На правах рукописи

КЛИМОВА АНЖЕЛИКА СЕРГЕЕВНА

МОДЕЛИ И МЕТОДЫ ГИБРИДНОЙ РЕЛЯЦИОННОЙ КЛАСТЕРИЗАЦИИ ДАННЫХ

Специальность 05.13.18 – Математическое моделирование, численные методы и комплексы программ

АВТОРЕФЕРАТ

диссертации на соискание ученой степени кандидата технических наук

Казань - 2013

Работа выполнена на кафедре информатики и прикладной математики федерального государственного бюджетного образовательного учреждения высшего профессионального образования «Казанский национальный исследовательский технологический университет»

Научный руководитель: доктор физико-математических наук, профессор, заслуженный деятель науки Республики Татарстан Батыршин Ильдар Закирзянович

Официальные оппоненты: Ярушкина Надежда Глебовна, доктор технических наук, профессор, заведующий кафедрой «Информационные системы» ФГБОУ ВПО «Ульяновский государственный технический университет»

Ибятов Равиль Ибрагимович, доктор технических наук, профессор, заведующий кафедрой прикладной информатики и математики ФГБОУ ВПО «Казанский государственный аграрный университет»

Ведущая организация: ФГБОУ ВПО «Тверской государственный университет», г. Тверь

Защита состоится 28 июня 2013 года в 16.00 часов на заседании диссертационного совета Д 212.080.13 при Казанском национальном исследовательском технологическом университете по адресу: 420015, г.Казань, ул. К. Маркса, 68, зал заседаний Ученого совета (А-330).

Отзыв на автореферат в двух экземплярах, заверенный гербовой печатью, просим направлять по адресу: 420015, г. Казань, ул. К. Маркса, 68, Казанский национальный исследовательский технологический университет, ученому секретарю диссертационного совета Д 212.080.13.

С диссертацией можно ознакомиться в фундаментальной библиотеке Казанского национального исследовательского технологического университета.

Автореферат разослан 28 мая 2013 г.

Ученый секретарь диссертационного совета Д 212.080.13, доктор технических наук, профессор Клинов Александр Вячеславович

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ[1]

Актуальность. Кластеризация данных играет большую роль в анализе данных в технике, в экономике, в социальных науках, биологии, геологии, астрономии и в других научных областях. Кластеризация позволяет представить исследуемые данные в виде разбиения на классы сходных объектов, что является одним из важных этапов формирования знаний об исследуемой предметной области, ее моделирования и анализа. Но структура реальных данных не всегда может быть адекватно представлена в виде разбиения на классы сходных объектов, также данные могут допускать различные разбиения на классы сходных объектов, кроме того, помимо структуры классов сходства, формируемой кластерным алгоритмом, часто желательно выявить дополнительную информацию о связях между объектами. Поэтому актуальной является задача разработки гибридных методов кластеризации, дающих новые методы представления структуры данных на основе кластерного анализа. В настоящее время активно развиваются методы сочетания нескольких кластерных процедур, методы комбинирования кластеризации с визуализацией данных и др. [Strehl, Jain, Murty, Babu, Крускал, Ярушкина и др.]. При этом важным является использование кластерных алгоритмов, удовлетворяющих условиям инвариантности относительно исходной нумерации (перестановки) кластеризуемых объектов и инвариантности относительно монотонных преобразований значений сходства [Jardine, Johnson, Батыршин и др.]. К сожалению, эти условия инвариантности не выполняются для большинства известных кластерных алгоритмов. Поэтому важной является задача разработки гибридных кластерных алгоритмов, удовлетворяющих указанным условиям инвариантности. В работах Батыршина И.З. и др. разработана реляционная схема иерархических инвариантных процедур кластеризации, основанная на преобразовании заданного взвешенного отношения сходства во взвешенное (нечеткое) отношение эквивалентности, определяющее иерархию разбиения множества объектов на кластеры определенного типа. Перспективной является задача расширения этой схемы для выделения кластеров новых типов и разработки гибридных кластерных процедур на ее основе.

Цель работы: Разработка моделей, методов и комплекса программ гибридной кластеризации данных на основе реляционных инвариантных процедур кластеризации.

---

[1] Автор выражает благодарность к.т.н. Богомолову В.А. за помощь, оказанную при работе над диссертацией.

Задачи исследования.

1. Теоретическое исследование свойств реляционной схемы инвариантных иерархических кластерных процедур с целью исследования возможности ее расширения на новые типы кластеров.
2. Разработка и реализация новых реляционных инвариантных процедур иерархической кластеризации.
3. Разработка методов гибридной кластеризации на основе реляционных кластерных процедур.
4. Создание комплекса программ гибридной реляционной кластеризации данных.

Методы исследования: кластерный анализ, теория нечетких множеств, теория графов, теория генетических алгоритмов.

Научная новизна работы.

1. Получено теоретическое обоснование реляционной схемы инвариантных иерархических кластерных процедур.
2. Предложено расширение реляционной схемы инвариантных иерархических кластерных процедур, содержащее новые кластерные процедуры.
3. Разработаны методы построения моделей данных в виде инвариантных (относительно исходной нумерации и относительно монотонных преобразований значений сходства) кластеров сходных объектов, сетей сходства и их визуализации в двумерном и трехмерном пространствах.
4. Разработаны численные методы поиска оптимальных представлений данных в виде кластеров и визуализации этих кластеров.

Достоверность результатов диссертационной работы подтверждается результатами экспериментов и практического использования материалов диссертации, и разработанного пакета программ.

Практическая значимость работы состоит в разработке в среде МАТЛАБ пакета программ гибридной реляционной кластеризации данных, позволяющего исследовать инвариантные структуры сходства в задачах кластеризации данных и анализа структуры систем, в разработке методов визуализации и гибридной кластеризации данных, в разработке методов анализа взаимодействия скважин нефтяного месторождения на основе данных добычи нефти и сопутствующей воды. Результаты работы внедрены в Институте проблем информатики АН РТ, Министерстве образования и науки РТ и введены в состав учебной дисциплины СД.04 "Нечеткая логика и мягкие вычисления" на кафедре информатики и прикладной математики Казанского государственного технологического университета.

Апробация работы. Основные положения и результаты работы обсуждены на международых конференциях "East West Fuzzy Colloquium" (Германия, Циттау, 2002, 2006) и "Fuzzy Sets and Soft Computing in Economics and Finance" (Санкт-Петербург, 2004, 2006), на II Всероссийской научно-

технической конференции "Проблемы информатики в образовании, управлении, экономике и технике" (Пенза, 2002), III Международном научно-практическом семинаре "Интегрированные модели и мягкие вычисления в искусственном интеллекте" (Коломна, 2005), на Всероссийской конференции "Нечеткие системы и мягкие вычисления" (Тверь, 2006).

Публикации результатов работы. Основные выводы и положения диссертации изложены в 17 печатных работах. Среди них 3 статьи в журналах из перечня ВАК, 8 – в материалах конференций, 6 - в журналах и сборниках научных работ академических и центральных изданий.

Ряд результатов диссертационной работы получен в рамках проектов фонда НИОКР и АН РТ (05-5.2-173/2002) и РФФИ (03-01-96-245-р200) по теме "Разработка методов моделирования процессов и систем на основе нечеткой логики, нечетких отношений и нейронных сетей", 02-01-00092-а "Разработка моделей и методов вычисления словами на основе гранулирования информации о нечетких зависимостях и оптимизации нечетких моделей по параметрам операций", а также совместных исследований с учеными Мексиканского нефтяного института.

Структура и объем работы.

Диссертационная работа состоит из введения, четырех глав, заключения и списка литературы, изложенных на 105 страницах машинописного текста. Содержит 3 таблицы, 18 рисунков, список литературы из 96 наименований.

СОДЕРЖАНИЕ РАБОТЫ

Во **введении** рассмотрена актуальность проблемы разработки методов гибридной реляционной кластеризации данных. Введение содержит формулировку целей и задач диссертационной работы: разработка моделей, методов и комплекса программ гибридной кластеризации данных на основе реляционных инвариантных процедур кластеризации.

Первая глава "Обзор методов кластеризации и визуализации данных и постановка задачи исследования" посвящена анализу работ по кластерному анализу, методам визуализации многомерных данных, гибридному кластерному анализу и формулировке целей исследования.

Задача гибридной кластеризации в реляционной форме формулируется в данной работе следующим образом. Пусть X – множество исследуемых объектов, S – отношение сходства, заданное на X, (S может быть получено из матрицы расстояний между объектами), RC - множество обыкновенных или взвешенных (нечетких) отношений на X, содержащее отношение эквивалентности, PC – множество описаний элементов из X в пространствах различной размерности. Гибридной реляционной кластеризацией X называется преобразование

$$HC(S) = <RC; PC>. \quad (1)$$

В диссертационной работе в качестве базовых методов гибридной кластеризации разработаны инвариантные процедуры реляционной иерархической кластеризации и генетические процедуры визуализации данных.

Вторая глава "Инвариантные процедуры реляционной иерархической кластеризации" посвящена изложению теоретического обоснования схемы инвариантных реляционных иерархических кластерных процедур, описанию новой инвариантной реляционной кластерной процедуры, основанной на идее «разрыва мостов между кластерами», описанию гибридной процедуры реляционной кластеризации с визуализацией сильных связей между объектами.

В качестве исходной информации для кластерной процедуры рассматривается взвешенное отношение сходства $S:X\times X\to R$, где X – множество из n объектов, $R=[0,I]$, I - некоторое положительное вещественное число, и S удовлетворяет условиям рефлексивности: $S(x,x) = I$, и симметричности: $S(x,y)=S(y,x)$ для всех x,y из X. Взвешенное отношение сходства называется взвешенным отношением эквивалентности, если оно удовлетворяет условию $(\vee,\wedge)$-транзитивности: $S(x,z)\geq S(x,y)\vee S(y,z)$ для всех x,y из X.

Для любого значения (уровня) a из R взвешенное отношение S определяет отношение уровня $S_{[a]}$ и взвешенное отношение S_a следующим образом: $S_{[a]}=\{(x,y)\in X \mid S(x,y)\geq a\}$; $S_a(x,y)=1$, если $S(x,y)\geq a$; $S_a(x,y)=0$, если $S(x,y)<a$.

Подмножество A множества X называется классом сходства отношения сходства S на X, если $S(x, y)>S(x, z)$ для всех $x, y\in A$ и $z\notin A$.

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

Реляционная схема иерархических кластерных процедур, удовлетворяющих условиям инвариантности относительно исходной нумерации объектов и инвариантности относительно монотонных преобразований значений сходства, была введена в работах Батыршина И.З. Схема (1) для данного случая имеет вид: $HC(S) = \langle E; \varnothing\rangle$, где E – взвешенное отношение эквивалентности. Преобразование $HC(S)$ задается процедурой кластеризации $Q(S)=E$, которая определяется так: $E = Q(S) = TC(F(S))$, где F это некоторая "коррекция" данного отношения сходства S такая, что $F(S)\subseteq S$, и TC обозначает процедуру транзитивного замыкания значений отношения сходства.

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

***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.

***Утверждение* 4.** Кластерные процедуры из представленной схемы "сохраняют классы сходства", если функции f_1 и f_2, используемые в этой процедуре, суть тождественные функции.

В разделе 2.5 описывается новый метод, основанный на идее "обрыва мостов между кластерами". Эта кластерная процедура использует другой тип множества поддерживающих точек. Вместо функции сходства S будем использовать функцию различия $D:X\times X\to[0,I]$, удовлетворяющую на X условиям: $D(y,x) = D(x,y)$, $D(y,y) = 0$. Если D удовлетворяет на X ультраметрическому неравенству, двойственному $(\vee,\wedge)$-транзитивности: $D(x,y) \le max\{(D(x,z),D(z,y)\}$, то D называется ультраметрикой.

Две точки считаются точками из "моста", если отсутствуют поддерживающие точки "вдоль граней моста", соединяющего эти точки. Дополнительно к отсутствию поддерживающих точек мы используем также некоторое топологическое восприятие точек моста. Этот метод использует следующие множества: $V^*_y(x)=\{z\in V_y(x)\mid f_2(D(x,y))\le D(x,z)\}$,

$$V^*_x(y)=\{z\in V_x(y)\mid f_2(D(x,y))\le D(y,z)\},$$

где $f_2(D(x,y)) \le f_1(D(x,y))$. Множество поддерживающих точек W определяется равенством: $W(x,y)= V^*_y(x)\cap V^*_x(y)$. Определение топологии точек мостов использует следующие множества:

$U(x)=\{z\in X\setminus\{x,y\}\mid D(x,z)\le f_3(D(x,y))\}$, $U(y)=\{z\in X\setminus\{x,y\}\mid D(y,z)\le f_3(D(x,y))\}$.

Мы будем говорить, что точки x и y различны на уровне k, если $H(x,y)\ge k$, где $H(x,y)= abs(|U(x)|-|U(y)|)$, и k – некоторый параметр. Процедура коррекции выглядит следующим образом:

$$F(D(x,y)) = \begin{cases} D(x,y), & \text{если } |W(x,y)|>t \text{ или } H(x,y)\ge k, \\ F_j(x,y), & \text{в остальных случаях,} \end{cases}$$

где t – некоторый параметр и $F_j(x,y)$ определяется коррекциями, описанными выше для значений $j=1, 2, 3$.

В разделе 2.6 изложен, разработанный в диссертации новый метод гибридной кластеризации, сочетающий инвариантную кластеризацию с визуализацией сильных связей между объектами. Идея метода заключается в визуализации сильных связей между объектами в виде графа сходства таким образом, что объекты с взаимным сходством не меньше, чем некоторый уровень сходства α, были соединены ребром. Уровень сходства выбирается из соображений минимизации расхождения графа сходства с полученной кластеризацией. Способ нахождения оптимального уровня α определяется ниже. Обозначим E_1 характеристическую функцию отношения эквивалентности E:

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

Пусть X содержит n_X объектов. Обозначим $N_\alpha = \sum_{x,y\in X} S_\alpha(x,y) - n_X$ общее число «сильных» связей $S_\alpha(x,y)= 1$ в 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 $N_\alpha^+ = \sum_{k=1}^{u}\left(\sum_{x,y\in A_k} S_\alpha(x,y)\right) - n_X$ - число внутри-кластерных сильных связей отношения S_α. Тогда $N_\alpha^- = N_\alpha - N_\alpha^+$ равно числу межкластерных сильных связей. Обозначим $N_E = \sum_{x,y\in X} E_1(x,y) - n_X$ 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:

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

Третья глава "Эволюционные процедуры визуализации многомерных данных" описывает процедуры двух- и трехмерной визуализации данных, основанные на генетических алгоритмах, и метод гибридной кластеризации с двух- и трехмерной визуализацией результатов кластеризации данных.

Задача визуализации рассматривается как задача минимизации искажений исходных расстояний между объектами, при их представлении в двух- или трехмерном пространстве. В работе предложено применение генетического алгоритма оптимизации. Схема (1) имеет вид: $HC(S) = \langle \varnothing; P\rangle$, где P задает координаты объектов в двух- или трехмерном пространстве.

Рассмотрим процедуру двумерной визуализации данных. Предположим, что R – расстояние между объектами n-мерного пространства признаков M, $n>3$. Ищем двумерное представление этих объектов. Генерируем начальную матрицу координат объектов в двумерном пространстве с осями X и Y. Основываясь на этой матрице, с помощью стандартной процедуры оптимизации определяем матрицу координат P объектов с минимальной ошибкой аппроксимации начальной матрицы D расстояниями матрицы R, вычисленными для матрицы координат P. Обычно полученное представление дает локальный оптимум. Затем определяем два объекта $a(a_x, a_y)$ и $b(b_x, b_y)$ из множества M, с максимальным значением $R(a, b)$. Систему координат $\langle X, Y\rangle$ перемещаем и поворачиваем так, чтобы центр «новой» системы координат $\langle X', Y'\rangle$ находился в точке a, а точка b располагалась на оси X.

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.

Затем выбираем объект c с максимальным по абсолютной величине значением y_c, в зависимости от значения координаты по оси Y' которого осуществляется зеркальное отображение всех объектов относительно оси Y'. Объекты a, b, c будут опорными элементами для всех последующих матриц координат объектов. Полученную матрицу координат P^* объектов из M назовем решением и ошибку аппроксимации матрицы D матрицей расстояний R, вычисленной на основе матрицы координат P^*, назовем ошибкой решения.

Случайным образом генерируем множество m матриц начальных координат объектов в двумерном пространстве и для каждой из них вычисляем ошибку аппроксимации. Полученное множество матриц назовем популяцией. Наилучшие q матриц с наименьшей ошибкой аппроксимации, выбранных из популяции, назовем элитой.

Для каждой матрицы из элиты применяем перемещение и поворот относительно опорных точек a, b, c так, чтобы их расположение было таким же, как и в матрице P^*. Полученные решения затем используем для генерации новых решений, называемых потомками. Потомков получаем 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 бесконечно малых величин.

Новая популяция получается с помощью добавления к старой элите: 1) потомков, полученных применением операции рекомбинации к старой элите; 2) потомков, полученных из элиты применением операции мутации; 3) потомков, полученных с помощью применения мутации после применения операции рекомбинации к старой элите.

Для новых элементов элиты, определяющих матрицы координат объектов в двумерном пространстве, вычисляются матрицы расстояний и ошибка аппроксимации исходной матрицы расстояний в n-мерном пространстве. 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. Генерация популяции повторяется заданное число раз или до тех пор, пока ошибка аппроксимации не будет меньше заданного значения.

Поиск трехмерного представления объектов осуществляется по аналогичной схеме. Отличие данного метода состоит в том, что в данном случае

выбираются четыре опорные точки и зеркальное отображение при необходимости осуществляется относительно двух осей.

Рассмотрим структуру генетического алгоритма для двумерного представления результатов классификации.

Рассмотрим первый этап алгоритма. Если кластер состоит из двух объектов, то эти объекты получают координаты *(0, 0)* и *(d, 0)*, соответственно, где *d* есть расстояние между этими объектами в *n*-мерном пространстве. Если кластер состоит более чем из двух объектов, то к множеству этих объектов применяется генетический алгоритм, описанный выше.

На втором этапе расстояния между двумя классами, которые объединяются вместе в кластерной процедуре, оптимизируются следующим образом. Координаты объектов из одного из них остаются неизменными. Этот класс будем называть фиксированным классом. Второй класс будем перемещать, так чтобы расстояния между объектами внутри этого класса, полученные на первом этапе, не изменялись. Этот класс будем называть перемещаемым классом. Координаты объектов из этого класса изменим 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 двумерном пространстве до перемещения класса, X_{new}, Y_{new} – координаты объектов после перемещения, dX, dY – величины сдвига объектов вдоль осей X и Y, A – угол поворота класса вокруг начала координат. Строка параметров *(dX, dY, A)* представлена как элемент популяции в генетическом алгоритме.

Теперь рассмотрим случай трехмерного представления объектов. Поиск трехмерного представления осуществляется по аналогичной схеме, с тем отличием, что координаты перемещаемого класса вычисляются по следующим формулам:

$$X_{new} = X^* \cos(A) + Y^* \sin(A) + dX,$$
$$Y' = -X^* \sin(A) + Y^* \cos(A) + dY,$$
$$Y_{new} = Y' \cos(A) + Z^* \sin(A) + dY,$$
$$Z_{new} = -Y' \sin(A) + Z^* \cos(A) + dZ,$$

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. Строка параметров *(dX, dY, dZ, A)* рассматривается как элемент популяции в генетическом алгоритме.

Четвертая глава "Описание комплекса программ гибридной кластеризации" содержит описание комплекса программ, реализующего разработанные методы, и его применение к задачам гибридной кластеризации временных рядов на основе меры ассоциаций локальных трендов временных рядов. Ком-

плекс программ гибридной кластеризации реализован в пакете Matlab. На рис 1. представлена совокупность методов, реализованных в этом пакете (обозначены М1-М7).

Разбиение исходного множества объектов на классы осуществляется инвариантными процедурами реляционной иерархической кластеризации (М1, М2), описанными в главе 2. Для представления исходного множества объектов в пространстве размерности два или три используются эволюционные процедуры двух- и трехмерной визуализации данных (М3, М4) из разделов 3.2 и 3.3. Метод гибридной кластеризации с визуализацией сильных связей между объектами (М5), из раздела 2.6, позволяет анализировать отклонения полученной кластеризации от структуры данных. Эволюционные процедуры двух- и трехмерной визуализации результатов кластеризации (М6, М7) из разделов 3.4 и 3.5 позволяют дополнить кластеризацию объектов визуализацией их взаимного расположения.

Методы гибридной кластеризации применялись к анализу взаимодействия нефтяных скважин одного из Мексиканских месторождений. Рассматривались 9 скважин, расположенных на общей территории, из которых 2 были нагнетающими.

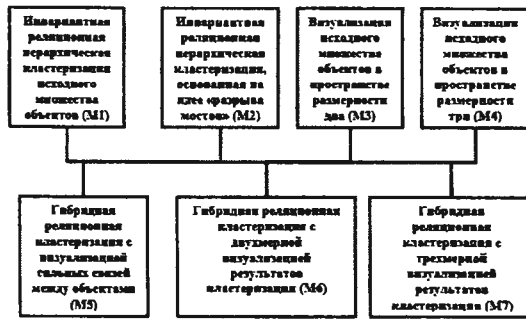

Рис. 1

Анализ проводился на основе временных рядов объемов помесячной добычи нефти и сопутствующей воды, а также объемов нагнетания воды за 32 месяца, начиная с января 2004 года. В качестве меры ассоциаций временных рядов использовалась мера

$$AM(y,x) = \max(AF_K(y,x)) = \max_{k\in K}(\cos s_k(y,x)),$$ где $K = \{2,3\}$ задает

размер скользящих окон. На Рис. 2 приведены графики временных рядов, где 1i и 7i обозначают временные ряды, соответствующие нагнетающим скважинам (инъекторы), k_p и $k_в$ означают, соответсвенно, временные ряды добычи нефти и сопутствующей воды в скважине k. Рис.3 дает графическое представление найденного ассоциативного графа между временными рядами,

для значения порога 0,468. Это значение равно величине ассоциации между нагнетающей скважиной 1i и добывающей скважиной 2p. Выбор такого порога для представления ассоциативного графа дает нетривиальное разбиение графа на связные компоненты и позволяет выявить ассоциации между нагнетающими и добывающими скважинами. Вершины ассоциативной сети расположены так, чтобы минимизировать пересечения дуг. В соответствии с этим порядком приведены также временные ряды на Рис.2, что позволяет увидеть сходство между формой временных рядов с высоким значением ассоциации локальных трендов (например между 6p и 6a, 3p и 3a, 7i и 6a).

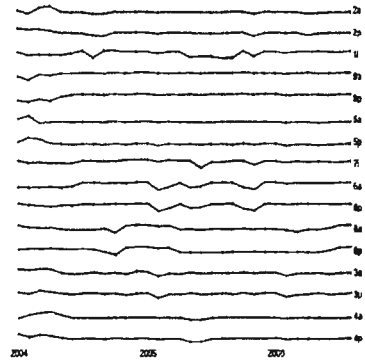

Рис. 2

На Рис. 4 представлены результаты гибридной кластеризации скважин на основе меры ассоциаций локальных трендов. Две скважины X и Y считаются взаимосвязанными, если существует значимая ассоциативная связь между хотя бы одной парой временных рядов (Xa, Ya), (Xa, Yp), (Xp,Ya), (Xp, Yp). Получено следующее разбиение скважин на 5 кластеров: {1i},{2,3,4,9},{5},{6,7i},{8}. На Рис. 4 одиночные кластеры помечены кружком, а скважины из кластеров {2,3,4,9} и {6,7i} квадратом и ромбом соответственно. Значимые ассоциативные связи между скважинами (5, 3 и 1i, 2), принадлежащим разным кластерам, показаны пунктиром. Интерпретация полученных результатов позволяет судить о характере взаимодействия между скважинами рассматриваемого месторождения. В частности, установлено хорошее взаимодействие между нагнетающей скважиной 7 и добывающей скважиной 6, объединенными в один кластер, что может объясняться высокой проницаемостью пластов в окрестности этих скважин. Отсутствие значимого взаимодействия между скважиной 8, попавшей в одиночный кластер, и другими скважинами, может характеризоваться разными причинами, и ввиду этого служить основой для дополнительного анализа свойств пород вокруг этих скважин. Объединение скважин 2, 3, 4, 9 в один кластер также может

служить основой для выдвижения гипотез о причинах значимого взаимодействия между скважинами: высокой проницаемости пород в окрестности этих скважин, наличия трещин или сдвигов пластов в окрестности этих скважин, и др. Результаты гибридной кластеризации дают основы для выдвижения различных гипотез о характере взаимодействия между скважинами и в результате этого служить основанием для проведения дополнительных специфических исследований на скважинах и в месторождении, а также для изменения режимов закачки воды в скважины.

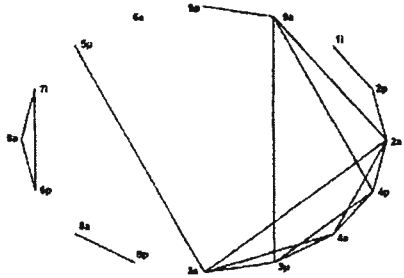

Рис.3

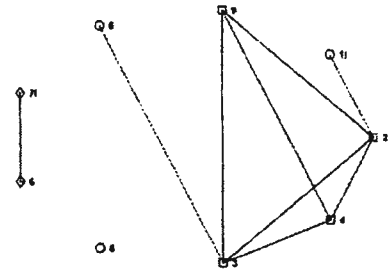

Рис.4

Метод сочетания иерархической кластеризации с одномерной визуализацией данных по некоторому показателю был использован в задаче анализа уровня потребления электроэнергии странами бывшего Советского Союза за 1992-2004 [World Energy Consumption in Standard U.S. Physical Units. http://www.eia.doe.gov/iea/wec.html.]. Рассмотрим применение метода гибридной кластеризации для анализа данных по потреблению гидроэлектроэнергии (World Net Hydroelectric Power Consumption). На Рис. 5 представлены результаты иерархической кластеризации временных рядов потребления гидроэлектроэнергии методом M1. В качестве расстояния между временными рядами использовалось Евклидово расстояние между ними.

Рис. 5

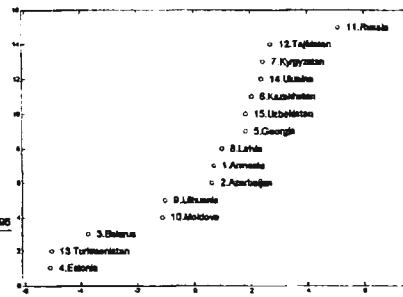

Рис. 6

Для каждого кластера приведено значение среднего потребления электроэнергии странами из этого кластера за рассматриваемый период. Как видно из построенной дендрограммы, страны сгруппированы в кластеры не только по сходству расстояний между соответствующими временными рядами, но также и по величине среднего потребления электроэнергии. Это дает основание проанализировать упорядочение стран по этому показателю, используя его как показатель P в одномерной гибридной кластеризации. Результаты представлены на Рис. 6. Как видно из Рис. 5 и 6 сочетание иерархической кластеризации с одномерным упорядочением данных по выбранному показателю позволяет более детально исследовать и визуализировать структуру исследуемых данных.

В заключении приведены основные результаты исследований, представленные в диссертационной работе.

Основные результаты работы и выводы

1. Разработана общая схема гибридной реляционной кластеризации.
2. Получено обоснование реляционной схемы инвариантных иерархических кластерных процедур и разработана новая инвариантная иерархическая кластериая процедура этой схемы.
3. Разработаны методы построения моделей данных в виде инвариантных кластеров сходных объектов, сетей сходства и их визуализация в двухмерном и трехмерном пространствах.
4. Разработан пакет программ гибридной кластеризации данных в среде Matlab.
5. Разработанные в работе алгоритмы кластеризации являются инвариантными относительно монотонного преобразования значений сходства и исходной нумерации объектов (глава 1), и параметрическими, что позволяет решать широкий класс задач анализа данных.
6. Алгоритмы визуализации позволяют представить многомерные данные в пространстве размерности два или три с ошибкой аппроксимации исходной матрицы расстояний матрицей расстояний двумерного или трехмерного представления объектов, меньшей чем ошибка аппроксимации представления, полученного другими методами (например, методами неметрического шкалирования, методами оптимизации (разделы 3.2 и 3.3)).

Основные результаты работы представлены в публикациях

Публикации в рецензируемых научных журналах и изданиях Перечня ВАК

1. Климова А.С. Гибридная кластеризация на основе реляционной схемы инвариантных кластерных процедур/Батыршин И.З., Климова А.С.//Вестник Тверского государственного университета. – 2007. - вып. 7 - С. 27-42. - Серия: Прикладная математика.
2. Климова А.С. Методы гибридной реляционной кластеризации в анализе среднего потребления электроэнергии странами бывшего Советского Союза/Климова А.С.//Известия высших учебных заведений. Проблемы энергетики. – 2008. - вып. 5-6 - С. 124-127.
3. Климова А.С. Применение методов гибридной кластеризации к анализу нефтяных скважин/ Климова А.С., Батыршин И.З., Шайдуллина Н.К.//Вестник Казанского технологического университета. – Казань, 2013.- Т. 8. – С. 241 – 245.

Публикации в журналах и сборниках научных работ академических и центральных изданий

4. Батыршин И.З. О визуализации результатов классификации/ Батыршин И.З., Климова А.С//Труды Международн. научно-технических конференций «Интеллектуальные системы (IEEE AIS'03)» и «Интеллектуальные САПР (CAD-2003)» - Москва: Физматлит, 2003. - Т. 2 - С. 172-177.
5. Батыршин И.З. Эволюционные процедуры иерархической двухмерной визуализации данных/ Батыршин И.З., Климова А.С//В сб.: Исследования по информатике. Институт проблем информатики АН РТ. - Казань, Отечество, 2004. - N 7 - С. 119–124.
6. Батыршин И.З Инвариантные кластерные процедуры, основанные на нечетком отношении сходств/ Батыршин И.З., Климова А.С//В кн.: Интегрированные модели и мягкие вычисления в искусственном интеллекте. Сб. научн. трудов III-го Межд. научно-практич. семинара.– Москва: Физматлит, 2005 - С. 119 - 125.
7. Батыршин И.З. Гибридная реляционная кластеризация и визуализация данных/ Батыршин И.З., Климова А.С//Труды Всеросс. научн. конф. по нечетким системам и мягким вычислениям, НСМВ-2006. - Москва: Физматлит, 2006 - С. 193-209.
8. Батыршин И.З. Преобразование скользящих аппроксимаций и ассоциативные сети в сравнительном анализе статистических рядов динамики/ Батыршин И.З., Шереметов Л.Б., Панова А.М., Климова А.С.//В сб.: Исследования по информатике. Институт проблем информатики АН РТ - Казань, Отечество, 2006. - вып. 11 - С. 35-48.
9. Батыршин И.З. Анализ взаимодействия нефтяных скважин на основе гибридной кластеризации временных рядов продуктивности скважин/ Ба-

10—

тыршин И.З., Кошульски А., Шереметов Л.Б., Климова А.С., Панова А.М.//Нечёткие системы и мягкие вычисления. – 2007. - том 2, №4 - С. 63-73.

Публикации в трудах международных и Российских конференций

10. Батыршин И.З. О визуализации многомерных данных/ Батыршин И.З., Климова А.С.//в кн.: Проблемы информатики в образовании, управлении, экономике и технике. II Всероссийская научно-технич. конф. - Пенза, 2002. - С. 156 - 158.
11. Batyrshin I.Z. New invariant relational clustering procedures/ Batyrshin I.Z., Klimova A.S.//In: Proceedings of East West Fuzzy Colloquium 2002, 10th Zittau Fuzzy Colloquium - Zittau, Germany, 2002 - P. 264 – 269.
12. Batyrshin I.Z. On two dimensional visualization of hierarchical clustering/Batyrshin I.Z., Klimova A.S., Sheremetov L.B.//In: IEEE International Conference on Computational Cybernetics, ICCC 2003. - Siofok, Hungary, 2003. - P. 337 – 342.
13. Батыршин И.З. Трехмерная визуализация результатов кластеризации/Батыршин И.З., Климова А.С.//XIV Международная научно-техническая конференция «Математические методы и Пенза, 2004. - С. 272-275.
14. Klimova A.S. Evolutionary procedures of visualization of multidimensional data./Klimova A.S.//In: FSSCEF 2004, Proc. Intern. Conf. on Fuzzy Sets and Soft Computing in Economics and Finance - St. Petersburg, Russia, 2004. - vol. I - P. 130– 139.
15. Batyrshin I.Z. On general scheme of invariant clustering procedures based on fuzzy similarity relation./Batyrshin I.Z., Rudas T., Klimova A.S.//In: FSSCEF 2004, Proc. Internat. Conf. on Fuzzy Sets and Soft Computing in Economics and Finance - St. Petersburg, Russia, 2004. - vol. I - P. 122-129.
16. Batyrshin I.Z. Combining local trend association network and clustering in visualization of relationships in time series data bases./ Batyrshin I.Z., Klimova A.S., Sheremetov L.B., Velasco-Hernandez J.X.// In: FSSCEF 2006, Proc. Intern. Conf. Fuzzy Sets and Soft Computing in Economics and Finance - St. Petersburg, Russia, 2006. - P. 242-251.
17. Batyrshin I.Z. Hybrid clustering of time series/ Batyrshin I.Z., Sheremetov L.B., Velasco-Hernandez J.X., Klimova A.S.//In: East West Fuzzy Colloquium 2006, 13th Zittau Fuzzy Colloquium - HS Zittau/Gorlitz, Germany, 2006. - P. 140-146.

Соискатель А.С. Климова

Заказ №167 Тираж 100 экз.

Отпечатано в типографии «Деловая полиграфия», 420111, г.Казань, ул.М.Межлаука, 6