УДК 004.8

ГИБРИДНАЯ КЛАСТЕРИЗАЦИЯ НА ОСНОВЕ РЕЛЯЦИОННОЙ СХЕМЫ ИНВАРИАНТНЫХ КЛАСТЕРНЫХ ПРОЦЕДУР

Климова А.С., Батыршин И.З.
Институт проблем информатики Академии наук Республики Татарстан

Поступила в редакцию 10.09.2007, после переработки 29.09.2007.

Дается описание комплекса методов гибридной реляционной кластеризации. Изучаются теоретические свойства схемы реляционных кластерных процедур, обосновывающие методы обобщения этой схемы и построения новых инвариантных процедур кластеризации. Дается описание генетических алгоритмов двух- и трехмерной визуализации результатов реляционной кластеризации. Предлагается метод гибридной кластеризации с одномерной визуализацией данных. Рассматривается применение гибридной кластеризации к анализу статистических временных рядов из области экономики.

A set of methods of hybrid relational clustering is considered. Theoretical properties of the scheme of relational clustering procedures are studied, that ground the methods of generalization of this scheme and construction new invariant clustering procedures. It is given a description of genetic algorithms of two- and three-dimensional visualization of results of relational clustering. A method of hybrid clustering with one-dimensional visualization of data is proposed. It is considered an application of hybrid clustering to analysis of statistical time series from economics.

Ключевые слова: иерархическая кластеризация, инвариантность, нечеткое отношение эквивалентности, визуализация данных, генетический алгоритм, временной ряд.
Keywords: hierarchical clustering, invariance, fuzzy equivalence relation, data visualization, genetic algorithm, time series.

Введение

Кластерные процедуры находят широкие приложения в анализе данных в различных областях человеческой деятельности [1 - 3]. Кластеризация объектов позволяет разбить множество исследуемых объектов на классы сходных объектов и перейти от описания объектов к описанию классов объектов, что является одним из важных этапов формирования знания об исследуемой области, ее анализа и моделирования. Поэтому кластерный анализ часто используется на начальном этапе исследования данных, когда представления о структуре данных еще не сформированы, и классы, выявляемые кластерным анализом, могут служить основой

для дальнейшего более детального анализа данных. Кластерный анализ активно развивается последние десятилетия, что обусловлено следующими причинами: 1) появлением новых областей его приложения со специфическими объектами исследования такими, как Интернет и биоинформатика [4 - 6]; 2) разработкой новых моделей анализа, обработки и представления данных таких, как нечеткая логика, нейронные сети и генетические алгоритмы [7, 8]; 3) разработкой методов гибридной кластеризации, сочетающих результаты разных кластерных процедур, или сочетающие кластерный анализ с другими методами анализа данных [9 -12].

В работе [13] предложена общая схема гибридных реляционных кластерных процедур на основе методов инвариантной реляционной кластеризации и генетической визуализации данных, разработанных авторами. В данной работе изучаются теоретические свойства схемы инвариантных реляционных кластерных процедур, обосновывающие методы обобщения этой схемы и построения новых инвариантных процедур кластеризации. Дается описание генетических алгоритмов двух- и трехмерной визуализации результатов реляционной кластеризации. Предлагается метод гибридной кластеризации данных с одномерной визуализацией по обобщенному признаку. Один из предложенных методов иллюстрируется на примере анализа базы данных статистических временных рядов из области экономики. Статья организована следующим образом. В Разделе 1 даются основные определения для взвешенных (нечетких) отношений сходства и эквивалентности и дается общая схема гибридной реляционной кластеризации и визуализации данных. В Разделе 2 дается описание схемы инвариантных реляционных кластерных алгоритмов. В Разделе 3 исследуются теоретические свойства этой схемы. В Разделе 4 описываются эволюционные процедуры двух- и трехмерной визуализации результатов реляционной кластеризации. В разделе 5 обсуждаются методы гибридной кластеризации с двух- и трехмерной визуализацией кластеров. В разделе 6 кратко описывается метод гибридной одномерной кластеризации временных рядов. В Разделе 7 описывается метод гибридной кластеризации с визуализацией сильных связей. Метод иллюстрируется на примере кластеризации временных рядов из области экономики. В Заключении отмечаются основные результаты статьи.

1. Основные понятия

Пусть $M = \{x, y, \ldots\}$ - множество, содержащее n объектов, I - положительное число и $L = [0, I]$. Функция $S : M \times M \to L$, удовлетворяющая на M условиям симметричности $S(y, x) = S(x, y)$ и рефлексивности $S(y, y) = I$, называется взвешенным (нечетким) отношением сходства [14]. Без ограничения общности будем полагать $I = 1$. Это отношение будет рассматриваться как исходная информация о сходстве между объектами из M, на основе которой кластерная процедура строит разбиение множества объектов на классы сходных объектов. На практике отношение сходства может быть задано непосредственно, например, экспертом, получено с помощью подходящей меры сходства, или на основе расстояния, если объекты заданы векторами в некотором пространстве признаков. В последнем случае мера расстояния будет рассматриваться как отношение различия $D : M \times M \to L$, удовлетворяющее на M условиям симметричности $D(y, x) = D(x, y)$ и антирефлексивности: $D(y, y) = 0$. Функции S и D могут быть получены друг из друга, например, таким образом: $D(x, y) = I - S(x, y)$; $S(x, y) = I - D(x, y)$.

Взвешенным (нечетким) отношением эквивалентности называется отношение сходства, удовлетворяющее на M условию $(\vee, \wedge)$-транзитивности:

$$S(x,z) \geq \min\{(S(x,y), S(y,z)\}. \tag{1}$$

$(\vee, \wedge)$-транзитивность отношения сходства двойственна ультраметрическому неравенству отношения различия:

$$D(x,z) \leq \max\{(D(x,y), D(y,z)\}.$$

Нечеткие отношения эквивалентности и ультраметрики определяют иерархию разбиений множества M на классы. Они изучались в работах [14-18]. Процедура кластеризации рассматривается здесь как преобразование отношение сходства в отношение эквивалентности или, двойственно, как преобразование отношения различия в ультраметрику. Гибридной реляционной кластеризацией множества объектов M называется преобразование [13]

$$HC(S) =< RC; PC >, \tag{2}$$

где S - взвешенное отношение сходства на M, RC - множество четких или нечетких отношений на исследуемом множестве объектов M, содержащее взвешенное отношение эквивалентности E, и PC - множество описаний элементов из M в пространстве размерности m. В общем случае преобразование HC должно удовлетворять некоторым условиям рациональности. Например, отношения из RC или расстояния между элементами множества M, определяемые представлением PC, должны быть согласованы с заданным отношением сходства S. Кластеризация и визуализация являются частными случаями схемы (2). Если $PC = \oslash$ и $RC = \{E\}$, где E - четкое или взвешенное отношение эквивалентности на M, то (2) определяет, соответственно, обычную или иерархическую кластеризацию множества M. Иерархическая кластеризация может быть задана также множеством четких отношений эквивалентности $RC = \{E_1, E_2, \ldots, E_m\}$ таких, что $E_1 \subset E_2 \subset, \ldots, \subset E_m$. Если $RC = \oslash$, а PC задает описание элементов из M в пространстве размерности $m \in \{1,2,3\}$, то HC задает визуализацию данных в пространстве малой размерности.

2. Инвариантная иерархическая реляционная кластеризация

Общая схема иерархических кластерных процедур, удовлетворяющих условиям инвариантности относительно исходной нумерации объектов и инвариантности относительно монотонных преобразований значений сходства исследовалась в работах [19-21]. Схема (2) имеет вид:

$$HC(S) =< E; \oslash > .$$

где E - взвешенное отношение эквивалентности. Преобразование $HC(S)$ задается процедурой кластеризации $Q(S) = E$, которая определяется так:

$$Q(S) = TC(F(S)) = E, \tag{3}$$

где ТС - процедура транзитивного замыкания взвешенных отношений [14- 16], а F - процедура «коррекции» заданного отношения сходства S, такая что $F(S) \subseteq S$.

Процедура транзитивного замыкания удовлетворяет обоим условиям инвариантности. Кластерная процедура Q также удовлетворяет этим условиям, если процедура коррекции F удовлетворяет им. Предложены различные варианты таких процедур коррекции [19-21]. Изложим их в обобщенной форме [20-21]. Пусть $f_1, f_2, f_3 : [0, I] \to [0, I]$ - монотонные, неубывающие, положительные функции. Процедуры коррекции зависят от следующих множеств и параметров:

$$V_y(x) = \{z \in M \ \{x, y\} | S(x, z) \geq f_1(S(x, y))\},$$

$$V_x(y) = \{z \in M \ \{x, y\} | S(y, z) \geq f_1(S(x, y))\},$$

Множества $V_y(x)$ и $V_x(y)$ обозначают множества объектов «похожих» на x и на y соответственно, где значение $f_1(S(x, y))$ служит критерием этого сходства. Множество

$$V(x, y) = \{z \in X \ \{x, y\} | \max\{S(x, z), S(y, z)\} \geq f_2(S(x, y))\},$$

содержит объекты из M, которые «похожи» по крайней мере на один из объектов x и y. Когда $f_1 \equiv f_2$, имеем $V(x, y) = V_y(x) \cup V_x(y)$. Это множество представляет множество «соседей» x и y. «Голоса» объектов из $V(x, y)$ учитываются при принятии решения о коррекции значения $S(x, y)$. Множество

$$W(x, y) = \{z \in X \ \{x, y\} | \min\{S(x, z), S(y, z)\} \geq f_3(S(x, y))\},$$

обозначает множество «сильных» или «общих» соседей, т. е. объектов, которые «похожи» на оба объекта x и y. Объекты из $W(x, y)$ «поддерживают» значение $S(x, y)$. Когда $f_1 \equiv f_3$, имеем $W(x, y) = V_y(x) \cap V_x(y)$. Решение о коррекции значения $S(x, y)$ зависит от доли объектов, «поддерживающих» значение сходства $S(x, y)$. Эта величина может быть вычислена следующим образом: $h = \frac{|W(x,y)|}{|V(x,y)|}$, где $h = 1$, если знаменатель равен 0.

Пусть $L_V(x, y)$ обозначает упорядоченный в убывающем порядке список всех значений $S(x, z)$, $S(y, z)$, $(z \in V)$, которые имеют значение меньше, чем $S(x, y)$. Обозначим $m = |L_V(x, y)|$ количество элементов в $L_V(x, y)$, и $l_k (k = 1, \ldots, m)$ элементы $L_V(x, y)$ т.ч. при $m > 1$, $l_k \geq l_{k+1}$ для всех $k = 1, \ldots, m - 1$.

Возможные коррекции при m>1 будут определяться параметром j:
$j = 1 : F_j(x, y) = l_m$, т. е.$F_j(x, y)$ минимальное значение в $L_V(x, y)$;
$j = 2 : F_j(x, y) = l_1$, т. е. $F_j(x, y)$ максимальное значение в $L_V(x, y)$;
$j = 3 : F_j(x, y) = (\sum l_k)/m$, т. е. $F_j(x, y)$ есть среднее всех значений из $L_V(x, y)$;
$j = 4 : F_j(x, y) = l_k$, где $k \in 1, \ldots, m$ - параметр, $F1$ и $F2$ это частные случаи $F4$;
$j = 5 : F_j(x, y) = median(L_V(x, y))$.

Процедура коррекции $F(S) = S_c$ кластерной процедуры Q определяется следующим образом:

$$S_c(x, y) = \begin{cases} S(x, y), & \text{если } h \geq p; \\ F_j(x, y), & \text{в противном случае.} \end{cases}$$

где $p \in [0, 1]$ это параметр, и $F_j(x, y)$ это скорректированное значение, такое что $F_j(x, y) \subseteq S(x, y)$. Мы будем предполагать, что $F_j(x, y)$ зависит от значений $S(x, z)$, $S(y, z)$ для объекта z, принадлежащего множествам соседей x и y: $V_y(x)$, $V_x(y)$ и

$V(x,y)$. Все процедуры коррекции инвариантны относительно нумерации объектов и все $F_j(x,y)$ для $j = 1,2,4,5$ инвариантны относительно монотонного преобразования значений сходства. Инвариантность процедуры коррекции относительно нумерации объектов обеспечивается тем, что процедура коррекции применяется ко всем парам объектов (x,y) независимо от их нумерации или порядка рассмотрения, и параметры, использованные в процедуре коррекции, не зависят от нумерации объектов. Свойство инвариантности относительно монотонных преобразований значений сходства является дополнительным к свойству инвариантности относительно нумерации объектов.

3. Свойства схемы иерархической реляционной кластеризации

Рассмотрим некоторые свойства схемы иерархических реляционных кластерных процедур. Далее полагаем, что на M задано взвешенное отношение сходства S. Отношение линейного порядка $\leq$ на L определяет операции min и max обозначенные, как $\wedge$ и $\vee$ соответственно: $a \wedge b = a$, если $a \leq b$ и $a \wedge b = b$, если $b \leq a$; $a \vee b = b$, если $a \leq b$ и $a \vee b = a$, если $b \leq a$. Для любого значения (уровня) a из L взвешенное отношение S определяет отношение уровня a (a-срез) - отношение $S_{[a]}$ и взвешенное отношение S_a следующим образом: $S_{[a]} = \{(x,y) \in X | S(x,y) \geq a\}$; $S_a(x,y) = 1$, если $S(x,y) \geq a$; и $S_a(x,y) = 0$, если $S(x,y) < a$.

Подмножество $A \subseteq M$ будет называться классом сходства отношения сходства S на M, если $S(x,y) > S(u,z)$ для всех $x,y,u \in A$ и $z \notin A$. Класс сходства может рассматриваться как естественный кластер в множестве M. Значение $s = \min\limits_{x,y \in A} S(x,y)$ будет называться силой класса сходства A.

Утверждение 1. *Множество классов сходства взвешенного отношения эквивалентности S совпадает с множеством классов эквивалентности отношений уровня $S_{[a]}$, $a \in L$.*

Доказательство. Пусть S взвешенное отношение эквивалентности, a - некоторый элемент из L, и A класс эквивалентности отношения $S_{[a]}$. Тогда для всех $x,y,u \in A$ и всех $z \notin A$ мы имеем $(x,y) \in S_{[a]}$, $S(x,y) \geq a$, $(u,z) \notin S_{[a]}$, $S(u,z) < a$ и, следовательно, $S(x,y) > S(u,z)$. Следовательно, каждый класс эквивалентности $S_{[a]}$ будет классом сходства S, и разбиение M на классы эквивалентности, определенное отношением уровня $S_{[a]}$, будет определять разбиение M на классы сходства. Предположим A - класс сходства взвешенного отношения эквивалентности S, s - сила A, и $x*$, $y*$ некоторые элементы из A, для которых выполнено $S(x*,y*) = s$. Тогда для всех $x,y \in A$ мы имеем $S(x,y) \geq s$, и $(x,y) \in S_{[s]}$ и для всех $z \notin A$ мы имеем $S(x,z) < S(x*,y*) = s$ и $(x,z) \notin S_{[s]}$. Следовательно, класс сходства A совпадает с некоторым классом эквивалентности S.

Два объекта называются идентичными в S, если $S(x,y) = I$ и $S(x,z) = S(y,z)$ для всех z из $M \backslash \{x,y\}$. Два объекта x и y называются неразличимыми на уровне $a \in L$ если $S(x,y) \geq a$ и для всех $z \in M$ выполняется $S(x,z) \geq a$ тогда и только, когда $S(y,z) \geq a$. Ясно, что два объекта неразличимые на некотором уровне a будут идентичными в отношении сходства S_a. Ясно также, что все объекты неразличимы на минимально возможном уровне 0, и максимально возможный уровень неразличимости двух объектов x и y в отношении S равен $a = S(x,y)$. Два объекта x и y неразличимые на уровне $a = S(x,y)$ будут называться неразличимыми в S.

Утверждение 2. *Отношение сходства S, определенное на M, будет взвешенным отношением эквивалентности тогда и только тогда, когда все объекты из M неразличимы в S.*

Доказательство. Пусть x,y,z - произвольные объекты из M. Покажем, что если все объекты неразличимы в S, то выполняется (1). Если $S(x,z) \geq S(x,y)$, то очевидно (1) выполняется. Пусть $S(x,z) < S(x,y)$, тогда из условия неразличимости объектов в M следует $S(y,z) < S(x,y)$. Из этих неравенств, неразличимости объектов в M и симметричности S следует, соответственно, $S(x,z) \leq S(z,y)$, и $S(y,z) \leq S(y,x)$, $S(y,z) \leq S(z,x)$, откуда следует $S(x,z) = S(y,z)$ и выполнение (1). Пусть S - $(\vee,\wedge)$-транзитивно. Покажем, что все объекты неразличимы в S. Можно показать [22], что отношение сходства S будет взвешенным отношением эквивалентности тогда и только тогда, когда для всех $x,y,z \in M$ два наименьших из трех значений $S(x,y)$, $S(y,z)$, $S(z,x)$ равны. Откуда следует, что выполнено одно из двух: $S(x,y) > S(y,z) = S(x,z)$ или $S(y,z) \geq S(x,y)$ и $S(x,z) \geq S(x,y)$, т.е. каждая пара объектов x,y неразличима в S.

При разработке процедур реляционной кластеризации $E = TC(F(S))$ из схемы (3) естественным пожеланием является внесение минимальных изменений в исходное отношение сходства S при его преобразовании в результирующее отношение эквивалентности E. Из свойств процедуры транзитивного замыкания TC [14] следует, что $TC(F(S))$ преобразует отношение сходства $F(S)$ во взвешенное отношение эквивалентности E такое, что $F(S) \subseteq E$ и минимальное взвешенное отношение эквивалентности, содержащее $F(S)$. Следовательно, процедура транзитивного замыкания дает минимальное увеличение значений $F(S)(x,y)$ для преобразования $F(S)$ во взвешенное отношение эквивалентности E. Из утверждения 2 мы можем заключить, что эта процедура преобразует все пары объектов из M в неразличимые. Следовательно, мы можем предположить, что суммарное изменение значений сходства $F(S)(x,y)$ при преобразовании $TC(F(S)) =$ зависит от числа пар элементов x,y, которые не являются неразличимыми в $F(S)$ и преобразуются в неразличимые пары $E(x,y)$, и от «степени неразличимости» этих элементов, если мы можем измерить эту степень. Следовательно, процедура коррекции F в процедуре кластеризации (3), уменьшающая значения сходства $S(x,y)$, должна давать такие минимальные изменения этих значений, которые будут увеличивать число неразличимых пар объектов или увеличивать «степень неразличимости» пар объектов. В этом случае изменения, производимые процедурой транзитивного замыкания TC в отношении $F(S)$, будет небольшими, и итоговые изменения, производимые процедурой кластеризации (3), также будут небольшими. Эти соображения лежали в основе процедур коррекции, рассмотренных выше.

Утверждение 3. *Для кластерных процедур Q с тождественными функциями f_1 - f_3 выполнено $Q(S) = S$ тогда и только тогда, когда S -взвешенное отношение эквивалентности.*

Доказательство. Из конструкции кластерной процедуры Q видно, что если $Q(S) = S$ для некоторого отношения сходства S, то S будет отношением эквивалентности. Предположим, что S - взвешенное отношение эквивалентности. Покажем, что $Q(S) = S$. Из Утверждения 2 следует, что все пары объектов в S являются неразличимыми. Из определения процедуры коррекции следует, что если все функции f_1 - f_3 есть тождественные функции, то значения сходства $S(x,y)$ между объектами не будут изменяться процедурой коррекции F, потому что для

неразличимых объектов все множества $V(x,y)$, $V_y(x)$, $V_x(y)$ и $W(x,y)$ будут совпадать, $h=1$, и, следовательно, $S_c(x,y)=S(x,y)$ для всех $p\in[0,1]$. Тогда, из утверждения 2 получаем, что $S_c(x,y)=S(x,y)$ для всех пар объектов, следовательно, $F(S)=S$ и из свойства операции транзитивного замыкания мы получим $TC(F(S))=S$.

Утверждение 4. *Кластерные процедуры из рассматриваемой схемы сохраняют классы сходства, если функции f_1 и f_2 - тождественные.*

Доказательство. Предположим A - класс сходства отношения сходства S, s - сила этого класса, и $x,y\in A$, тогда $S(x,y)\geq s$, и для любого z из отношение $\max(S(x,z),S(y,z))\geq s$ выполнено тогда и только тогда, когда $z\in A$. Следовательно, множества $V(x,y)$, $V_y(x)$ и $V_x(y)$ будут содержать только элементы из . В результате, $F_j(x,y)\geq\min\limits_{z\in V}\{S(x,z),S(y,z)\}<s$, где $V=V(x,y)\cup V_y(x)\cup V_x(y)$. Следовательно, для всех $x,y\in A$ скорректированное значение $F(S)(x,y)$ будет больше или равно, чем s и, более того, мы будем иметь $E(x,y)=TC(F(S))(x,y)\geq s$. Покажем, что $E(x,z)=TC(F(S))(x,z)<s$ для всех $x\in A$ и $z\notin A$ и, следовательно, будет классом сходства взвешенного отношения эквивалентности . Из свойств транзитивного замыкания [14] имеем

$$E(x,z)=max_q\{min\{F(S)(y_0,y_1),F(S)(y_1,y_2),\ldots,F(S)(y_{n-2},y_{n-1}),\}\},$$

где $q=(y_0,y_1,y_2,\ldots,y_{n-2},y_{n-1})$ это путь длиной $n-1$ из $y_0=x$ в $y_{n-1}=z$ в множестве M, и max берется от всех таких путей в [14]. Для каждого такого пути существует индекс k такой, что $y_k\in A$ и $y_{k+1}\notin A$. Для этих двух объектов мы имеем $F(S)(y_k,y_{k+1})\leq S(y_k,y_{k+1})<s$ и, в результате, мы имеем $E(x,z)<s$.

Доказанные утверждения, во-первых, обосновывают кластерные процедуры с тождественными функциями f_1 и f_2 как процедуры, сохраняющие классы сходства, во-вторых, эти утверждения легли в основу разработки кластерных процедур, основанных на идее «разрыва мостиков между кластерами» [21,22]. Часто такие мостики содержат классы сходства, соединяющие более крупные кластеры. Из Утверждения 4 следует, что для построения реляционной кластерной процедуры на основе рассматриваемой выше схемы разрывающей такие мостики, необходимо чтобы функции f_1 и f_2 отличались от тождественных. В [23] были предложены такие процедуры.

4. Эволюционные процедуры двух- и трехмерной визуализации

Задача визуализации может рассматриваться как задача минимизации искажений исходных расстояний между объектами при их представлении в двух- или трехмерном пространстве. Схема (2) имеет вид:

$$HC(S)=<\oslash;P>,$$

где P задает координаты объектов в двух- или трехмерном пространстве.

Рассмотрим процедуру двухмерной визуализации данных. Предположим, что D - исходное отношение различия (матрица расстояний) между объектами из M. Генерируем начальную матрицу P координат объектов в двухмерном пространстве с осями X и Y. Используя ее как начальное приближение, с помощью стандартной процедуры оптимизации определяем новую матрицу координат объектов P^*,

которой соответствует матрица расстояний между объектами R с минимальной ошибкой аппроксимации начальной матрицы D. Обычно полученное представление P^* дает локальный оптимум. Затем определяем два объекта a и b из множества M, с максимальным значением $R(a,b)$. Систему координат $< X, Y >$ перемещаем и поворачиваем так, чтобы центр «новой» системы координат находился в точке a, а точка b располагалась на оси X. Пусть a и b имеют координаты (a_x, a_y) и (b_x, b_y) соответственно. Перемещение объектов осуществляется параллельным переносом вдоль оси X на величину $-a_x$ и вдоль оси Y на величину $-a_y$. Затем осуществляется поворот системы координат на угол между вектором (a,b) и осью X.

В новой системе координат X' и Y'выбираем объект c с максимальным абсолютным значением координаты по оси $Y^{'}$. В зависимости от знака этой координаты осуществляется зеркальное отображение всех объектов относительно оси $Y^{'}$. Объекты a, b, c будут опорными элементами для всех последующих матриц координат объектов. Полученная матрица координат P^* объектов из называется решением и ошибка аппроксимации матрицы D матрицей расстояний R, вычисленной на основе матрицы координат P^*, называется ошибкой решения.

Случайным образом генерируется множество m матриц начальных координат объектов в двухмерном пространстве и для каждой из них вычисляется ошибка аппроксимации. Полученное множество матриц называется популяцией. Наилучшие q матриц с наименьшей ошибкой аппроксимации, выбранные из популяции, называются элитой.

Для каждой матрицы из элиты применяется перемещение и поворот относительно опорных точек a, b, c так, чтобы их расположение было таким же, как и в матрице P^*. Полученные решения затем используются для генерации новых решений, называемых потомками. Потомки получаются в результате применения следующих шагов. Из элиты случайным образом выбирается пара решений («родители»), которые впоследствии используются для построения новых решений («потомков») с помощью операций рекомбинации и мутации, осуществляемых следующим образом.

Операция рекомбинации: случайным образом выбирается объект x из множества M, и все объекты одного из родителей, с номерами меньше, чем номер объекта x принимают координаты тех же объектов другого родителя. Операция мутации: к матрице координат решения добавляется матрица, составленная из нормально распределенных малых величин.

Новая популяция получается с помощью добавления к старой элите потомков, полученных применением операции рекомбинации к старой элите, потомков, полученных из элиты применением операции мутации и потомков, полученных с помощью применения мутации после применения операции рекомбинации к старой элите.

Для новых элементов элиты, определяющих матрицы координат объектов в двухмерном пространстве, вычисляются матрицы расстояний и ошибка аппроксимации исходной матрицы расстояний D.

Новая элита выбирается из полученной популяции следующим образом: половина элиты состоит из наилучших решений популяции, а вторая половина выбирается из популяции случайным образом. Для новой элиты повторяются все шаги, описанные выше: перемещение и поворот всех координат, выбор «родителей» и

генерация «потомков».

Генерация популяции повторяется заданное число раз или до тех пор, пока ошибка аппроксимации не станет меньше заданного значения.

Поиск трехмерного представления объектов осуществляется по аналогичной схеме, с тем отличием, что опорных элементов выбирается четыре и зеркальное отображение элементов выполняется относительно двух осей [24,25].

5. Гибридная кластеризация с двух- и трехмерной визуализацией результатов кластеризации

Рассмотрим совместное использование методов иерархической кластеризации с методами двух- и трехмерной визуализации данных. В результате иерархической агломеративной кластеризации все объекты и кластеры поэтапно будут объединятся в один кластер. Необходимо отметить, что обычно невозможно представить многомерные данные в двухмерном или трехмерном пространстве без изменений значений расстояния. По этой причине рассматриваемый подход к гибридной визуализации данных пытается уменьшить изменения расстояний между объектами внутри маленьких классов за счет увеличения изменений расстояний между объектами из разных классов на высших уровнях иерархии.

Схема гибридной кластеризации (2) имеет вид:

$$HC(S) =< E; P >,$$

где E - взвешенное отношение эквивалентности, получаемое реляционной процедурой иерархической кластеризации, а P - описание объектов множества M в двух- или трехмерном пространстве, согласованное в определенном смысле с реляционной процедурой иерархической кластеризации. P последовательно строится так, чтобы расстояния между объектами из кластеров, объединяемых в соответствии с иерархической процедурой кластеризации, отклонялись от исходных расстояний на минимальную величину. Вычисление координат P осуществляется генетическим алгоритмом. Рассмотрим работу этого алгоритма для двухмерного представления результатов классификации.

На первом этапе кластеризации все кластеры иерархической кластеризации содержат малое число объектов и сосредоточены на нижних уровнях формируемой иерархии. Если кластер состоит из двух объектов, то эти объекты получают координаты (0,0) и (d,0), соответственно, где $d = D(x, y)$. Если кластер состоит более, чем из двух объектов, то к множеству этих объектов применяется генетический алгоритм, описанный в предыдущем разделе. Алгоритм пытается найти координаты объектов из некоторого класса так, чтобы расстояния между ними отличались от соответствующих исходных расстояний из D на минимальную величину.

На втором этапе минимизируются изменения расстояний между кластерами, объединяющимися вместе в соответствии с последовательностью построения иерархии кластеров кластерной процедурой. Один из двух объединяемых кластеров остается неподвижным, и координаты объектов из этого класса остаются неизменными. Другой кластер перемещается так, чтобы расстояния между объектами из неподвижного кластера и объектами из перемещаемого кластера стали как можно более похожи на исходные расстояния между этими объектами, причем чтобы расстояния между объектами внутри перемещаемого класса, полученные на первом

этапе, не изменялись. Координаты объектов из этого класса изменяются следующим образом:

$$X_{new} = X * \cos(a) + Y * \sin(a) + dX,$$

$$Y_{new} = -X * \sin(a) + Y * \cos(a) + dY,$$

где X^*, Y^* - это координаты объектов в двухмерном пространстве до перемещения класса, X_{new}, Y_{new} - координаты объектов после перемещения, dX, dY - величины сдвига объектов вдоль осей X и Y, а - угол поворота класса вокруг начала координат. Строка параметров (dX, dY, a) представлена как элемент популяции в генетическом алгоритме.

Были использованы следующие характеристики генетического алгоритма: размер начальной популяции (количество строк) равен 500, количество строк, отбираемых в элиту, равно 15, количество генераций новых популяций равно 300. Элементы строк начальной популяции - случайные нормально распределенные числа из интервала [-100, 100] для dX и dY, $[-\pi, \pi]$ для угла a. Для каждой из 500 сгенерированных строк было вычислено среднее изменение расстояний между рассматриваемыми классами следующим образом:

$$E = \frac{sum_j(sum_i(abs(R_{ij} - D_{ij})))}{(n * m)},$$

где D_{ij} и R_{ij} - расстояния между объектами i из фиксированного класса и объектами j из перемещаемого класса, исходное и вычисленное в двухмерном пространстве, соответственно. Наилучшие 15 строк параметров с минимальным значением были отобраны в элиту. Генерация новых строк из строк, отобранных в элиту, основана на операциях мутации и рекомбинации.

Две строки-родители из элиты и один параметр из (dX, dY, a) выбираются случайно. Две новых строки-потомки, полученные из двух строк-родителей как результат обмена значениями выбранного параметра между строками-родителями. Этот процесс повторяется 15 раз и, в результате, мы имеем 30 строк-потомков. Новая популяция получается объединением элиты, потомков и мутированной элиты, так что размер популяции становится равным 90.

Была использована следующая операция мутации. Каждому параметру мутируемых строк добавляется малая нормально распределенная ошибка:

$$e = K * r/T,$$

где r - случайное число и - номер текущей генерации, принимающий значения от 1 до 300. Для dX и dY коэффициент равен 1000. Для a он равен $0.2 * \pi$.

Для каждой новой популяции выбирается новая элита, с помощью которой генерируется новая популяция и т. д.

Результатом этой процедуры является оптимальный сдвиг и поворот перемещаемого класса относительно неподвижного класса. Если перемещаемый класс состоит более чем из двух объектов, то трех операций, определяемых параметрами (dX, dY, a) может быть не достаточно, для наилучшего расположения этого класса, потому что может быть необходимо так же сделать зеркальное отображение этого класса относительно одной из осей X или Y. По этой причине эволюционная процедура применяется так же к множеству $2D$ координат, полученных из

множества координат объектов перемещаемого класса заменой всех координат Y на $-Y$. Наилучшие решения, полученные после применения генетического алгоритма к перемещаемому классу и его зеркальное отображение, дают оптимальные координаты перемещаемого класса.

Аналогично можно определить процедуру трехмерной визуализации результатов кластеризации. Эта процедура отличается от описанной выше только тем, что координаты всех объектов из перемещаемого класса изменяются следующим образом:

$$X_{new} = X * \cos(a) + Y * \sin(a) + dX,$$

$$Y^{'} = -X * \sin(a) + Y * \cos(a) + dY,$$

$$Y_{new} = Y^{'} * \cos(a) + Z * \sin(a) + dY,$$

$$Z_{new} = -Y^{'} * \sin(a) + Z * \cos(a) + dZ,$$

где X, Y, Z суть $3D$ координаты объектов до перемещения класса, X_{new}, Y_{new}, Z_{new} - координаты объектов после перемещения, dX, dY, dZ - значения сдвига объектов из перемещаемого класса вдоль осей X, Y, Z, a - угол поворота класса. Эволюционный алгоритм визуализации ищет такие значения четырех параметров (dX, dY, dZ, a), определяющих перемещение кластера, при которых отклонения между исходными и $3D$ расстояниями между объектами неподвижного и перемещаемого кластеров минимальны. Поиск оптимального вектора (dX, dY, dZ, a) происходит по общей схеме генетического алгоритма: случайным образом генерируется начальное множество таких векторов (начальная популяция); для каждого из них вычисляется значение функции F, равное суммарному отклонению между исходными и $3D$ расстояниями; выбирается элита - заданное число наилучших векторов с минимальными значениями F; из векторов элиты генерируются новые вектора (новая популяция) с помощью операций кроссинговера и мутации; для новой популяции вычисляются значения F; выбирается новая элита и т.д.

Операция кроссинговера: случайным образом из элиты выбирается пара векторов-родителей, часть координат одного из них заменяется координатами другого, и наоборот, в результате чего получаются два новых вектора- потомка.

Операция мутации: вектору координат прибавляются приращения $e = K * S/T$, где S - стандартная нормально распределенная случайная величина, T - номер популяции, $= 1000$ для dX, dY, dZ, и $= 0.2 * \pi$ для a. Операция мутации применяется к векторам-родителям и векторам-потомкам.

Новая популяция образуется из элиты, потомков и мутированных векторов элиты и потомков [24,25].

6. Гибридная одномерная кластеризация временных рядов

Здесь предлагается метод сочетания иерархической кластеризации с одномерной визуализацией данных по некоторому показателю. Общая схема метода:

$$HC(S) =< E; P >,$$

где E - нечеткое отношение эквивалентности, определяющее иерархическую кластеризацию данных, а P - описание объектов по определенному показателю такое,

что $P(x)$ - значение этого показателя для объекта x. Согласованность E и P может быть достигнута при наличии связи между расстоянием между объектами, используемым в алгоритме кластеризации, и показателем P. Если объектами являются временные ряды, и алгоритм кластеризации использует Евклидово расстояние между ними, то в качестве такого показателя P может быть выбрано среднее значение временного ряда за рассматриваемый интервал времени.

7. Гибридная кластеризация с визуализацией сильных связей

Этот метод исходит из того, что структура данных может не содержать «естественной» кластеризации, и визуализация сильных связей может быть полезна для анализа отклонения полученной кластеризации от структуры данных. Идея метода заключается в визуализации сильных связей между объектами в виде графа таким образом, что объекты с большим взаимным сходством соединены ребром. Уровень сходства выбирается из соображений минимизации расхождения графа сходства с полученной кластеризацией.

Схема гибридной кластеризации имеет вид:

$$HC(S) =< E, S_\alpha; \oslash >$$

где E - четкое отношение эквивалентности на M, задаваемое реляционным кластерным алгоритмом, а S_α - оптимальный α-срез исходного взвешенного отношения сходства S на уровне сходства α. Критерий оптимальности α определяется ниже. α-срез отношения S определяется взвешенным отношением:

$$S_\alpha(x,y) = \begin{cases} 1, & \text{если } S(x,y); \\ 0, & \text{в противном случае.} \end{cases}$$

Обозначим E_1 характеристическую функцию отношения E:

$$E_1(x,y) = \begin{cases} 1, & \text{если } (x,y) \in E; \\ 0, & \text{в противном случае.} \end{cases}$$

Пусть M содержит n_M объектов. Обозначим $N_\alpha = \sum\limits_{x,y \in M} S_\alpha(x,y) - n_M$ общее число «сильных» связей $S_\alpha(x,y) = 1$ в α-срезе S_α за вычетом рефлексивных связей $S_\alpha(x,x) = 1$. Пусть $\{A_1, \ldots, A_u\}$ - классы эквивалентности (кластеры) отношения E, содержащие по $n_1, \ldots, n_u$ объектов, соответственно. Обозначим ${N_\alpha}^+ = \sum\limits_{k=1}^{u} (\sum\limits_{x,y \in A_k} S_\alpha(x,y)) - n_M$ - число внутри-кластерных сильных связей отношения S_α. Тогда ${N_\alpha}^- = N_\alpha - {N_\alpha}^+$ равно числу межкластерных сильных связей. Обозначим $N_E = \sum\limits_{x,y \in M} E_1(x,y) - n_M$ - число связей в E_1 за вычетом рефлексивных связей. В гибридной процедуре кластеризации оптимальное значение α выбирается таким образом, чтобы минимизировать число рассогласований между сильными связями в S_α и сильными связями в E_1:

$$G = N_E - {N_\alpha}^+ + {N_\alpha}^- = N_E + N_A - 2{N_\alpha}^+.$$

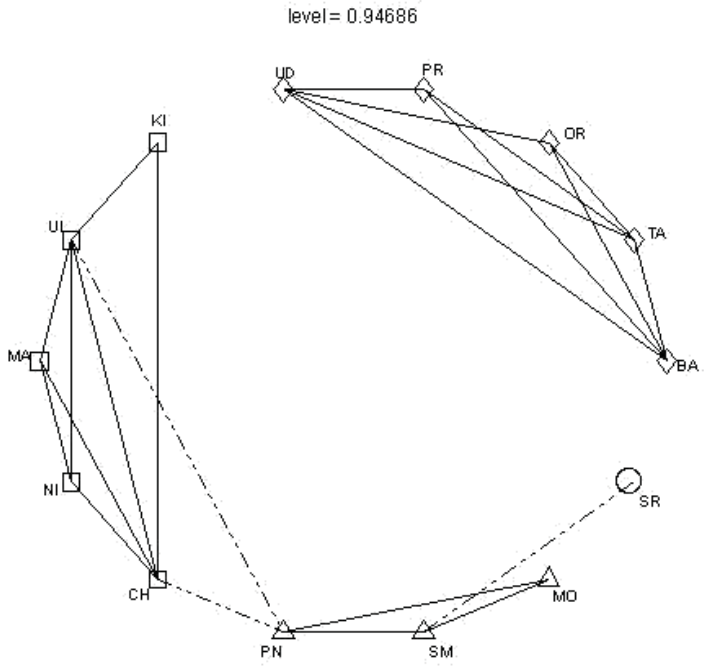

Рис. 1: Оптимальная гибридная кластеризация регионов ПФО. Регионы из разных кластеров обозначены разными фигурами. Внутри-кластерные и межкластерные сильные связи показаны сплошными и пунктирными линиями, соответственно.

Поскольку может быть несколько оптимальных значений α , в качестве дополнительного критерия для выбора α может рассматриваться максимизация ${N_\alpha}^+$ или минимизация ${N_\alpha}^-$[13].

Предлагаемый метод гибридной кластеризации проиллюстрирован на примере гибридной кластеризации 14 регионов Приволжского Федерального Округа на основе статистических данных об их инвестициях в основной капитал за период 1999-2004 гг. [26]. В качестве меры сходства между временными рядами инвестиций использовалась мера ассоциаций локальных трендов, основанная на преобразовании скользящих аппроксимаций [27]. Для получения кластеризации $R_1 = E$ применялись реляционные кластерные алгоритмы, рассмотренные в Разделе 2. Рис. 1 показывает полученную оптимальную гибридную кластеризацию для $\alpha = 0,94686$. Сплошными и пунктирными линиями показаны сильные связи между временными рядами из одинаковых кластеров и из разных кластеров, соответственно. Используемая аббревиатура обозначает следующие регионы: BA - Башкортостан, MA - Марий Эл, MO - Мордовия, TA - Татарстан, UD - Удмуртия, CH - Чувашия, KI - Кировская область, NI - Нижегородская область, OR - Оренбургская область, PN - Пензенская область, PR - Пермская область, SM - Самарская область, SR - Саратовская область, UL - Ульяновская область. После применения процедуры

кластеризации, были получены четыре различных кластера: C1={BA, TA, OR, PR, UD}, C2={KI, UL, MA, NI. CH}, C3={PN, SM, MO}, C4={SR}. Регионы, объединенные в один кластер, обозначены одинаковыми фигурами. Визуализация сильных связей показывает, что полученная кластеризация достаточно хорошо отражает структуру сильных связей между регионами. Рассогласование имеется лишь по трем внутри-кластерным связям {PR,OR}, {KI,MA}, {KI,NI} и трем межкластерным связям {UI,PN}, {CH,PN}, {SR,SM}.

Заключение

В работе рассмотрены общая постановка гибридной реляционной кластеризации и визуализации данных и различные частные случаи общей схемы. Гибридная реляционная кластеризация позволяет сочетать кластеризацию данных с их пространственной визуализацией либо с выделением сильных связей, что дает более полную информацию о структуре исследуемых данных. Предложенный в данной работе метод гибридной одномерной визуализации и кластеризации данных целесообразно использовать, когда может быть введен обобщенный показатель, определяющий одномерную визуализацию данных. Разработанные методы гибридной кластеризации реализованы в виде комплекса программ на языке МАТЛАБ.

Список литературы

[1] Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Классификация и снижение размерности. Справочное издание /Под ред. С.А. Айвазяна. – М.: Финансы и статистика, 1989. 607 с.

[2] Дюран Б., Оделл П. Кластерный анализ. – М.:Статистика, 1977. 128с.

[3] Мандель И.Д. Кластерный анализ. – М.: Финансы и статистика. 1988. 176 с.

[4] Fielding A.H. Cluster and Classification Techniques for the Biosciences, Cambridge University Press, 2007.

[5] Knudsen S. Biologist's Guide to Analysis of DNA Microarray Data. Wiley-Interscience, 2002.

[6] Chen Y., Qiu L., Chen W., Nguyen L., Katz R.H. Clustering Web content for efficient replication. In: Proc. 10th IEEE Intern. Conf. on Network Protocols, 2002, pp. 165-174.

[7] Friedman M., Kandel A. Introduction to Pattern Recognition. Statistical, Structural, Neural and Fuzzy Logic Approaches. World Scientific, 1999.

[8] Pedrycz W. Knowledge-Based Clustering. From Data to Information Granules. Wiley-Interscience, 2005.

[9] Jain A.K., Murty M.N., Flynn P.J. Data clustering: A review. - ACM Computing Surveys, Vol. 31, 3, 1999, pp. 264–323.

[10] Strehl A., Ghosh J. Cluster ensembles – A knowledge reuse framework for combining multiple partitions. – Journal of Machine Learning Research 3, 2002, pp. 583-617.

[11] Zhou A., Qian W., Qian H., Wen J., Zhou Sh., Fan Y. A hybrid approach to clustering in very large databases, in: PAKDD 2001, LNAI 2035, 2001, pp. 519-524.

[12] Крускал Дж. Взаимосвязь между многомерным шкалированием и кластер-анализом. В кн.: Классификация и кластер /Под ред. Дж.Вэн Райзина. - М: Мир, 1980. С. 20-41.

[13] Батыршин И.З., Климова А.С. Гибридная реляционная кластеризация и визуализация данных. Труды Всеросс. научн. конф. по нечетким системам и мягким вычислениям НСМВ-2006. М.: Физматлит, 2006. С. 193-209.

[14] Аверкин А.Н., Батыршин И.З., Блишун А.Ф., Силов В.Б., Тарасов В.Б. Нечеткие множества в моделях управления и искусственного интеллекта/ Под ред. Д.А. Поспелова. – М.: Наука. Гл. ред. физ.-мат. лит., 1986. 312 с.

[15] Tamura S., Higuchi S., Tanaka K. Pattern classification based on fuzzy relations.// IEEE Trans. Systems, Man, Cybernetics, SMC-1. 1971. pp. 61-66.

[16] Zadeh L.A. Similarity relations and fuzzy orderings.// Information Sciences. 1973. № 3. pp. 177-200.

[17] Jardine N., Sibson R. Mathematical taxonomy.// London: John Wiley and Sons. 1971.

[18] Johnson S.C. Hierarchical clustering schemes.// Psychometrika. 1967. № 32. pp. 241-254.

[19] Батыршин И.З., Шустер В.А. Структура семантических пространств вербальных оценок действий // Acta et Commentationes Universitas Tartuensis, Transactions on Artificial Intelligence, Principle Questions of Knowledge Theory, Tartu. 1984. № 688. pp. 20-38.

[20] Батыршин И. З, Климова А. С. Инвариантные кластерные процедуры, основанные на нечетком отношении сходства. Интегрированные модели и мягкие вычисления в искусственном интеллекте. Сборник научных трудов III-го Международного научно-практического семинара, 2005. – Москва, Физматлит, 2005, С. 119 - 125.

[21] Batyrshin I.Z., Rudas T., Klimova A. On general scheme of invariant clustering procedures based on fuzzy similarity relation // Proc. Int. Conf. Fuzzy Sets and Soft Computing in Economics and Finance, FSSCEF 2004, St. Petersburg, Russia. 2004. vol. I. pp. 122-129

[22] Батыршин И.З. О транзитивности размытых упорядочений// Исследование операций и аналитическое проектирование в технике. - Казань: Казанск.авиац.ин-т, 1979. С. 67-73.

[23] Batyrshin I., Klimova A. New invariant relational clustering procedures, in: Proceedings of East West Fuzzy Colloquium 2002, 10th Zittau Fuzzy Colloquium, Zittau, Germany, 2002, 264 - 269.

[24] Батыршин И.З., Климова А.С. Эволюционные процедуры иерархической двухмерной визуализации данных. В сб.: Исследования по информатике. Институт проблем информатики АН РТ. - Казань, Отечество, N 7, 2004, 119-124.

[25] Klimova A. Evolutionary procedures of visualization of multidimensional data // Proc. Int. Conf. Fuzzy Sets and Soft Computing in Economics and Finance, St. Petersburg, Russia. 2004. vol. I. pp. 130-139.

[26] Россия в цифрах. Федеральная служба государственной статистики http://www.gks.ru/scripts/db_inet/dbinet.cgi?pl=2702005.

[27] Batyrshin I.Z., Klimova A.S., Sheremetov L.B., Velasco-Hernandez J.X. Combining local trend association network and clustering in visualization of relationships in time series data bases. In: Proc. Intern. Conf. Fuzzy Sets and Soft Computing in Economics and Finance, FSSCEF 2006, June - July 1, 2006, St. Petersburg, Russia, pp. 242-251.