Методы интеллектуальной обработки качественных данных*

***И. В. Покровская*[1], *М. Д. Гольдовская*[1], *Ю. А. Дорофеюк*[1,2], *Н. Е. Киселева*[1]**
ivp750@mail.ru
[1]Москва, Институт проблем управления им. В. А. Трапезникова РАН (ИПУ РАН)
[2]Москва, Научно-исследовательский университет Высшая школа экономики (НИУ ВШЭ)

Исследуются задачи интеллектуальной обработки качественных данных. Рассмотрено два примера постановок задач и алгоритмов обработки качественных данных, представленных в виде признаков долевого типа и эмпирических графов большой размерности. Разработана методика интеллектуальной обработки признаков долевого типа, проведено тестирование на реальных данных. Исследованы возможности точного и приближенного представления графа большой размерности через его описание. На задачу агрегирования распространен оптимизационный подход к построению размытой классификации. В рамках структурно-классификационной методологии интеллектуального анализа сложно организованных данных разработаны оригинальные алгоритмы решения задачи обработки информации с помощью агрегирования графов большой размерности.

Ключевые слова: *качественные данные; интеллектуальный анализ данных; экспериментальные графы большой размерности; параметры долевого типа; размытая упорядоченная классификация*

Intellectual methods of processing qualitative data*

***I. V. Pokrovskaya*[1], *M. D. Goldovskaya*[1], *J. A. Dorofeyuk*[1,2] *and N. E. Kiseleva*[1]**
[1]ICS RAS; [2]SIU HSE

Intellectual processing of qualitative data problem is investigated. Two examples of the states of the problems and algorithms for qualitative data processing, presented in the form of the equity-type characteristics and the large dimension empirical graphs, are considered. The methodology of data mining (group) characteristics of equity-type (equivalent blurred classifications) is developed; this method was tested on real data. The possibilities of the exact and approximate representation of the large dimension graph through its description are studied. The optimization approach to the construction of the fuzzy classification is distributed to the problem of aggregation. In the framework of the structural-classification mining methodology of complex data, the original information processing algorithms by large dimension graphs aggregation methods are developed.

Keywords: *qualitative data; data mining; large dimension experimental graphs; options equity-type; fuzzy ordered classification*

Введение

В последнее время существенно возрос интерес к исследованию и моделированию слабо формализованных социально-экономических систем. Для многих систем такого рода

*Работа выполнена при частичной финансовой поддержке РФФИ, гранты № 14-07-00463-а, № 13-07-00992-а и № 12-07-00540-а.

значительная часть исходных параметров, описывающих состояние системы, имеют качественную природу. К таким параметрам относятся, например, бинарные, ранговые и номинальные признаки. Очевидно, что решение стандартных задач моделирования подобных систем, например, идентификации или структурного описания, невозможно получить методами, использующими только количественные (числовые) показатели. Здесь возможны два пути выхода из создавшейся ситуации. Первый путь — это разработка своеобразных преобразователей, позволяющих использовать алгоритмическую базу методов моделирования для количественных признаков. Типичным примером такого случая является модификация процедур расчета расстояний между объектами в многомерном пространстве бинарных признаков при решении задач структуризации множества исследуемых объектов. А именно, предлагается для таких расчетов вместо евклидовой метрики использовать метрику Хэмминга. Легко также преобразуется процедура расчета расстояний для случая, когда часть признаков — числовые, а другая часть — бинарные. Второй путь — это разработка принципиально новых алгоритмов решения стандартных задач моделирования и анализа систем, описываемых качественными параметрами. Здесь часто определяющую роль играет схема представления исходных данных или, другими словами, качественная модель порождения данных. Примером реализации этого пути является представление многих социологических данных в виде направленных бинарных или взвешенных графов. В этом случае задача структуризации, например, сводится к известной задаче декомпозиции графа на подграфы по степени связности, не требующей подсчета в явном виде каких-либо расстояний. В настоящей работе рассмотрены две задачи, на примере которых продемонстрированы особенности реализации первого и второго пути. Первая задача — структуризация специального типа качественных параметров — параметров долевого типа. Вторая задача — исследование структуры множества взаимосвязанных объектов многоагентной системы, когда сама система и взаимосвязь входящих в нее объектов характеризуется ориентированным не взвешенным графом большой размерности.

Группировка качественных параметров долевого типа

Под структуризацией (группировкой) некоторого множества параметров имеется в виду разбиение его на группы «близких», «взаимозависимых» параметров на основе выбранной меры близости (зависимости) между параметрами. Так построены алгоритмы экстремальной группировки параметров, измеряемых в количественных, ранговых и номинальных шкалах [1]. В работе предлагается подход к решению этой задачи для особого вида параметров — качественных параметров долевого типа.

Задача структуризации множества исходных параметров. Опыт использования алгоритмов структурно-классификационного анализа показывает, что классификация по всем исходным параметрам далеко не всегда приводит к желаемым результатам. Действительно, при сравнительно небольших выборках экспериментальных наблюдений и наличии помех (ошибки в определении значений параметров, сознательное искажение информации и т. д.) использование для классификации большого числа входных параметров приводит к сильному «перемешиванию» классов, а сами классы при этом плохо поддаются интерпретации. По этой причине классификацию целесообразно проводить не в исходном пространстве, а в пространстве наиболее существенных (информативных) параметров, имеющем значительно меньшую размерность. Для структуризации параметров обычно используются алгоритмы экстремальной группировки параметров [2]. При этом необходимо определить, нужна ли группировка с фоновой группой или без нее (т. е. отсекать или нет сильно шумящие параметры) [3]. Результатом группировки являются группы

параметров и факторы — обобщенные характеристики групп. На основе результатов группировки строятся интегральные показатели исследуемой структуры. В качестве таковых выбираются либо сами факторы, либо параметры в определенном смысле ближайшие к факторам. Основное условие — они должны быть легко интерпретируемы. Для удобства использования интегральных показателей по каждому из них делается одномерная классификация объектов. Благодаря этому интегральный показатель преобразуется в качественный, так как его значения можно качественно характеризовать в терминах типа «низкие», «средние», «высокие».

Другое применение метода экстремальной группировки — выбор информативных параметров для структуризации на последующих этапах. В качестве набора информативных параметров выбирается либо набор факторов, либо набор, в который входят один или небольшое число параметров из каждой группы экстремальной группировки. Обычно окончательное решение о выборе информативных параметров производится экспертом-пользователем [4].

Структуризация результатов классификации. Практически все алгоритмы структурно-классификационного анализа содержат свободные параметры, значения которых трудно выбрать заранее из теоретических соображений. Кроме того, эти алгоритмы находят лишь локальный экстремум соответствующего критерия качества структуризации, поэтому результаты их работы зависят от начальных условий (начального разбиения объектов на классы или параметров на группы). В связи с этим при решении практических задач свободные параметры алгоритмов, начальные условия, а часто и состав переменных, образующих исходное пространство, варьируются в широких пределах. Это приводит к тому, что в результате получается достаточно обширное множество различных вариантов классификации. Число классификаций часто оказывается столь большим, что для их анализа приходится применять компьютерные методы, вводя меру близости между классификациями и разбивая их на группы «похожих» классификаций. Легко показать, что размытые классификации можно рассматривать как признаки долевого типа. Следовательно, можно для структуризации множества классификаций использовать методы группировки признаков долевого типа.

Признаки долевого типа. Рассмотрим некоторый «агрегированный объект» (например, город), включающий множество «индивидуальных объектов» (например, жителей города). Пусть каждый житель характеризуется некоторым качественным показателем, измеряемым в номинальной шкале (например, уровнем образования, имеющим три значения: ниже среднего, среднее, высшее). Тогда для города этот же показатель, уровень образования, естественно характеризовать набором из трех чисел, показывающих, какую долю его населения составляют жители с соответствующими уровнями образования.

Рассмотрим множество из n агрегированных объектов, каждый из которых состоит из ряда индивидуальных объектов. Будем считать, что индивидуальные объекты описываются двумя параметрами x и y, измеренными в номинальных шкалах. Пусть параметр x принимает одно из значений $(x_1, \ldots, x_k)$, а y — одно из значений $(y_1, \ldots, y_m)$.

Рассмотрим соответствующие параметры долевого типа и , описывающие агрегированные объекты. Их значения для t-го объекта представляют собой векторы $A_t = = (\alpha_t^{(1)}, \ldots, \alpha_t^{(k)})$ и $B_t = (\beta_t^{(1)}, \ldots, \beta_t^{(m)})$. Здесь $\alpha_t^{(i)}$ — доля индивидуальных объектов, принадлежащих t-му агрегированному объекту, для которых $x = x_i$, а $\beta_t^{(j)}$ — доля индивидуальных объектов, принадлежащих t-му агрегированному объекту, для которых $y = y_j$.

Вначале предположим, что все индивидуальные данные нам известны. Тогда, кроме этих параметров, мы можем подсчитать параметр $G_t = (g_t^{(i,j)}, i = \overline{1,k},\ j = \overline{1,m})$, где $g_t^{(i,j)}$ — доля индивидуальных объектов, принадлежащих t-му агрегированному объекту, для которых $x = x_i$, а $y = y_j$.

Если интерпретировать долю $\alpha_t^{(i)}$ как вероятность i-й градации параметра x для t-го агрегированного объекта, а $\beta_t^{(j)}$ как вероятность j-й градации параметра y, то $g_t^{(i,j)}$ интерпретируется как совместная вероятность i-й градации параметра x и j-й градации параметра y.

Введем матрицы условных вероятностей

$$Q_t^{(i,j)} = P_t(y = y_j \,|\, x = x_i) = \frac{g_t^{(i,j)}}{\alpha_t^{(i)}},\ R_t^{(i,j)} = P_t(x = x_i \,|\, y = y_j) = \frac{g_t^{(i,j)}}{\beta_t^{(j)}},\quad i = \overline{1,k},\ j = \overline{1,m}.$$

Справедливы соотношения:

$$\beta_t^{(j)} = \sum_{i=1}^{k} Q_t^{(i,j)} \alpha_t^{(i)};\ \alpha_t^{(i)} = \sum_{j=1}^{m} R_t^{(i,j)} \beta_t^{(j)};\ i = \overline{1,k},\ j = \overline{1,m}. \tag{1}$$

Матрицы Q_t и R_t отражают вероятностную зависимость между долевыми параметрами A и B. Однако во многих практических задачах индивидуальные данные недоступны, имеются только значения долевых показателей A и B. В этом случае показатель G и матрицы Q_t и R_t напрямую подсчитать нельзя, тогда возникает следующая задача: восстановить матрицы Q_t и R_t по параметрам A и B.

Алгоритм восстановления матриц. Для решения указанной задачи сделаем следующее допущение: матрицы условных вероятностей Q_t и R_t не зависят от агрегированного объекта, т. е. $Q_t = Q$ и $R_t = R$. Рассмотрим алгоритм восстановления матрицы Q (восстановление матрицы R производится аналогично).

Будем считать, что соотношения (1) выполняются не точно, а с некоторой случайной погрешностью, имеющей характер аддитивного шума, в частности:

$$\beta_t^{(j)} = \sum_{i=1}^{k} Q_t^{(i,j)} \alpha_t^{(i)} + \varepsilon_t^{(j)}, i = \overline{1,k},\ j = \overline{1,m}. \tag{2}$$

Уравнение (2) имеет вид линейной регрессии и отличается от нее только ограничением:

$$\sum_{j=1}^{m} Q^{(i,j)} = 1,\ i = \overline{1,k};\ Q^{(i,j)} \geqslant 0,\ i = \overline{1,k},\ j = \overline{1,m}.$$

Оценочная матрица $\widehat{Q}$, минимизирующая суммарную дисперсию случайной погрешности, находится стандартными процедурами квадратичного программирования.

Качество линейной регрессионной модели обычно измеряют коэффициентом детерминации. В рассматриваемой задаче это:

$$D(\widehat{Q}) = \frac{T(Y) - F(\widehat{Q})}{T(Y)},$$

где $T(Y)$ — сумма квадратов отклонений компонент вектора B от своих средних значений, а $F(\widehat{Q})$ — сумма квадратов невязок в (2). Преимущество коэффициента детерминации по сравнению с некоторыми другими мерами качества модели состоит в том, что он

Таблица 1. Значения параметра X

X	$\mathbf{A_1}$	$\mathbf{A_2}$	$\mathbf{A_3}$	$\mathbf{A_4}$	$\mathbf{A_5}$	$\mathbf{A_6}$	$\mathbf{A_7}$	$\mathbf{A_8}$	$\mathbf{A_9}$	$\mathbf{A_{10}}$
α_1	0,1	0,3	0,0	0,5	0,2	0,4	0,3	0,0	1,0	0,4
α_2	0,3	0,4	0,2	0,3	0,1	0,4	0,1	0,3	0,0	0,2
α_3	0,2	0,2	0,4	0,1	0,4	0,0	0,2	0,4	0,0	0,3
α_4	0,4	0,1	0,4	0,1	0,3	0,2	0,4	0,3	0,0	0,1
X	$\mathbf{A_{11}}$	$\mathbf{A_{12}}$	$\mathbf{A_{13}}$	$\mathbf{A_{14}}$	$\mathbf{A_{15}}$	$\mathbf{A_{16}}$	$\mathbf{A_{17}}$	$\mathbf{A_{18}}$	$\mathbf{A_{19}}$	$\mathbf{A_{20}}$
α_1	0,2	0,7	0,1	0,3	0,8	0,9	0,1	0,2	0,4	0,1
α_2	0,1	0,1	0,1	0,5	0,1	0,0	0,6	0,1	0,2	0,6
α_3	0,1	0,0	0,4	0,1	0,0	0,0	0,2	0,6	0,0	0,1
α_4	0,6	0,2	0,4	0,1	0,3	0,1	0,1	0,1	0,4	0,2

достаточно нагляден и легко интерпретируем — меняется от 0 до 1, и чем он ближе к 1, тем лучше модель.

Матрица R восстанавливается аналогично. Однако в общем случае $D(\widehat{Q}) \neq D(\widehat{R})$, т. е. зависимость между долевыми параметрами A и B несимметрична (A может зависеть от B сильнее, чем B от A, и наоборот). Поэтому при структуризации множества долевых параметров в качестве меры зависимости целесообразно использовать сумму коэффициентов детерминации.

Связь с размытыми классификациями. Напомним, что четкой классификацией множества объектов на k классов называется такое разбиение объектов на классы, что каждый объект отнесен к одному и только одному классу. Будем приписывать каждому объекту номер класса, в который он попал. Тогда получается, что на множестве объектов задан номинальный признак — номер класса. Наоборот, если есть номинальный признак, то множество объектов разбивается по нему на классы эквивалентности. Следовательно, понятия номинальный признак и четкая классификация можно интерпретировать друг через друга. Соответственно, понятие рангового признака можно интерпретировать как упорядоченную классификацию, т. е. такую, у которой классы упорядочены.

Наряду с четкими классификациями широкое применение получили размытые классификации, у которых объекты с разными весами могут принадлежать сразу нескольким классам. Размытая классификация задается через вектор-функцию принадлежностей. Заметим, что формальный объект такого рода можно интерпретировать как параметр долевого типа.

Таким образом, номинальные признаки на индивидуальных данных можно интерпретировать как четкие классификации, а параметры долевого типа на агрегированных данных как размытые классификации.

Компьютерное моделирование. Для проверки работоспособности предложенной методики было проведено компьютерное моделирование как на модельных, так и на реальных данных, содержащих данные о $g_t^{(i,j)}$. Результаты моделирования показывают, что оценки матриц условных вероятностей получаются близкими к реальным матрицам, а при отсутствии шума $\varepsilon_t^{(j)}$ совпадают с ними.

Компьютерное моделирование на модельных данных. Исследовались 20 агрегированных объектов. На них был построен параметр долевого типа X, состоящий из четырех градаций, значения параметра X приведены в табл. 1.

Таблица 2. Матрица условных вероятностей Q

Q	β_1	β_2	β_3
α_1	$P_{11} = 0{,}1$	$P_{21} = 0{,}8$	$P_{31} = 0{,}1$
α_2	$P_{12} = 0{,}3$	$P_{22} = 0{,}6$	$P_{32} = 0{,}1$
α_3	$P_{13} = 0{,}5$	$P_{23} = 0{,}4$	$P_{33} = 0{,}1$
α_4	$P_{14} = 0{,}4$	$P_{24} = 0{,}2$	$P_{34} = 0{,}4$

Таблица 3. Результаты экспериментов

δ	$F_1(\widehat{Q})$	$D_1(\widehat{Q})$	$\rho(Q, \widehat{Q})$
0,0	0,000	1,000	0,000
0,1	0,010	0,731	0,011
0,2	0,038	0,465	0,042
0,3	0,077	0,393	0,105

Значения параметра долевого типа Y, состоящего из трех градаций, рассчитывались по следующей схеме. Задана матрица условных вероятностей Q, приведенная в табл. 2.

По параметру X и матрице Q строились величины:

$$\tilde{\beta}_t^{(j)} = \max\left(0, \sum_{i=1}^{4} Q_t^{(i,j)} \alpha_t^{(i)} + \delta z_t^{(j)}\right), \ j = 1, 2, 3; \ t = \overline{1, 20}.$$

где δ — константа, задающая уровень шума; $z_t^{(j)}$ — величины, полученные датчиком случайных чисел, распределенных равномерно на отрезке $[-1; 1]$.

Наконец, значения компонент (градаций) параметра Y вычислялись по формуле:

$$\beta_t^{(j)} = \frac{\tilde{\beta}_t^{(j)}}{\sum_{l=1}^{4} \tilde{\beta}_t^{(l)}}.$$

Такая схема введения шума в зависимость параметра Y от параметра X гарантирует выполнение ограничений (2) и (3).

В эксперименте строилась оценка $\widehat{Q}$ для матрицы условных вероятностей Q для четырех разных уровней шума. Величина δ равнялась последовательно 0; 0,1; 0,2 и 0,3. Результаты эксперимента приведены в табл. 3.

В табл. 3 величина $\rho(\widehat{Q}, Q) = (\widehat{Q} - Q)^2$ является мерой отличия модельной матрицы $\widehat{Q}$, полученной при разных уровнях шума, от исходной матрицы Q.

Во-первых, следует отметить, что если шума нет, то исходная матрица восстанавливается точно. Во-вторых, при возрастании уровня шума возрастает значение $F_1(\widehat{Q})$ и падает значение коэффициента детерминации $D_1(\widehat{Q})$. В-третьих, следует отметить хорошую корреляцию между вторым и четвертым столбцами табл. 3, т. е. между $F_1(\widehat{Q})$ и $\rho(Q, \widehat{Q})$. Следовательно, величина $F_1(\widehat{Q})$ достаточно хорошо отражает качество оценивания матрицы Q.

Компьютерное моделирование на реальных данных. Для моделирования использовались данные переписи населения России [5]. Анализировались некоторые демографические показатели 77 регионов России. В качестве первого параметра долевого типа X рассматривалась степень урбанизации региона ($\alpha_t^{(1)}$ — доля городских жителей в t-м

Таблица 4. Значения элементов матрицы $\widehat{Q}$, минимизирующие критерий $F_1(\widehat{Q})$

Q	$P(\beta_1)$	$P(\beta_2)$	$P(\beta_3)$
α_1	$P_{11} = 0{,}156$	$P_{21} = 0{,}638$	$P_{31} = 0{,}206$
α_2	$P_{12} = 0{,}271$	$P_{22} = 0{,}542$	$P_{32} = 0{,}187$

Таблица 5. Значения элементов выборочной матрицы $P(Y/X)$

$Q = P(Y/X)$	$P(\beta_1)$	$P(\beta_2)$	$P(\beta_3)$
α_1	$P_{11} = 0{,}165$	$P_{21} = 0{,}631$	$P_{31} = 0{,}204$
α_2	$P_{12} = 0{,}221$	$P_{22} = 0{,}560$	$P_{32} = 0{,}218$

регионе, а $\alpha_t^{(2)}$ — доля сельских). В качестве второго долевого параметра Y рассматривалась возрастная структура населения соответствующего региона ($\beta_t^{(1)}$ — доля людей в t-м регионе, чей возраст меньше трудоспособного; $\beta_t^{(2)}$ — доля людей трудоспособного возраста; $\beta_t^{(3)}$ — доля людей старше трудоспособного возраста). При рассмотрении демографических данных доля населения в регионе от суммарной численности населения во всех рассматриваемых регионах является тем масштабирующим коэффициентом d_t, который используется в критерии $F_3(\widehat{Q})$.

В табл. 4 приведены значения элементов матрицы $\widehat{Q}$, минимизирующие критерий $F_1(\widehat{Q})$. Для этой матрицы $F_1(\widehat{Q}) = 0,0025$ и $D_1(\widehat{Q}) = 0,245$. Расчеты показывают, что коэффициент детерминации получился значимым.

Из данных переписи можно извлечь также данные о пересечении рассматриваемых параметров. Отметим, что в построении матрицы $\widehat{Q}$ эти данные не использовались, поэтому их можно рассматривать как тестовый материал для модели. По этим данным была построена выборочная матрица условных вероятностей параметра Y от параметра X. Значения ее элементов приведены в табл. 5.

Сравнение табл. 4 и 5 показывает, что матрица оценок $\widehat{Q}$ достаточно хорошо соответствует реальной матрице Q (в данном случае $P(Q, \widehat{Q}) = 0,0039$).

Агрегирование графов большой размерности

Пусть задан ориентированный невзвешенный граф большой размерности, полученный как результат экспериментального исследования группы взаимосвязанных объектов (например, некоторой многоагентной системы). Задача состоит в выявлении основных пучков дуг в этом графе, т. е. в выделении пар подмножеств множества вершин графа, таких, что из первого подмножества во второе идут почти все дуги. Особый интерес представляет случай, когда совокупность всех пучков можно рассматривать как некоторый малый граф, множество вершин которого является набором подмножеств множества вершин исходного графа. Набор подмножеств некоторого множества можно интерпретировать, как кластеризацию с перекрывающимися кластерами. Задача агрегирования исходного графа заключается в нахождении такой кластеризации множества вершин исходного графа и такого малого графа построенного на кластерах, которые в некотором смысле оптимально описывают исходный граф.

Формально задача ставится следующим образом. Обозначим исходный граф через G, множество его вершин через $X = \{x_1, ..., x_n\}$, а его матрицу смежности через $M(G) = \|m_{i,j}; i = \overline{1,n}; j = \overline{1,n}\|$.

Пусть $H = \{H_1, ..., H_r\}$ $(H_i \subseteq X)$ — некоторая кластеризация множества X с перекрывающимися кластерами. Такую кластеризацию можно задавать с помощью матрицы $B(H) = \| b_{pi} \|$, элемент b_{pi} которой, находящийся на пересечении p-ой строки и i-го столбца, равен 1, если i-я вершина принадлежит p-му кластеру, а в противном случае он равен 0. Пусть на H как на множестве вершин построен малый граф , матрицу смежности которого обозначим через $M(\Gamma) = \|\mu_{i,j}; i = \overline{1,r}; j = \overline{1,r}\|$. По кластеризации H и графу Γ строится аппроксимирующий граф $G(\Gamma, H)$ с помощью следующего **алгоритма построения графа** $G(\Gamma, H)$. Выбирается дуга графа Γ, пусть она для определенности идет из вершины H_p в вершину H_q. Затем в графе $G(\Gamma, H)$ проводятся дуги из всех элементов кластера H_p во все элементы кластера H_q. Такая процедура выполняется со всеми дугами графа Γ. Матрица смежности $M(G(\Gamma, H)) = \|\hat{m}_{i,j}\|$ графа $G(\Gamma, H)$ вычисляется по формуле $M(G(\Gamma, H)) = B(H)^T * M(\Gamma) * B(H)$. Здесь знак «$*$» означает булево произведение матриц. Отсюда следует, что элементы матрицы смежности $M(G(\Gamma, H))$ определяются следующим выражением $\hat{m}_{i,j} = \bigvee_{p,q=1}^{r} b_{pi}\mu_{pq}b_{qj}$. Возможны две постановки задачи агрегирования, которые далее рассмотрены более подробно.

Первая задача агрегирования — оптимальное сужение исходного графа. Для заданного исходного графа G найти граф Γ и соответствующую кластеризацию $H = \{H_1, ..., H_r\}$ минимального размера такие, что $G = G(\Gamma, H)$. Другими словами это постановка задачи точного представления графа G через граф меньшего размера. Будем его называть сужением графа G, а сужение с минимальным числом вершин будем называть оптимальным сужением графа G. Если рассматривать эту задачу без каких либо дополнительных ограничений, то она представляет собой NP-полную задачу.

Ограничим поиск сужений графа G его подграфами. Для этого будем рассматривать пары «граф Γ — кластеризация $H = \{H_1, ..., H_r\}$» только следующего специального вида: в каждом кластере H_p выделяется один из элементов x_{ip} (напомним, что такой элемент — одна из вершин графа G). Далее считается, что на графе Γ дуга из вершины H_p идет в вершину H_q тогда и только тогда, когда из вершины x_{ip} идет дуга в вершину x_{iq} на графе G. Другими словами подграф графа G, построенный на множестве вершин $\{x_{i_1}, ..., x_{i_r}\}$ изоморфен графу Γ. Такие сужения будем называть внутренними.

Для нахождения оптимального внутреннего сужения графа G построим матрицу $D(G)$, у которой n строк и $2n$ столбцов. Первые n столбцов этой матрицы соответствуют матрице $M(G)$, а последующие n столбцов соответствуют матрице $M(G)^T$ — транспонированная матрица $M(G)$, т. е. матрица $D(G)$ имеет следующую структуру: $D(G) = (M(G) : (M(G)^T))$. Среди строк этой матрицы выделим строки, которые нельзя представить в виде булевой суммы никакого набора других строк этой же матрицы. В содержательном смысле этот набор строк составляет «независимый базис». Пусть $\{x_{i_1}, ..., x_{i_r}\}$ — подмножество вершин графа G, соответствующих выделенным строкам.

Теорема 1. Граф в оптимальном внутреннем сужении графа G — изоморфен подграфу G, построенному на множестве вершин $\{x_{i_1}, ..., x_{i_r}\}$.

Вторая задача агрегирования — аппроксимационный подход к декомпозиции графа. Для заданного исходного графа G найти такие граф Γ и соответствующую кластеризацию $H = \{H_1, ..., H_r\}$ с фиксированным числом классов r, чтобы граф $G(\Gamma, H)$ был как можно ближе к графу G в смысле заранее выбранного критерия J: $J = J(G(\Gamma, H)) = (M(G) - M(G(\Gamma, H)))^2 = \sum_{i,j=1}^{n} [m_{ij} - \hat{m}_{ij}]^2$. Такая постановка реализует аппроксимационный подход к задаче декомпозиции графа G. Алгоритм решения этой задачи

Таблица 6. Матрица смежности модельного графа

$M(G)$	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}
x_1	1	1	0	1	0	1	1	1	1	0	0	0	0	0	0	0	1	1
x_2	1	1	1	0	0	1	0	1	0	0	0	1	0	1	1	0	1	0
x_3	1	1	1	1	1	1	1	1	0	1	0	1	0	0	0	0	1	1
x_4	1	1	1	1	0	1	0	0	0	0	0	0	0	0	0	1	1	1
x_5	0	1	1	1	1	1	0	0	1	0	0	0	0	1	0	0	0	0
x_6	1	0	0	0	1	1	1	1	1	0	0	1	1	0	0	0	1	1
x_7	1	1	0	1	0	0	1	1	1	0	0	1	0	1	0	1	0	0
x_8	1	1	0	0	0	0	1	1	1	0	0	0	0	0	1	1	0	0
x_9	1	0	0	0	0	0	1	1	1	0	0	0	0	0	1	0	1	1
x_{10}	0	1	1	0	0	1	0	0	1	1	0	0	1	0	0	0	1	1
x_{11}	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	0	0
x_{12}	0	1	1	0	0	1	1	1	1	0	1	1	0	1	0	1	1	0
x_{13}	1	0	1	1	1	1	0	1	0	1	0	1	1	0	0	0	0	1
x_{14}	0	1	0	0	0	1	1	1	0	0	1	1	0	1	1	1	0	0
x_{15}	0	1	1	0	0	0	0	0	1	0	1	0	0	1	1	0	0	0
x_{16}	0	0	1	1	0	1	0	0	0	0	1	1	0	1	1	1	1	0
x_{17}	1	1	0	0	0	1	0	1	1	0	0	0	0	0	0	1	1	1
x_{18}	1	1	0	1	0	1	0	0	1	0	0	0	1	0	0	0	1	1

состоит из последовательного выполнения двух процедур, которые выполняются на каждом шаге алгоритма:

1. При заданном графе Γ ищется такая кластеризация $H = \{H_1, ..., H_r\}$ с заданным числом классов r, которая позволяет уменьшить значение критерия J.
2. При фиксированном числе вершин графа Γ и при заданной кластеризации H ищется такое изменение структуры графа Γ, которое уменьшает значение критерия J.

В работе предложены эффективные алгоритмы оптимизации для обеих процедур. Предложенный подход является продвижением идей размытой классификации для анализа больших экспериментальных графов. Отметим, что кластеризация с перекрывающимися классами является важным частным случаем размытой классификации [3].

Компьютерное моделирование. Было проведено компьютерное моделирование алгоритма декомпозиции исходного графа (в рамках аппроксимационного подхода) на модельном материале. В качестве исходного материала была рассмотрена группа детей (18 чел.) одного из детских садов г. Москвы. Каждого ребенка попросили ответить на вопрос, с кем из детей из предъявленного списка он хочет играть. В результате был получен ориентированный граф G с матрицей смежности, приведенной в табл. 6.

Затем с помощью описанного выше алгоритма находились малый граф Γ и соответствующая кластеризация H на два класса (т. е. аппроксимирующий граф Γ состоял из двух вершин). В итоге было получено два следующих результата:

1. Граф Γ_1 состоит из двух несвязанных вершин с петлями; это означает, что есть два основных класса детей, желающих играть с детьми из «своего» класса (отметим, что классы пересекающиеся), этот результат отражен в табл. 7.
2. В графе Γ_2 есть петли у обеих вершин, и из второй вершины идет дуга в первую. Это означает, что дети хотят играть с детьми из «своего» класса, но, кроме того, все дети

Таблица 7. Структура графа Γ_1

$M(\Gamma_1)$	H_1	H_2	Кластеризация
H_1	1	0	$H_1 = \{x_2, x_7, x_{11}, x_{12}, x_{14}, x_{15}, x_{16}\}$
H_2	0	1	$H_2 = \{x_1, x_2, x_3, x_4, x_6, x_7, x_8, x_9, x_{17}, x_{18}\}$

Таблица 8. Структура графа Γ_2

$M(\Gamma_2)$	H_1	H_2	Кластеризация
H_1	1	0	$H_1 = \{x_1, x_2, x_6, x_7, x_8, x_9, x_{17}, x_{18}\}$
H_2	1	1	$H_2 = \{x_2, x_3, x_4, x_7, x_{12}, x_{14}, x_{18}\}$

из второго класса хотят играть со всеми детьми из первого класса. Данный результат дает более детальную картину взаимоотношений между детьми. Этот результат отражен в табл. 8.

Заключение

Разработаны новые методы исследования данных качественной природы: методика структурной обработки признаков долевого типа, а также алгоритмы точного и приближенного представления графа большой размерности через его описание. В настоящее время полученные результаты распространяются на случай взвешенных ориентированных графов динамического типа, широко используемых в мультиагентных системах управления. Разрабатываются также специализированные алгоритмы интеллектуальной обработки результатов многовариантного экспертного оценивания, масштабных социологических обследований и структурного анализа информационных потоков в Интернете.

Литература

[1] *Бауман Е. В.* Структуризация номинальных признаков в задаче экспертизы // *Экспертные оценки в задачах управления.* М.: ИПУ, 1982. С. 16–23.

[2] *Браверман Э. М., Мучник И. Б.* Структурные методы обработки эмпирических данных. М.: Наука, 1983. 464 с.

[3] *Дорофеюк А. А., Бауман Е. В., Дорофеюк Ю. А.* Методы интеллектуальной обработки информации на базе алгоритмов стохастической аппроксимации // *Математические методы распознавания образов: 15-я Междунар. конф.* М: МАКС ПРЕСС, 2011. С. 108–112.

[4] *Дорофеюк Ю. А., Киселева Н. Е., Покровская И. В.* Комплекс алгоритмов интеллектуального анализа данных для исследования функционирования сложных систем // *Управление развитием крупномасштабных систем MLSD'2013): Тр. 7-й Междунар. конф.* М.: ИПУ РАН, 2013. Т. 1. С. 220–232.

[5] Итоги переписи населения 2002 г. Т. 2: Возрастно-половой состав и состояние в браке. М.: Статистика России, 2004.

References

[1] *Bauman E. V.* 1982. Strukturizatsiya nominal'nykh priznakov v zadache ekspertizy. *Ekspertnye otsenki v zadachakh upravleniya.* M.: IPU Publ. 16–23. (In Russian.)

[2] *Braverman E. M., Muchnik I. B.* 1983. *Strukturnye metody obrabotki empiricheskikh dannykh.* M.: Nauka. 464 p. (In Russian.)

[3] *Dorofeyuk A. A., Bauman E. V., Dorofeyuk Yu. A.* 2011. Metody intellektual'noy obrabotki informatsii na baze algoritmov stokhasticheskoy approksimatsii. *Matematicheskie Metody Raspoznavaniya Obrazov: 15-aya Mezhdunar. Konf.: Sb. dokladov.* Moscow. 108–112. (In Russian.)

[4] *Dorofeyuk Yu. A., Kiseleva N. E., Pokrovskaya I. V.* 2013. Kompleks algoritmov intellektual'nogo analiza dannykh dlya issledovaniya funktsionirovaniya slozhnykh sistem. *Upravlenie razvitiem krupnomasshtabnykh sistem MLSD'2013): Tr. 7-y Mezhdunar. Konf.* 1:220–232. (In Russian.)

[5] *Itogi perepisi naseleniya 2002 g. Tom 2: Vozrastno-polovoy sostav i sostoyanie v brake.* M.: Statistika Rossii, 2004. (In Russian.)