Интеллектуальные системы и технологии

Методы интеллектуального анализа данных при исследовании сложных систем управления*

Ю.А. Дорофеюк, А.А. Дорофеюк, И.В. Покровская, А.Г. Спиро

Аннотация. Рассматривается задача исследования системы управления заданного множества объектов, каждый из которых характеризуется фиксированным (исходным) набором разнородных параметров. Для этого исследуется структура расположения управляемых объектов в пространстве информативных параметров. Для выявления такой структуры разработан комплекс алгоритмов интеллектуального анализа данных (ИАД), а также процедур экспертной коррекции. Проведен теоретический анализ алгоритмов ИАД, доказаны теоремы об их сходимости.

Ключевые слова: *интеллектуальный структурно-классификационный анализ данных, информативные параметры, начального разбиение, выбор числа классов, заполнение пропущенных наблюдений, процедуры экспертной коррекции.*

1. Введение

В последнее время для исследования сложных систем управления стали широко использоваться структурно-классификационные методы интеллектуального анализа данных, базирующиеся на алгоритмах классификационного анализа [1,4]. Это объясняется тем, что многие системы управления, в первую очередь организационно-административные, функционируют в условиях большой информационной размытости и неопределенности.

В работе рассматриваются задача анализа функционирования системы управления заданного множества объектов, каждый из которых характеризуется фиксированным (исходным) набором разнородных параметров. Основная идея предлагаемого метода решения подобных задач состоит в следующем. В работе предлагается исследовать не точные значения параметров, описывающих состояние каждого объекта системы, а лишь структуру взаиморасположения этих объектов в пространстве параметров. Такое интегральное описание управляемых объектов позволяет существенно повысить эффективность анализа поведения системы, а также устойчивость и робастность процедур принятия управленческих решений. Для формализации такой задачи используется методология классификационного анализа данных [1,4].

Пусть исследуемая система состоит из n объектов, каждый из которых характеризуется набором из k параметров. Вводится в рассмотрение k-мерное пространство параметров X, в котором каждый объект представляется точкой $x_j = (x_j^{(1)}, x_j^{(2)}, \ldots, x_j^{(k)})$, $j = 1, \ldots, n$. Предполагается, что вектор значений параметров x_j достаточно полно характеризует состояние j-го объекта, то есть взаиморасположение множества точек $x_1, \ldots, x_n$ в пространстве параметров X отражает реальную структуру исследуемого множества объектов. Для выявления такой структуры был разработан комплекс алгоритмов интеллектуального анализа данных и процедур экспертной коррекции, включаю-

* Работа выполнена при частичной поддержке РФФИ, гранты 14-07-00463-а, 15-07-06713-а, 16-07-00896-а, 16-07-00895-а, 16-29-12880-офи; РНФ: грант 14-19-01772.

щий алгоритмы: структурно-классификационного анализа данных, выбора информативных параметров, выбора начального разбиения, выбора числа классов, заполнения пропущенных наблюдений, а также процедуры экспертной коррекции результатов работы этих алгоритмов. Далее каждый из этих алгоритмов рассматривается отдельно.

2. Алгоритм структурно-классификационного анализа данных (СКАД)

Пусть задано R_0 – некоторое начальное разбиение (классификация) точек классифицируемой выборки $x_1, .., x_n$ на r классов A_i, i=1÷ r. Алгоритм циклический, многоэтапный, итерационный, – на j-ом шаге l-го этапа рассматривается некоторый набор из l точек X_j^l из исходной последовательности $x_1, ..., x_n$, принадлежащих одному и тому же классу, j – номер этого набора. Номер этапа l равен мощности множества точек, которые «перебрасываются» на каждом шаге этого этапа из класса в класс, т.е. числу точек в наборе. На j-ом шаге происходит пробная «переброска» из класса в класс множества точек X_j^l. Тогда X_j^l относится к тому классу A_s, значение критерия качества классификации J для которого будет наибольшим, т.е. $X_j^l \in A_s$, для которого $A_s = arg\,max\, J(X_j^l \in A_i)$, $i = 1 \div r$, $j = 1 \div N_l$, где N_l – число различных наборов из l точек в исходной выборке, принадлежащих одному и тому же классу. На следующем шаге l-го этапа процедура повторяется для множества X_{j+1}^l. Число шагов (итераций) на l-ом этапе равно N_l и определяется выражением $N_l = \sum_{i=1}^{r} C_{n_i}^l$ для $n_i \geq (l+2)$, где C_m^k – число сочетаний из m по k. Из этого выражения следует, что для всех итераций l-го этапа процедура не применяется для таких классов A_i, число точек n_i в которых меньше, чем (l+2). Число этапов (глубина перебора) $l_{\max}$ либо фиксируется заранее $l_{\max} = m$, либо выбирается из условия: в классификации, полученной после (l-1)-го этапа, должен быть хотя бы один класс, число точек в котором не меньше (l+2). Это правило обеспечивает автоматический выбор максимально возможной глубины перебора $l_{\max}$. Для повышения эффективности алгоритма СКАД используется следующая циклическая процедура. После завершения последнего этапа (либо m-ый, либо $l_{\max}$-ый) весь описанный выше цикл повторяется заново, только в качестве начальной классификации используется не R_0, а классификация, полученная на последнем этапе первого цикла. Алгоритм СКАД заканчивает работу, если на некотором цикле среди точек $x_1, ..., x_n$ не будет сделано ни одной «переброски» из класса в класс, т.е. для этого цикла начальная классификация совпадает с конечной. Доказана следующая теорема о сходимости этого алгоритма .

Теорема 1. *Алгоритм СКАД сходится за конечное число шагов к локальному максимуму критерия J.*

Доказательство (без ограничения общности даётся для случая двух классов). По процедуре работы алгоритма СКАД значения критерия J образуют монотонно не убывающую, ограниченную сверху последовательность. Величина ограничения $C_1 \geq J_i$, $i \in D$, где D – множество номеров всех возможных дихотомий исходной выборки $x_1, .., x_n$, зависит от вида выбранного критерия J. С другой стороны, в силу конечности исходной выборки существует такая константа C_2, что $C_2 \leq |J_i - J_j|$, $i, j \in D, i \neq j$. Таким образом, может быть только конечное число шагов N, на которых критерий J возрастает: $N \leq C_1 / C_2$. На некоторых шагах работы алгоритма СКАД возможны ситуации, когда при равенстве значений критерия J до и после «переброски» l точек происходит изменение принадлежности этих точек к классу. Для того чтобы не допустить возможности циклической последовательности таких «перебросок» (что приводит к бесконечному числу таких шагов), в алгоритме введено специальное правило для таких случаев: при равенстве значений критерия J до и после «переброски» соответствующие l точек относятся к классу с меньшим номером. Это означает, что таких перебросок с нулевым приращением значения критерия J также будет конечное число. Достижимость локального максимума непосредственно следует из самой процедуры «переброски» точек. Действительно, предположим противное – после останова значение критерия $J_{êîí}$ не является l–локальным максимумом. А это по определению локального экстремума означает, что существует, по крайней мере, один набор из l точек исходной последовательности, для которого изменение принадлежности к классу приведет к увеличению значения критерия J. Однако правило останова гарантирует, что такого набора в исходной последовательности не существует, поскольку на последнем перед остановом цикле не было ни одной «переброски» l точек. Теорема доказана.

Алгоритм сокращенного перебора – эвристический вариант выбора множеств X_j^l. На каждом шаге алгоритма для пробной «переброски» использует точки в определенном смысле ближайшие к границе между классами. Иллюстрация идеи

работы алгоритма представлена на рис. 1. Четыре точки, обведенные кружочками – это как раз и есть те l точек (рассматривается случай $l = 4$), которые на j-ом шаге ближе всего расположены к границе (квадратиками обозначены центры классов на j-ом шаге). Если уравнение границы в явном виде неизвестно, то выбираются l точек, ближайших к эталону другого класса. Обычно в качестве эталона выбирается точка «центра тяжести» всех точек исходной выборки, принадлежащих на j-ом шаге соответствующему классу (то есть среднему этого класса):

$$a_s(j) = \frac{1}{n_s(j)} \sum_{x_i \in As(j)} x_i, \qquad (1)$$

где s – номер класса.

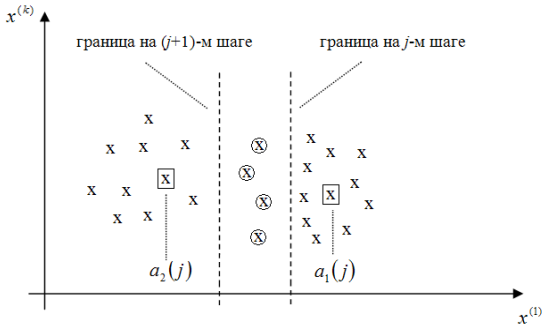

Рис. 1. Иллюстрация идеи сокращенного перебора

Алгоритм СКАД для одномерного случая. Необходимо специально отметить этот частный, но весьма распространенный в прикладных задачах случай, связанный со структурным анализом временных рядов [2]. Дело в том, что одномерный случай имеет уникальное свойство, существенно упрощающее процедуру целенаправленного перебора, используемую при структурном анализе. А именно: ввиду одномерной упорядоченности классов границей между двумя классами (в детерминированном случае) служит только одна точка, и таких границ может быть не более двух (для крайних правого и левого классов – только одна).

Теорема 2. *Одномерный вариант алгоритма СКАД сходится за конечное число шагов к глобальному максимуму критерия J.*

Доказательство. Одномерный случай существенно отличается от многомерного тем, что классы упорядочены на оси X. Это, в свою очередь, позволяет декомпозировать процесс минимизации функционала J для всей выборки на независимые процедуры его минимизации для подвыборок, каждая из которых составляет одну из всех соседних пар классов. Таким образом, этот алгоритм фактически является реализацией схемы динамического программирования, обеспечивающей нахождение глобального экстремума функционала J [9].

Действительно, предположим противное, – после завершения работы одномерного алгоритма СКАД получена классификация на r классов (обозначим ее через $H_{лок}$), доставляющая не глобальный, а лишь локальный экстремум $J_{лок}$ функционала J. Это означает, что существует такая классификация $H_{глоб}$, для которой значение функционала $J_{глоб}$ будет больше, чем $J_{лок}$.

Введем в рассмотрение пересечение двух классификаций $H_1=\{A_{11}, ..., A_{1r}\}$ и $H_2=\{A_{21}, ..., A_{2r}\}$: это множество $H_1 \cap H_2=\{A_{1i} \cap A_{2i},\ i=1\div r\}$; а также их разность $H_1 \setminus H_2=\{A_{1i} \setminus A_{2i},\ i=1\div r\}$.

Рассмотрим пересечение классификаций $H_{глоб} \cap H_{лок}$, а затем вычтем его из классификации $H_{лок}$. Обозначим получившийся в результате набор множеств точек через $B_1, \ldots, B_r$. Далее рассматриваются только непустые множества такого вида. Рассмотрим для примера множество B_1 (пусть оно содержит m_1 точек) из первого класса классификации $H_{лок}$. Рассмотрим этап алгоритма СКАД для одномерного случая, на котором анализируются точки только первого и второго классов, остальные границы считаются фиксированными. В качестве начальных условий выберем границу между первым и вторым классами из классификации $H_{глоб}$. В соответствии с работой алгоритма, точки множества B_1 должны быть «переброшены» в первый класс, так как по построению такой переброске будет соответствовать большее значение функционала J. Аналогичные рассуждения проводятся для всех множеств B_i, $i=1\div r$. Следует подчеркнуть, что на каждом цикле рассмотрения пары соседних классов используется правило выбора максимально возможной глубины перебора l_{max}, обеспечивающее глобальный экстремум критерия J для рассматриваемой пары классов (при фиксированных остальных классах).

Из вышеизложенного можно сделать вывод, что предположение о существовании классификации $H_{глоб}$, доставляющее большее значение функционалу J, чем классификация $H_{лок}$, неверно. Таким образом, полученная в результате работы алгоритма классификация доставляет глобальный экстремум функционалу J. Теорема доказана.

При моделировании и в приложениях в качестве критерия качества классификации J использовался функционал J_1 средней близости точек в классах, определяемый через потенциальную функцию $K(x,y)$ близости точек x и y [3]:

$$K(x,y) = \frac{1}{1+\alpha R^p(x,y)}, \qquad (2)$$

где α и p – настраиваемые параметры алгоритма. Средняя близость точек в классе определяется как:

$$K(A_i, A_i) = \frac{2}{n_i(n_i-1)} \sum_{i=1}^{n_i} \sum_{j>i} K(x_i, x_j), \quad (3)$$

где $K(x_i, x_j)$ определяется формулой (2), n_i – число точек в классе A_i. Тогда критерий J_1 определяется как:

$$J_1 = \sum_{i=1}^{r} \frac{n_i}{n} K(A_i, A_i). \quad (4)$$

Во многих задачах структурно-классификационного анализа объекты по самой постановке задачи могут относиться к разным классам с различной степенью «достоверности». Для таких случаев была разработана постановка задачи размытого классификационного анализа [1,4].

Вариант алгоритма СКАД в размытой постановке. Размытой классификацией множества X на r классов называется r-мерная вектор-функция $H(x) = (h_1(x),...,h_r(x))$, где $h_i(x)$ – функция принадлежности объекта x к i-му классу, удовлетворяющая условию нормировки: $\sum_{i=1}^{r} h_i(x) = 1$, $0 \le h_i(x) \le 1$ [1]. Критерий оценки качества классификации содержательно остается прежним, только видоизменяется процедура подсчета его значений. А именно, функционал принимает следующий вид: $J = \sum_{i=1}^{r} B_i K(A_i, A_i)$, где $B_i = \frac{1}{n} \sum_{j=1}^{n} h_i(x_j)$ – нормирующий множитель, аналогичный n_i/n для детерминированного случая. Величина $K(A_i,A_i)$ средней близости точек в классе A_i для размытого случая определяется по формуле:

$$K(A_i, A_i) = c_i \sum_{j=1}^{n} \sum_{l>j} K(x_j, x_l) h_i(x_j) h_i(x_l), \quad (5)$$

где $c_i = 2 / \sum_{j=1}^{n} (h_i(x_j))^2$ – нормирующий множитель.

Рассмотрим вкратце работу ***размытого алгоритма СКАД***. Для простоты изложения и без ограничения общности рассмотрим случай двух классов (r = 2). Пусть задано начальное размытое разбиение $H_0=\{h_i(x_j)$, i=1,2, j=1,…,$n\}$ точек классифицируемой выборки $x_1,\dots, x_n$, которое может задаваться либо изначально, либо при помощи специального алгоритма выбора начального разбиения. Для начального размытого разбиения H_0 подсчитывается значение критерия $J(H_0)$. Как и в детерминированном случае размытый алгоритм является циклическим, многоэтапным, итерационным, – на j-ом шаге l-го этапа рассматривается некоторый набор из l точек X_j^l из исходной последовательности $x_1,..., x_n$. При этом для всех точек множества X_j^l справедливо неравенство $h_l(x_s) > h_k(x_s)$, $l \ne k$, что соответствует требованию для детерминированного алгоритма: все точки множества X_j^l принадлежат одному и тому же классу. Для множества точек X_j^l выполняется следующая операция, аналогичная «переброске» точек из класса в класс. Вводится понятие «старой» и «новой» функций принадлежности $h_i(x_s)_{\text{стар}}$ и $h_i(x_s)_{\text{нов}}$ соответственно, $x_s \in X_j^l$. Тогда, если для всех $x_s \in X_j^l$ выполняется $h_1(x_s)_{\text{стар}} > h_2(x_s)_{\text{стар}}$ (аналог того, что набор точек X_j^l принадлежит первому классу), то $h_1(x_s)_{\text{нов}} = h_2(x_s)_{\text{стар}}$ и $h_2(x_s)_{\text{нов}} = h_1(x_s)_{\text{стар}}$ (аналог того, что набор точек X_j^l «переброшен» во второй класс). Далее подсчитывается значение критерия $J(H_1)$ с «переброшенным» набором точек X_j^l. Если значение $J(H_1) > J(H_0)$, то изменения функций принадлежности для набора точек X_j^l остается в силе, в противном случае произведенные изменения значений функций принадлежности отменяются. Аналогичная операция выполняется для случая, когда $h_1(X_j^l)_{\text{стар}} < h_2(X_j^l)_{\text{стар}}$ (аналог того, что набор точек X_j^l принадлежит второму классу). Алгоритм прекращает работу, если на каком-то цикле не было произведено изменений функций принадлежности ни для одной из точек исходной выборки.

3. Алгоритм построения начального разбиения

В составе комплекса алгоритмов интеллектуального анализа данных был разработан алгоритм построения начального разбиения – как для детерминированного, так и для размытого случая. Рассмотрим эти алгоритмы более подробно.

Детерминированный случай. Для простоты изложения без ограничения общности, алгоритм описан для случая двух классов – A_1 и A_2. На первом шаге из всех точек выборки $x_1, \dots, x_n$ находится пара наиболее удаленных друг от друга точек, x_l и x_p, одна из которых – x_l, относится к классу A_1, а другая – x_p, относится к классу A_2. Если n достаточно велико, то используется усеченный вариант первого шага, а именно: x_l выбирается случайно, а x_p ищется как точка, наиболее от нее удаленная.

Затем последовательно рассматриваются все точки выборки, за исключением точек x_l и x_p. А именно, на втором шаге рассматривается точка x_1 (при условии, что $x_1 \ne x_l$, $x_1 \ne x_p$), которая относится к первому классу, если она ближе к x_l, чем к x_p, и ко второму классу в противном случае. Если $x_1 = x_l$ или $x_1 = x_p$, тогда переходим к рассмотрению следующей точки. На j-ом шаге рассматривается точка x_j (при

условии, что $x_j \neq x_l$, $x_j \neq x_p$), которая относится к одному из двух классов в соответствии с правилом:

$$x_j \in \begin{cases} A_1, \text{ если } K(x_j, A_1) \geq K(x_j, A_2) \\ A_2, \text{ если } K(x_j, A_1) < K(x_j, A_2), \end{cases}$$

$$j = 1 \div n\,,\ x_j \neq x_l\,,\ x_j \neq x_l\,.$$

Такая процедура повторяется до тех пор, пока не будут исчерпаны все точки выборки. Полученное разбиение принимается в качестве начального разбиения R_0.

Размытый случай. На первом шаге алгоритма находится начальное разбиение R_0 выборки x_1, …, x_n на r классов в детерминированном случае. На втором шаге определяются a_1, …, a_r – центры всех классов в полученном разбиении R_0. Далее для каждой точки x_j рассчитываются функции принадлежности $h_i(x_j)$, i=1, …, r, j=1, …, n, значения которых обратно пропорциональны расстояниям до центров соответствующих классов: $h_i(x_j) = K(a_i, x_j) / \sum_{i=1}^{r} K(a_i, x_j)$, где $K(a_i, x_j)$ – потенциальная функция вида (2), а $\sum_{i=1}^{r} K(a_i, x_j)$ – нормирующий множитель, обеспечивающий выполнения условия нормировки функций принадлежности: $\sum_{i=1}^{r} h_i(x) = 1$. Полученный набор функций принадлежности и определяет размытое начальное разбиение $H_0 = \{h_i(x_j), i=1,\ldots,r; j=1,\ldots,n\}$.

4. Алгоритм выбора числа классов

Одна из основных проблем использования структурно-классификационных методов при решении задач исследования сложных систем управления – это выбор числа классов. Дело в том, что чрезвычайно важным фактором в прикладных исследованиях является содержательная интерпретация элементов, получаемых в результате анализа структуры объектов (например, классов объектов). В работе описан специально разработанный алгоритм оптимального выбора числа классов. При этом оптимальность понимается в смысле максимизации содержательно обоснованного критерия качества классификации. Алгоритм, по сути, представляет собой экспертно-компьютерную процедуру, которая работает следующим образом. Сначала эксперт оценивает диапазон (r_{min}, r_{max}), в пределах которого заведомо находится искомое число классов. Далее, используя алгоритм СКАД, проводится разбиение анализируемого множества объектов на r_{min}, r_{min}+1, ..., r_{max} классов. Качество каждой из полученных классификаций оценивалось с помощью критерия

$$J_3 = J_1 - q\,J_2\,. \qquad (6)$$

В формуле (6) величина J_1 – это средняя (по классам) мера близости точек, принадлежащих одному и тому же классу, вычисляется по формуле (4); q – свободный (настраиваемый) параметр алгоритма; J_2 – средняя мера близости классов, определяемая соотношением:

$$J_2 = \frac{1}{r-1} \sum_{i=1}^{r} \sum_{j>i} \frac{n_i + n_j}{n} K(A_i, A_j)\,. \qquad (7)$$

В формуле (7) величина n_j – это число точек в классе A_i, а величина $K(A_i, A_j)$ – мера близости классов A_i, A_j – вычисляется по формуле:

$$K(A_i, A_j) = \frac{1}{n_i\, n_j} \sum_{x_l \in A_i} \sum_{x_p \in A_j} K(x_l, x_p)\,. \qquad (8)$$

В формуле (8) потенциальная функция $K(x_l, x_p)$ определяется формулой (2). Параметр q является масштабирующим параметром, приводящим значения функционалов J_1 и J_2 к соизмеримым величинам; на практике q имеет значения порядка 2–7 (во столько раз обычно отличается средняя близость внутри классов от средней близости между самими классами). Более подробно вопрос выбора значений настраиваемых параметров рассмотрен далее в специальном разделе.

В итоге получается последовательность $J_3(r_{min})$, ..., $J_3(r_{max})$. Формально в качестве наилучшего (оптимального) можно выбрать такое число классов r_{opt}, которое соответствует экстремальному значению критерия (6):

$$r_{opt} = r_j / max\, J_3(r_j) \quad r_j = r_{min}, \ldots, r_{max}\,.$$

Однако наличие существенной, но неиспользованной при классификации информации (например, ввиду отсутствия данных) может привести к тому, что полученное таким способом r_{opt} не будет наилучшим с точки зрения эксперта. Для компенсации этого недостатка предлагается использовать следующую экспертную процедуру. Экспертам выделяются значения $J_3(r_j)$, $r_j = r_{min}$, …,r_{max}, представленные для удобства в виде графика, на котором отмечается значение r_{opt} (оно соответствует максимальной точке на графике). Используя эту информацию, эксперты могут корректировать выбираемое число классов. В подавляющем числе случаев экспертное число классов либо совпадает с $r_{opt,}$, либо незначительно (±1) отличается от него.

При классификации многомерных объектов во время такой экспертизы анализируется также классификация каждого объекта. Для этой цели экспертам сообщается информация о мере близости $K(x_i, c_j)$ каждой точки x_i до центров классов c_j, j=1, …, r в оптимальной классификации, то есть матрица близости $\|K(x_i, c_j)\|$, i=1, …, n, j=1,…, r_{opt}. Перенесение точки (объекта) x_i из j-го класса в l-ый считается допустимым, если величины $K(x_i, c_j)$ и $K(x_i, c_l)$ отличаются незначительно. Другими словами, содержательно обоснованное перенесение допустимо для точек, расположенных вблизи границы между соответствующими классами.

5. Алгоритм выбора информативных параметров

Опыт использования алгоритмов структурно-классификационного анализа показывает, что классификация по всем исходным параметрам далеко не всегда приводит к желаемым результатам [4]. 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 объектов целесообразно проводить не в исходном пространстве, а в пространстве наиболее существенных (информативных) параметров, имеющем значительно меньшую размерность.

Для выбора информативных параметров в работе предлагается использовать результаты структуризации параметров. Далее, для того, чтобы отличать структуризацию объектов и параметров, будем говорить не о классификации объектов, а о группировке параметров.

Алгоритм СКАД в задаче группировки параметров. Для группировки параметров, как и в случае классификации объектов, предлагается использовать алгоритм СКАД. Формальная постановка задачи группировки параметров подразумевает определение: множества параметров, подлежащих группировке, множества решающих правил и критерия качества группировки [1].

Группируемое множество параметров – это конечный набор параметров $\{x^{(1)},...,x^{(k)}\}$, полученный из исходного набора после нормировки дисперсии каждого параметра на 1. Здесь $x_j^{(i)}, i = 1 \div k, j = 1 \div n$ определены как реализации случайной величины $x^{(i)}$ на множестве исследуемых объектов. ***Множество решающих правил***, как и в случае классификации объектов – единичный симплекс [1].

Для формулировки критерия качества группировки необходимо ввести меру близости между параметрами (случайными величинами) x и y. В качестве такой меры используется коэффициент ковариации (совпадающий с коэффициентом корреляции для нормированных параметров x и y), который будем обозначать через $cov_{x,y} = (x, y)$, понимая его как скалярное произведение случайных величин x и y. Для дисперсии $cov_{x,x}$ случайной величины x используется обозначение $cov_{x,x} = (x,x) = x^2$. ***Критерий качества группировки*** используется в виде следующего функционала:

$$J^* = \sum_{j=1}^{s} \sum_{\substack{x^{(i)}, x^{(l)} \in A_j \\ x^{(i)} \neq x^{(l)}}} cov^2_{x^{(i)}, x^{(l)}} = \sum_{j=1}^{s} \sum_{\substack{x^{(i)}, x^{(l)} \in A_j \\ x^{(i)} \neq x^{(l)}}} \left(x^{(i)}, x^{(l)}\right)^2, \quad (9)$$

где s – число групп. Максимизация функционала (9) соответствует интуитивному представлению о «хорошем» разбиении параметров, – когда в одну и ту же группу попадают наиболее близкие (в определенном выше смысле) параметры. В этом смысле функционал (9) полностью аналогичен функционалу (4), который используется как критерий качества классификации объектов.

Для выбора информативных параметров чрезвычайно важно знать интегральные характеристики (эталоны) полученных групп. Для классификации объектов такими эталонами обычно являются «центры тяжести» точек, попавших в один и тот же класс, которые вычисляются по формуле (1). Для группировки параметров такого типа эталонами являются «средние» (в определенном выше смысле) виртуальные нормированные параметры (случайные величины) $f_1, f_2, ..., f_s$ такие, что $f_j^2 = 1,\ j = 1 \div s$, которые будем называть факторами. Факторы (эталоны) некоторой группировки на s групп $A_1, A_2, \ldots, A_s$ определяются соотношением (10), являющемся, в определенном смысле, аналогом (1):

$$f_j = \arg\max_{f} \sum_{x^{(i)} \in A_j} \left(x^{(i)}, f\right)^2, \quad f^2 = 1. \quad (10)$$

При решении прикладных задач критерий качества группировки (9) иногда удобнее представить в эквивалентном виде:

$$J^* = \sum_{j=1}^{s} \sum_{x^{(i)} \in A_j} \left(x^{(i)}, f_j\right)^2. \quad (11)$$

Таким образом, задача группировки набора *k* параметров на заданное число групп *s* состоит в максимизации функционала (11) как по разбиению параметров на группы A_j, так и по выбору факторов f_j, $j=1\div s$, $f_j^2=1$, определяемых из соотношения (10) при фиксированной группировке.

Легко показать [3], что для фиксированной группировки на непересекающиеся группы $A_1, A_2, \ldots, A_s$ (детерминированная постановка задачи) факторы (эталоны групп) определяются по формуле:

$$f_j = \frac{\sum\limits_{x^{(i)}\in A_j} \alpha_i x^{(i)}}{\sqrt{\left(\sum\limits_{x^{(i)}\in A_j} \alpha_i x^{(i)}\right)^2}} = \frac{\sum\limits_{x^{(i)}\in A_j} \alpha_i x^{(i)}}{\sqrt{\sum\limits_{\substack{x^{(i)}\in A_j,\\ x^{(l)}\in A_j}} \alpha_i \alpha_j \left(x^{(i)}, x^{(l)}\right)}}, \; j=1\div s \quad (12)$$

где α_i – компоненты собственного вектора матрицы $R_j = \left\|\left(x^{(i)}, x^{(l)}\right)\right\|$, $x^{(i)}, x^{(l)} \in A_j$ соответствующего её наибольшему собственному значению. Из (12) непосредственно следует, что фактор группы – это линейная комбинация параметров, отнесенных к этой группе (знаменатель в (12) необходим для нормировки $f^2=1$), причем коэффициентами в этой комбинации являются компоненты «максимального» собственного вектора ковариационной матрицы параметров из этой группы.

Для одновременного определения групп $A_1, A_2, \ldots, A_s$ и факторов $f_1, f_2, \ldots, f_s$, удовлетворяющих этим условиям, используется описанный выше алгоритм СКАД, который в данном случае работает следующим образом (для простоты описан алгоритм СКАД в детерминированном случае для l=1, s=2).

Пусть задано некоторое начальное разбиение R_0^* группируемого множества параметров $\left\{x^{(1)}, \ldots, x^{(k)}\right\}$. Обозначим через $x^{(j)} \in A_1^*$ параметры, относящиеся к первой группе, а через $x^{(j)} \in A_2^*$ – ко второй. Алгоритм итерационный, на каждом шаге рассматривается один параметр из последовательности $x^{(1)}, \ldots, x^{(k)}, x^{(1)}, \ldots, x^{(k)}, x^{(1)}, \ldots$ («зацикленная» исходная последовательность параметров). Отнесение параметра $x^{(j)}$ к одной из двух групп обозначается с помощью индекса $\rho(x^{(j)})$, который равен 1, если $x^{(j)} \in A_1^*$, и равен -1 в противном случае. Тогда этот вариант алгоритма группировки записывается в виде:

$$\rho\left(x^{(j)}\right) = sign\left[J^*\left(x^{(j)} \in A_1^*\right) - J^*\left(x^{(j)} \in A_2^*\right)\right],$$
$$j = 1, \ldots, k, 1, \ldots, k, 1, \ldots . \quad (13)$$

Таким образом, на каждом шаге текущий параметр $x^{(j)}$ относится к той группе, при отнесении к которой, значение критерия J^* будет больше (если эти значения равны, то он относится к группе с наименьшим номером). Алгоритм (13) заканчивает работу, если на некотором цикле среди параметров $x^{(1)}, \ldots, x^{(k)}$ не будет сделано ни одной «переброски» параметра из группы в группу. При этом, если критерий качества имеет вид (9), то факторы групп f_1 и f_2 определяются с помощью (12) по завершении процедуры группировки. Если же используется критерий в виде (11), то для подсчета значений $J^*\left(x^{(j)} \in A_i^*\right)$ в (13) факторы групп необходимо определять с помощью (12) на каждом шаге.

Теорема 3. *Алгоритм СКАД в задаче группировки параметров сходится к локальному максимуму функционала J* за конечное число шагов (итераций).*

Доказательство этого утверждения аналогично доказательству Теоремы 1.

Выбор информативных параметров. В результате применения алгоритма СКАД к исходным *k* параметрам будет получено их разбиение на заданное число групп *s* (в прикладных задачах значение *s* колеблется в диапазоне 3–10), а также значения факторов для полученных групп. При решении прикладных задач в дальнейшем используются либо новые интегральные параметры – факторы групп (если удается получить их удовлетворительное содержательное описание), либо такой набор параметров из исходного множества параметров (число которых равно числу групп), каждый из которых является ближайшим (в определенном выше смысле) к фактору в соответствующей группе. В некоторых случаях (например, когда нет такого параметра в группе, значения коэффициента корреляции которого с фактором значимо больше, чем для других параметров этой группы) в отдельных группах может быть отобрано по 2, а для особо многочисленных групп – по 3 и более параметров, ближайших к соответствующему фактору и максимально удаленных друг от друга. Иногда при формировании набора информативных параметров используются процедуры экспертной коррекции [4,5].

В большинстве приложений исходные или выделенные информативные параметры имеют неравнозначную важность при анализе структуры объектов. Для формирования коэффициентов важности (весов) в работе предлагается использовать процедуры экспертного оценивания. Хорошие результаты дает процедура многовариантной экспертизы [6], когда для оценки таких весов используется несколько групп экспертов – специалистов в различных аспектах исследуемой проблемы. В результате экспертизы каждому параметру присваивается определенный вес (важности) при исследовании структуры объектов.

Выбор «оптимального» числа классов в задаче группировки параметров. Для выбора числа классов в задаче группировки параметров используется специальная экспертно-компьютерная процедура оптимизации критерия, аналогичного критерию выбора «оптимального» числа классов в задаче классификации объектов. Опишем вкратце эту процедуру.

Сначала эксперт-пользователь оценивает диапазон (s_{min}, s_{max}), в пределах которого заведомо находится искомое число групп. Далее, используя алгоритм СКАД, проводится разбиение группируемого множества параметров на s_{min}, $s_{min}+1$, …, s_{max} групп. Качество каждой из полученных группировок оценивается с помощью критерия:

$$J_3^*(s) = J_1^*(s) - qJ_2^*(s), \qquad (14)$$

где J_1^* – величина средней по группам меры близости параметров в группе, а J_2^* – величина средней меры близости между группами. Величина q в (14) является масштабирующим параметром, приводящим к одному масштабу средние значения функционалов J_1^* и J_2^*. На практике величина q выбирается в диапазоне значений 2-5 (обычно во столько раз отличается средняя близость параметров внутри групп от средней близости самих групп).

В качестве «оптимального» можно выбрать такое число групп $s_{opt}= s_j$, которое соответствует максимальному значению критерия (14) для $s_j = s_{min}$, …, s_{max}. Однако, наличие существенной, но неиспользованной при классификации информации может привести к тому, что так полученное s_{opt} не будет «истинно оптимальным». Для компенсации этого недостатка используется процедура экспертной коррекции [4,5].

6. Особенности реализации разработанного комплекса алгоритмов

В процессе реализации разработанного комплекса алгоритмов интеллектуального анализа данных возникает целый ряд проблем, для разрешения которых приходится разрабатывать специальные процедуры или использовать уже известные алгоритмы. В большинстве приложений, особенно связанных с социально-экономическими системами, пользователь сталкивается с проблемой качества исходных данных. Здесь, прежде всего, необходимо выявлять ошибки в исходных данных, в том числе имеющие случайный характер. Для этой цели используются разнообразные алгоритмы фильтрации. Например, для выявления существенных «выбросов» в значениях параметров строится гистограмма распределения значений каждого из параметров, и в зависимости от содержательной модели исследуемого объекта выбирается тот или иной тип функции распределения. Для структурно-классификационных алгоритмов наиболее адекватной моделью является смесь нормальных распределений. Существуют стандартные статистические методы для определения того, является ли анализируемое значение выбросом или согласуется с выбранной моделью порождения данных [7]. В любом случае, по виду гистограммы экспертным путем всегда можно определить какое из значений параметра заведомо является «выбросом». Так, например, во многих приложениях широко используется так называемое «правило 3σ». Правило действует следующим образом: для каждого числового параметра по имеющейся выборке определяется среднее значение $\hat{x}^{(i)} = \frac{1}{n}\sum_{j=1}^{n} x_j^{(i)}$ и стандартное отклонение $\sigma^{(i)} = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}\left(x_j^{(i)} - \hat{x}^{(i)}\right)^2}$. Все значения $x_j^{(i)}$, превосходящие $\hat{x}^{(i)} \pm 3\sigma^{(i)}$ считаются «выбросами». Обычно «выбросы» заменяются либо на среднее значение этого параметра, либо на соответствующую границу диапазона $\hat{x}^{(i)} \pm 3\sigma^{(i)}$. В работе предлагается выбросы считать пропущенными наблюдениями и использовать для их заполнения специально разработанную процедуру.

Процедура заполнения пропущенных наблюдений. Как уже говорилось выше, во многих приложениях имеются пропуски в данных, кроме того, в процессе фильтрации «выбросы» часто рассматриваются как пропущенные наблюдения. В этой ситуации нужно либо использовать специальные процедуры подсчета расстояний между объектами, в параметрах которых имеются пропуски, либо разрабатывать специальные процедуры заполнения таких пропусков. В подавляющем большинстве работ, пропуски по каждому параметру предлагается заполнять средним известных значений соответствующего параметра (для исходной выборки). В настоящей работе была разработана специальная процедура заполнения пропусков в исходных данных с использованием алгоритмов автоматической классификации. Основная идея процедуры состоит в следующем. Если множество изучаемых объектов структурировано (то есть их можно разделить на классы, достаточно компактно расположенные в пространстве параметров X), то дисперсия (диапазон) изменения каждого параметра в пределах каждой группы, как правило, будет

существенно меньше, чем показатель для значения этого параметра на всей выборке. Таким образом, если по данным с пропусками удастся определить реальную структуру взаиморасположения точек (т.е. провести классификацию, адекватную этой структуре), то заполнять пропущенное значение *l*-го параметра для объекта из *i*-го класса можно средним этого параметра по его известным значениям для всех объектов, попавших в *i*-ый класс. Исходя из сделанного предположения, отклонение полученного значения от «истинного» должно быть существенно меньше (в среднем), чем обычная схема заполнения по общему среднему.

Опишем процедуру более подробно. На первом шаге все пропуски заполняются средними значениями каждого параметра по всей выборке. Далее проводится классификация выборки с заполненными пропусками на r_0 классов, где r_0 выбирается из следующих соображений. В каждом классе число объектов должно быть достаточным для статистически значимой оценки среднего значения параметра, т.е. не меньше, чем 8-10 точек. Поэтому $r_0^{нач} = n/15$ (с учетом неоднородности распределения числа точек по классам). Если в полученной классификации для некоторого класса число входящих точек будет меньше 8, то такой класс присоединяется к ближайшему классу. Некоторые из таких классов могут объединиться между собой, тогда дальнейшее их объединение не производится, если число точек в образованном классе больше или равно 8. В качестве меры близости двух классов A_i, A_j используется величина $K(A_i, A_j)$, определяемая формулой (8). В итоге, получается разбиение на r_1 классов. Затем, в каждом из полученных классов ранее заполненные пропущенные наблюдения заполняются новыми значениями. А именно, пропущенное значение *i*-го параметра для *j*-го объекта заменяется средним известных значений *i*-го параметра для всех объектов из *l*-го класса (к которому принадлежит *j*-ый объект). Такое заполнение производится для всех значений параметров, пропущенных в исходной выборке. На втором шаге происходит точно такая же процедура для матрицы данных, полученной после первого шага. Процедура заканчивается на шаге, на котором классификация точек осталась неизменной относительно предыдущего шага.

Выбор свободных параметров алгоритма. Комплекс алгоритмов имеет несколько настраиваемых параметров, которые должны быть выбраны либо до его использования на конкретном материале с помощью экспертов, либо в процессе такого использования с привлечением экспертных процедур. Таковыми параметрами являются: α и *p* в формуле (2), определяющей значение потенциальной функции $K(x,y)$, а также параметр q в формуле (6), определяющей критерий J_3 выбора оптимального числа классов. При выборе α и *p* в (2) воспользуемся следующими соображениями. Введем в рассмотрение величину R_{cp} (расстояние «среза»), определяемую равенством:

$$\left.\frac{d^2K\left[R(x,y)\right]}{dR^2(x,y)}\right|_{R_{cp}} = 0\,. \qquad (15)$$

Значение величины R_{cp} в (15) определяет точку перегиба функции $K[R(x,y)]$, т.е. точку максимальной крутизны этой функции. На рис. 2 изображен график функции для различных значений *p* при одном и том же значении величины R_{cp}. Параметр *p* при фиксированном R_{cp} характеризует крутизну функции $K[R(x,y)]$ в районе точки перегиба. Для удобства счета в качестве *p* выбирают числа кратные 2 (2,4,6,…).

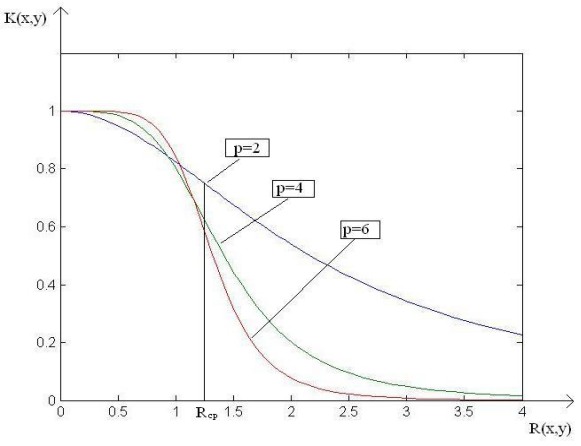

Рис 2. График функции K[R(x,y)] для различных значений ***p*** при одном и том же $\boldsymbol{R_{cp}}$ =1,25.

Параметр α в выражении (2) при известном R_{cp} и фиксированном *p* определяется из выражения: $\alpha = \dfrac{p-1}{(p+1)\,R_{cp}}$. Обычно R_{cp} выбирается из следующих соображений. По определению, точки, входящие в одну и ту же группу (класс), имеют высокие значения функции близости (значения потенциальной функции). А это означает, что расстояние между точками одной и той же группы в большинстве случаев меньше R_{cp}. И, наоборот, – значения потенциальной функции (меры близости) между точками из разных групп существенно меньше, чем аналогичные значения для точек из одной и той же группы, т.е. соответствующее значение расстояния будет больше, чем R_{cp}. Это означает, что R_{cp} должно

равняться расстоянию от границы группы до центра группы (в среднем по всем группам). Поскольку до самой группировки определить это значение невозможно, то обычно делается 2-4 пробных расчета для различных значений этого параметра. Начальное $R_{ср}$ обычно выбирается как функция от размерности k пространства X, числа классов r и «характерного» размера множества точек выборки, например, диаметр сферы, описывающей все точки исходной выборки. В работе для этой цели используется выражение $R_{ср}^{нач} = \frac{\sqrt{k}\ R_{max}(x_i, x_j)}{r}$, где $R_{max}(x_i, x_j)$ – расстояние между максимально удаленной пары точек исходной выборки (после фильтрации и замены «выбросов», о которых говорилось выше).

При выборе p необходимо иметь ввиду следующее обстоятельство. Если исследуемый материал достаточно хорошо структурирован, т.е. в пространстве X имеются хорошо обособленные друг от друга группы точек, то крутизна функции $K(x,y)$ в районе $R_{ср}$ может быть не очень большой, т.к. влияние далеких точек, находящихся на расстоянии существенно большем, чем $R_{ср}$, будет не существенно. С другой стороны, если такой явной структурированности нет (например, в случае сильной зашумленности данных), то в «промежутках» между группами будет достаточное количество точек (так называемые «мосты»). В этом случае, крутизна потенциальной функции в районе границы, т.е. в районе $R=R_{ср}$, должна быть достаточно высокой, чтобы минимизировать влияние точек в районе границы на процесс кластеризации. Для сильно зашумленных данных разработаны алгоритмы, в которых вводится специальный, так называемый, «фоновый» класс [1]. К фоновому классу относятся точки, которые расположены достаточно далеко от центров всех классов. В прикладных исследованиях величина p подбирается экспериментально: начальное значение $p = 2$ выбирается для случая хорошей структурированности, а $p = 4, \ldots, 8$ – для случаев слабой структуризации.

Следует отметить, что в прикладных задачах могут использоваться как числовые, так и качественные переменные. В первом случае, в качестве $R(x,y)$ в формуле (2) используется евклидово расстояние $R(x,y) = R_e(x,y) = \sqrt{\sum_{i=1}^{k}\left(x^{(i)} - y^{(i)}\right)^2}$.

В приложениях различные типы качественных параметров в подавляющем числе случаев приводятся к набору логических переменных, для них используется расстояние по Хеммингу: $R(x,y) = R_h(x,y)$, т.е. число несовпадающих разрядов в двоичных кодах векторов. Заметим, что для логических переменных x и y: $R_h(x,y) = R_e^2(x,y)$. Если среди входных параметров есть как числовые, так и логические переменные, то в качестве квадрата расстояния можно использовать величину $R^2(x,y) = R_h(\tilde{x},\tilde{y}) + R^2{}_e(\hat{x},\hat{y})$, где $\tilde{x},\tilde{y}$ – логические переменные, $\hat{x},\hat{y}$ – числовые.

При выборе масштабирующего параметра q в формуле (6) обычно руководствуются следующими соображениями. Из формул (3) и (4), определяющих J_1, и формул (7) и (8), определяющих J_2, непосредственно следует, что величина J_1 существенно больше, чем J_2. Значения J_1 и J_2 определяются конкретной структурой расположения точек в пространстве X и выбранными значениями параметров $R_{ср}$ и p. Моделирование разработанных алгоритмов, а также решение некоторых прикладных задач показало, что характер поведения функции $J_3 = J_1 - qJ_2$ мало меняется в широком диапазоне значений q. В зависимости от структурированности пространства X хорошие результаты получаются для значений q в диапазоне 2–7.

7. Заключение

Разработанный комплекс алгоритмов интеллектуального анализа данных использовался для анализа сложноорганизованных данных в рамках исследования сложных систем управления, а также при совершенствовании процедур принятия решений для нескольких крупных систем управления, в основном регионального характера. Во всех приложениях, а также при машинном моделировании [8], была подтверждена высокая эффективность разработанного комплекса.

Литература

1 *Дорофеюк А.А., Бауман Е.В., Дорофеюк Ю.А.* Методы интеллектуальной обработки информации на базе алгоритмов стохастической аппроксимации. // Математические методы распознавания образов. 15-ая международная конференция: Сб. докладов. М.: МАКС ПРЕСС, 2011. С. 108-112.

2 *Гольдовская М.Д., Дорофеюк Ю.А., Киселева Н.Е.* Методы структурного анализа в прикладных задачах исследования временных рядов. // Проблемы управления. 2013, № 3. С. 33-41.

3 *Браверман Э.М., Мучник И.Б.* Структурные методы обработки эмпирических данных. – М.: Наука, 1983.

4 *Дорофеюк А.А.* Методология экспертно-классификационного анализа в задачах управления и обработки сложноорганизованных данных (история и перспективы развития). // Проблемы управления. 2009. № 3.1. С. 19-28.
5 *Дорофеюк А.А., Покровская И.В., Чернявский А.Л.* Экспертные методы анализа и совершенствования систем управления // Автоматика и телемеханика. 2004, №10. – с. 172 – 188.
6 *Дорофеюк А.А., Дорофеюк Ю.А., Покровская И.В., Чернявский А.Л.* Метод независимой многовариантной экспертизы и его использование при решении прикладных задач. // Управление развитием крупномасштабных систем MLSD'2013): Труды Седьмой международной конференции. Том 1. – М.: ИПУ РАН, 2013. – с. 260-271.
7 *Крамер Г.* Математические методы статистики. // Пер. с англ. А.С.Монина и А.А.Петрова под редакцией академика А. Н. Колмогорова.// − М.: «МИР», 1975.
8 *Дорофеюк Ю.А., Гольдовская М.Д., Спиро А.Г.* Особенности компьютерной реализации и моделирования алгоритмов интеллектуального анализа сложноорганизованных данных. // Управление развитием крупномасштабных систем (MLSD'2013): Материалы Седьмой международной конференции. Т. 2. – М.: ИПУ РАН, 2013. – с. 328-331.
9 *Bellman R.* Dynamic Programming and Lagrange Multipliers. / Proc. Nat. Acad. of Sc. USA, 1956, vol.42, pp.767-769.

Дорофеюк Юлия Александровна. С.н.с. ИПУ РАН им. В.А.Трапезникова. К.т.н. Окончила в 2007 г. МИЭМ, ГТУ. Количество печатных работ: 124 (в т.ч. 2 монографии). Область научных интересов: интеллектуальные методы анализа данных, математическое моделирование в организационных, социальных, экономических, медико-биологических и технических системах; технологии интеллектуальной поддержки принятия решений; структурное прогнозирование. E-mail: dorofeyuk_julia@mail.ru.
Дорофеюк Александр Александрович. Гл.н.с. ИСА ФИЦ ИУ РАН. Зав. лабораторией ИПУ РАН им. В.А.Трапезникова. Профессор НИУ ВШЭ. Д.т.н., профессор. Окончил в 1965 г. МФТИ. Количество печатных работ: 232 (в т.ч. 15 монографий). Область научных интересов: математическая статистика, функциональный анализ, интеллектуальные методы анализа данных, методы экспертизы и анализа экспертных оценок, методы поддержки принятия решений, системный анализ. E-mail: daa2@mail.ru.
Покровская Ирина Вячеславовна. С.н.с. ИПУ РАН им. В.А.Трапезникова. К.т.н. Окончила в 1976 г. МГУ им. М.В.Ломоносова. Количество печатных работ: 61. Область научных интересов: интеллектуальные методы анализа данных, методы экспертизы и анализа экспертных оценок, методы поддержки принятия решений. E-mail: ivp750@mail.ru.
Спиро Арнольд Григорьевич. С.н.с. ИПУ РАН им. В.А.Трапезникова. К.т.н. Окончил в 1958 г. Ростовский политехнический институт (РосПИ). Количество печатных работ: 43. Область научных интересов: интеллектуальные методы анализа данных, методы экспертизы и анализа экспертных оценок, методы поддержки принятия решений. E-mail: spiro35@mail.ru.