – значения следующих формальных критериев оценки качества кластеризации [1], [2]:

1) критерий внутрикластерных дисперсий

$$J_1 = \sum_{j=1}^{k} \sum_{X_i \in K_j} d_E^2 \left(X_i, \mu_j\right),$$

где $\mu_j = \frac{1}{n_j} \sum_{X_i \in K_j} X_i$ – центр тяжести кластера K_j; n_j – число объектов в нем;

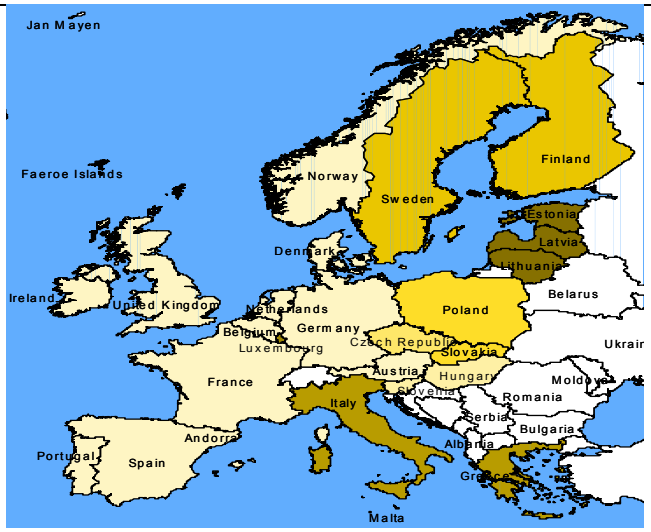
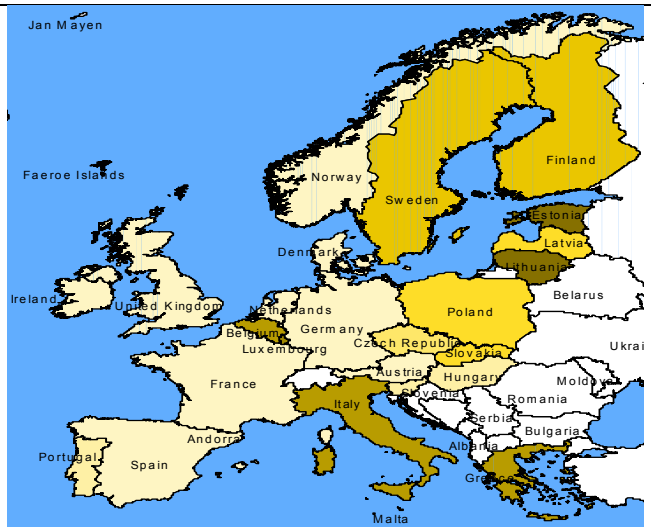
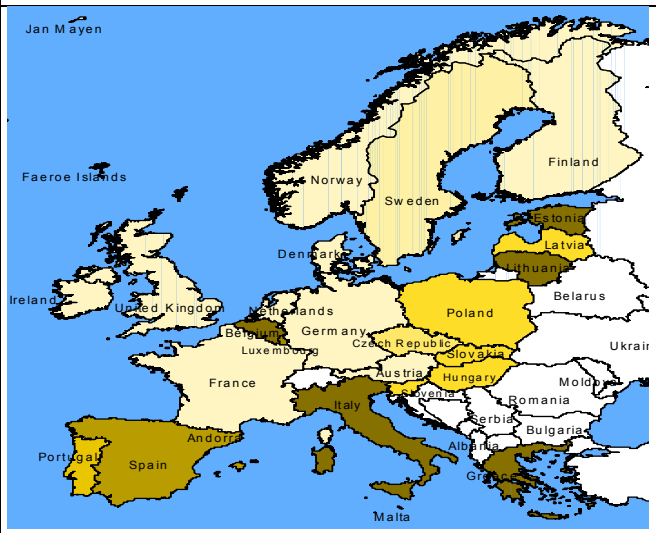
2) критерий попарных внутрикластерных расстояний между объектами:

$$J_2 = \sum_{j=1}^{k} \frac{1}{n_j} \sum_{X_i, X_g \in K_j} d_E^2\left(X_i, X_g\right);$$

3) критерий межкластерного разброса объектов (чем больше величина J_3 $(0<J_3<1)$, тем большая доля общего разброса объектов объясняется межклассовым разбросом и тем лучше качество разбиения):

$$J_3 = 1 - \frac{W}{S},$$

Таблица 2 – Результаты экспериментального сравнения методов кластеризации

Метод	Значения критериев качества	Геоиконическая модель	Регион, номер кластера
Геомати-ческий	J_1=415,8 J_2=171,1 J_3=0,78		Belgium 1 Czech Republic 3 Denmark 1 Germany 1 Estonia 8 Greece 6 Spain 1 France 1 Ireland 1 Italy 6 Latvia 8 Lithuania 8 Luxembourg 7 Hungary 2 Netherlands 1 Austria 1 Poland 4 Portugal 1 Slovenia 2 Slovakia 4 Finland 5 Sweden 5 United Kingdom 1 Norway 1
GeoTime	J_1=415,9 J_2=179,4 J_3=0,81		Belgium 6 Czech Republic 3 Denmark 1 Germany 1 Estonia 8 Greece 6 Spain 1 France 1 Ireland 1 Italy 6 Latvia 4 Lithuania 8 Luxembourg 7 Hungary 2 Netherlands 1 Austria 1 Poland 4 Portugal 2 Slovenia 2 Slovakia 4 Finland 5 Sweden 5 United Kingdom 1 Norway 1
k-средних	J_1=417,3 J_2=178,2 J_3=0,75		Belgium 7 Czech Republic 3 Denmark 1 Germany 1 Estonia 8 Greece 7 Spain 6 France 1 Ireland 1 Italy 7 Latvia 4 Lithuania 8 Luxembourg 7 Hungary 4 Netherlands 1 Austria 1 Poland 4 Portugal 5 Slovenia 4 Slovakia 4 Finland 1 Sweden 2 United Kingdom 1 Norway 1

где $W = \sum_{j=1}^{k} W_j$; $W_j = \sum_{X_i \in K_j} d^2(X_i, \mu_j)$ – внутрикластерный разброс;

$S = \sum_{i=1}^{n} d^2(X_i, \overline{X})$ – общее рассеивание.

Выводы

Разработан новый алгоритм GeoTime для кластеризации пространственно-временных данных, учитывающий географическое соседство объектов (на основе геоинформационного подхода, т.е. рассмотрения топологических свойств месторасположений регионов) и временное соседство их показателей (на основе индуктивного подхода).

Преимущества предлагаемого алгоритма:

1) нацеленность на выделение целостных географических зон, образуемых отдельными кластерами;

2) учет временных тенденций изменения показателей для выделения начальных ядер кластеров;

3) возможность содержательной интерпретации выделенных кластеров.

Многочисленные эксперименты на реальных данных ЭСЭ-мониторинга регионов показали, что геоматическая и GeoTime-кластеризации имеют несколько лучшие показатели известных формальных критериев качества, чем кластеризация, полученная методом *k*-средних (для иерархических методов результаты хуже, чем для *k*-средних).

Перечисленные выше результаты показывают целесообразность применения предлагаемого пространственно-временного подхода к кластеризации регионов по данным ЭСЭ-мониторинга.

Литература

1. Айвазян С.А., Енюков И.С, Мешалкин Л.Д. Прикладная статистика. Исследование зависимостей. – М.: Финансы и статистика, 1985. – 488 с.
2. Классификация и кластер / Под ред. Дж.Вэн Райзина.– М.: Мир, 1980.– 389 с.
3. Ту Дж., Гонсалес Р. Принципы распознавания образов. – М.: Мир, 1978. – 411 с.
4. Дюран Б., Оделл П. Кластерный анализ. – М.: Статистика, 1977. – 128 с.
5. Сарычева Л.В. Компьютерный эколого-социально-экономический мониторинг регионов. Геоинформационное обеспечение: Монография. – Днепропетровск: НГУ, 2003. – 174 с.
6. Сарычева Л.В. Компьютерный эколого-социально-экономический мониторинг регионов. Математическое обеспечение. – Днепропетровск: НГУ, 2003. – 222 с.
7. Мендель И.Д., Миркин Б.Г. Кластер-анализ и смежные вопросы (краткий обзор основных направлений) // Автоматика. – 1987. – № 2. – С. 72-82.
8. Бусыгин Б.С., Мирошниченко Л.В. Распознавание образов при геолого-геофизическом прогнозировании. – ДГУ, 1991. – 168 с.
9. Sarycheva L., Boyko A. Geomatic clusterization based on ecological-social-economical monitoring rates // Scientific Bulletin of NMU. – 2006. – № 5. – P. 76-81.
10. Eurostat // Эл. ресурс. URL: http://www.eurostat.com/

Статья поступила в редакцию 18.07.2006.