УДК 62-50

ЭКСПЕРТНО-КЛАССИФИКАЦИОННЫЙ АНАЛИЗ ДАННЫХ В ЗАДАЧЕ ОЦЕНКИ ЭФФЕКТИВНОСТИ ФУНКЦИОНИРОВАНИЯ КРУПНОМАСШТАБНЫХ СИСТЕМ УПРАВЛЕНИЯ[8]

© Покровская И.В., Гольдовская М.Д., Дорофеюк Ю.А.

Институт проблем управления РАН

Abstract. In the paper the complex-organized data structural analysis methods and the results expert correction procedures in connection with of large-scale control systems efficiency problems are described. Algorithms of such structuring were developed on the base of range data analysis methods.

Введение

В работе рассмотрены методы классификационного анализа сложноорганизованных данных и процедур экспертной коррекции результатов такого анализа применительно к задачам оценки эффективности крупномасштабных систем управления. Главная проблема, с которой сталкиваются разработчики алгоритмов и процедур анализа подобных данных – это проблема структуризации и сжатия такой информации, Необходимо учитывать также большую размерность и многопрофильность исходной информации, – число объектов в подобных задачах может достигать десятков тысяч, а число показателей – нескольких сотен, при этом используются не только числовые, но и ранговые, номинальные, экспертные показатели. Кроме того, часто требуется провести динамический анализ функционирования исследуемых объектов за несколько лет Это приводит к тому, что оказывается малоэффективным использование классических методов статистической обработки и анализа подобных данных. По этим причинам главным становится построение сжатого, структурированного описания исходных данных, содержательно хорошо интерпретируемого и пригодного для подготовки и обоснования принимаемых управленческих решений. Такую структуризацию исследуемой системы предлагается получать методами классификационного анализа данных [1], дополненных экспертными процедурами коррекции. В работе описаны подобные методы для структуризации исходных параметров; множества объектов, входящих в исследуемую систему; траекторий объектов (изменение значений характеристик объектов во времени).

1. Методы структуризации параметров

Практика использования алгоритмов структуризации показывает, что структуризация объектов с учётом всех исходных параметров редко приводит к хорошим результатам. Именно поэтому для структуризации объектов используются не исходные, а информативные параметры, которых обычно существенно меньше. Практика показывает, что решить эту задачу чисто экспертным путём не удаётся. Мнения экспертов по поводу того, какие показатели являются более, а какие менее информативными, часто расходятся.

[8]Работа выполнена при частичной финансовой поддержке РФФИ, проекты 08-07-00349-а, 08-07-00427-а.

Наиболее плодотворным оказалось использование идеи экстремальной группировки параметров [2]. Она заключается в разбиении всего множества показателей на группы таким образом, чтобы показатели внутри групп как можно больше коррелировали друг с другом, а показатели из разных групп – как можно меньше. Математически задача формулируется как задача поиска разбиения, доставляющего максимум соответствующему критерию качества разбиения (отсюда – название «экстремальная группировка») [2]. Результатом экстремальной группировки являются группы параметров и факторы – синтезированные параметры-характеристики групп, каждый из которых является линейной комбинацией исходных параметров соответствующей группы и, в определённом смысле, её «центром». Когда такое разбиение найдено, множество показателей каждой группы можно заменить расчётным, синтетическим «средним» показателем, или фактором, и характеризовать каждый объект уже не исходным набором показателей, а гораздо меньшим набором факторов (близкая идея, но в несколько иной интерпретации, реализуется в факторном анализе). Если какой-либо из исходных показателей имеет достаточно высокую корреляцию с фактором, это можно интерпретировать как высокую значимость или «важность» показателя и использовать его для дальнейшего анализа вместо соответствующего фактора. Таким образом, за счет структуризации множества показателей удается значительно (как правило, на порядок) уменьшить их количество и, соответственно, упростить задачу анализа данных. При этом используются специальные экспертно-корректирующие процедуры для выбора конкретного типа алгоритма и основных его характеристик, таких как: выбор типа группировки, алгоритма фильтрации (в зависимости от уровня «зашумлённости» параметров), вида меры связи между параметрами и др.

При решении прикладных задач структурного анализа сложноорганизованных данных в основном используется алгоритм экстремальной группировки «квадрат» [2]. Опишем вкратце схему его работы.

Будем в дальнейшем коэффициент корреляции (или ковариации) $\rho_{x,y}$ двух случайных величин x и y обозначать как: $\rho_{x,y} = (x, y)$, подчеркивая этим обозначением тот факт, что коэффициент корреляции может пониматься как скалярное произведение случайных величин x и y. Для дисперсии $\rho_{x,x}$ случайной величины x будем применять обозначение $\rho_{x,x} = (x, x) = x^2$.

Пусть множество параметров (случайных величин) $x_1,\ x_2, ...,\ x_k$ разбито на непересекающиеся группы $A_1,\ A_2, ...,\ A_s$ и заданы случайные величины $f_1, f_2, ..., f_s$ такие, что $f_1^2 = f_2^2 = ... = f_s^2 = 1$, которые будем называть факторами. Введем в рассмотрение функционал

$$J^* = \sum_{x_i \in A_1} (x_i, f_1)^2 + \sum_{x_i \in A_2} (x_i, f_2)^2 + ... + \sum_{x_i \in A_s} (x_i, f_s)^2. \qquad (1.1)$$

Алгоритм «квадрат» решает задачу максимизации этого функционала как по разбиению параметров на множества $A_1,\ A_2, ...,\ A_s$, так и по выбору случайных величин $f_1, f_2, ..., f_s,\ f_l^2 = 1\ \ (l = 1, ..., s)$.

Максимизация функционала (1.1) соответствует интуитивному требованию такого разбиения параметров, когда в одну группу попадают наиболее «близкие» между собой параметры. Действительно, при максимизации функционала (1.1) для каждого фиксированного набора случайных величин $f_1, f_2, ..., f_s$ в l-ую группу будут попадать такие параметры, которые наиболее «близки» к величине f_l; в то же время среди всех возможных наборов случайных величин $f_1, f_2, ..., f_s$ будет отбираться такой набор, что каждая из величин f_l в среднем наиболее «близка» ко всем параметрам из своей группы.

Если заданы группы параметров $A_1, A_2, \ldots, A_s$, то максимум функционала J^* может быть получен, если в качестве факторов $f_1, f_2, \ldots, f_s$ выбрать такие случайные величины, что каждая случайная величина $f_l\,(l = 1, ..., s)$ удовлетворяет условию

$$\max_{f_l} \sum_{x_i \in A_l} (x_i, f_l)^2\,,\ \ f_l^2 = 1. \tag{1.2}$$

Фактор f_l, удовлетворяющий условию (1.2) при фиксированном множестве параметров A_l, находится по формуле

$$f_l = \frac{\sum\limits_{x_i \in A_l} \alpha_i x_i}{\sqrt{\left(\sum\limits_{x_i \in A_l} \alpha_i x_i\right)^2}} = \frac{\sum\limits_{x_i \in A_l} \alpha_i x_i}{\sqrt{\sum\limits_{x_i \in A_l, x_j \in A_l} \alpha_i \alpha_j\,(x_i, x_j)}}, \tag{1.3}$$

где α_i – компоненты собственного вектора матрицы $R_l = \{\,(x_i, x_j)\,\}\,,\ x_i, x_j \in A_l$ соответствующего её наибольшему собственному значению. С другой стороны, если величины $f_1, f_2, ..., f_s$ заданы, то разбиение параметров на группы A_1, A_2, ..., A_s обеспечивающее максимум функционала J^*, должно удовлетворять условию: для каждого $x_i \in A_l$

$$(x_i, f_l)^2 \geq (x_i, f_q)^2\ \ (q = 1, 2, ..., s)\,, \tag{1.4}$$

так как в противном случае функционал J^* можно было бы увеличить, перебросив параметр x_i из группы A_l в ту группу A_q для которой соотношение (1.4) не выполнено. Соотношения (1.2) и (1.4) в совокупности являются необходимыми условиями максимума функционала J^*.

Можно предложить следующий итерационный алгоритм, определяющий одновременно группы A_1, A_2, ..., A_s и факторы $f_1, f_2, ..., f_s$, удовлетворяющие этим условиям.

Пусть на p-м шаге итерации построено разбиение параметров на группы $A_1^{(p)}, ..., A_s^{(p)}$. Для каждой такой группы параметров строят факторы $f_l^{(p)}$ по формуле (1.3) и новое, $(p+1)$-е разбиение параметров $A_1^{(p+1)}, ..., A_s^{(p+1)}$ в соответствии с правилом: параметр x_i относится к группе $A_l^{(p+1)}$, если

$$\left(x_i, f_l^{(p)}\right)^2 \geq \left(x_i, f_q^{(p)}\right)^2\ \ (q = 1, 2, ..., s)\,. \tag{1.5}$$

В том случае, когда существуют два или более факторов и такой параметр x_i, что для этих факторов и этого параметра в (1.5) имеет место равенство, параметр x_i,

относится к одной из соответствующих групп произвольно. Предложенный выше алгоритм сходится к максимуму (возможно, локальному) функционала J^*, поскольку, каковы бы ни были факторы $f_1^{(p-1)}, ..., f_s^{(p-1)}$, на каждом шаге итерации функционал J^* не убывает.

2. Методы структуризации объектов

Классификация объектов производится в пространства X интегральных показателей, полученных на предыдущем этапе. Как и в предыдущем разделе, используются специальные экспертно-корректирующие процедуры для: выбора конкретного алгоритма классификации объектов; определения вида критерия качества классификации; выбора типа фильтрации (например, классификация строится с фоновым классом или без такового); выбора типа размытости – четкая, размытая, с размытыми границами, четкая с размытым фоном, размытая с четким фоном и т.д. Результатом классификации является вектор функций принадлежности объектов к классам $(h_1(x), \ldots, h_r(x))$, r – число классов, а также описание самих классов (например, эталонов) [1]. Для того чтобы результаты структуризации можно было использовать в практических задачах, важно не только насколько экономно она представляет исходную информацию, но и насколько эта структуризация удобна для интерпретации в содержательных терминах. В этой связи в приложениях в последнее время широко используются экспертно-классификационные алгоритмы построения так называемых «хорошо интерпретируемых классификаций» [3]. В прикладных задачах мы в основном использовали комплексный алгоритм автоматической классификации [4], основу которого составляет алгоритм m-локальной оптимизации. Схема работы этого алгоритма состоит в следующем.

Вначале опишем работу алгоритма 1-локальной оптимизации. Для простоты изложения рассматривается случай двух классов $r = 2$. Пусть задано начальное разбиение R_0 всех объектов классифицируемой выборки $x_1, . . ., x_n$. Здесь каждый объект представляется точкой $x_j = (x_j^{(1.1)}, x_j^{(1.2)}, \ldots, x_j^{(k)})$, $j = 1, \ldots, n$ в $k-$мерном пространстве параметров X. Обозначим через $x_j \in A_1$ точки, относящиеся к первому классу, а через $x_j \in A_2$ – ко второму. Алгоритм итерационный, – на каждом шаге рассматривается одна точка из последовательности $x_1, . . ., x_n, x_1, . . ., x_n, x_1, . . .$ («зацикленная» исходная последовательность). Отнесение точки к одному из двух классов обозначается с помощью индекса $\rho(x_j) = \begin{cases} 1, \text{ если } x_j \in A_1 \\ -1, \text{ если } x_j \in A_2 \end{cases}$. Тогда алгоритм 1-локальной оптимизации определяется следующим образом: $\rho(x_j) = sign\,[J(x_j \in A_1) - J(x_j \in A_2)]$.

В итоге точка x_j относится к тому классу, при отнесении к которому, значение критерия J будет больше (если эти значения равны, то для определённости точка относится к классу с меньшим номером). Алгоритм заканчивает работу, если на некотором цикле среди точек $x_1, . . ., x_n$ не будет сделано ни одной «переброски» точки из класса в класс.

Алгоритм m – локальной оптимизации – это поэтапное применение к выборке алгоритмов s – локальной оптимизации, $s = 1 \div m$. На s – ом этапе алгоритм работает

по той же схеме, только на каждом его шаге происходит пробная «переброска» из класса в класс не одной, а s точек. Подсчитывается значение критерия J до и после «переброски», Принадлежность каждой из sточек к классу либо остаётся неименной (J до «переброски» больше, чем после), либо меняется на другой класс – в противном случае. В данном случае цикл – это число шагов, равное числу всевозможных различных наборов, в каждый из которых входит s точек, выбранных из n точек исходной выборки. Доказана сходимость алгоритма за конечное число шагов к локальному максимуму критерия J.

Разработан эвристический алгоритм сокращённого перебора, который на каждом шаге для пробной «переброски» использует s точек в определённом смысле ближайших к границе между классами.

В приложениях в качестве критерия J использовался функционал J_1 средней близости точек в классах, определяемый через потенциальную функцию близости точек x и y:

$$K(x, y) = 1/\{1 + \alpha\, R^p(x, y)\}, \tag{2.1}$$

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

$$K(A_i, A_i) = \frac{2}{n_i(n_i - 1)} \sum_{i=1}^{n_i} \sum_{j>i} K(x_i, x_j), \tag{2.2}$$

где $K(x_i, x_j)$ определяется формулой (2.1), n_i – число точек в классе A_i. Тогда критерий J_1 определяется как: $J_1 = \sum\limits_{i=1}^{r} \frac{n_i}{n} K(A_i, A_i)$, где $K(A_i, A_i)$ определяется с помощью формулы (2.2).

3. Методы структуризации динамических объектов

При исследовании многопараметрической информации, изменяющейся во времени, непосредственное использование обычных алгоритмов автоматической классификации невозможно. Были разработаны специальные алгоритмы динамического классификационного анализа (ДКА), когда каждый объект по каждому параметру характеризуется набором значений для некоторой последовательности моментов времени (траекторией) [4]. В рамках вариационного подхода разработан соответствующий алгоритм ДКА. Проведен теоретический анализ этого алгоритма. Показано, что оптимальную классификацию можно искать в узком классе так называемых эталонных классификаций, и она определяется градиентом исходного функционала (критерия качества).

Постановка задачи ДКА формулируется с использованием трех основных понятий: классифицируемое множество объектов, класс допустимых классификаций и функционал качества разбиения [1].

1) Классифицируемое множество объектов

В ДКА предлагается классифицировать конечное множество объектов, изменяющихся во времени. Пусть в каждый момент времени объекты описываются некоторым конкретным набором параметров $x^{(1.1)}, ..., x^{(k)}$. Считается, что для каждого объекта последовательно фиксируются m значений каждого из параметров в соответствующие моменты времени. Таким образом, каждый объект характеризуется серией из m векторов $x_1, ..., x_m$ в k-мерном пространстве параметров, представляющих собой траекторию изменения данного объекта в пространстве параметров. Такую траекторию будем обозначать через $\tilde{x} = (x_1, ..., x_m)$. Итак, в качестве классифицируемого множества будет рассматриваться конечное множество объектов, задаваемых своими траекториями фиксированной длины, т.е. необходимо классифицировать множество $X = \{\tilde{x}_1, ..., \tilde{x}_n\}$ $(X \subseteq \mathrm{R}^{k*m})$.

Заметим, что важной особенностью такого подхода является то, что моменты времени, в которые зафиксированы значения параметров, у разных объектов могут быть разные.

2) Класс допустимых классификаций

Размытой классификацией множества $X = \{\tilde{x}_1, ..., \tilde{x}_n\}$ на r классов с фоновым классом называется $(r+1)$-мерная вектор-функция $H(\tilde{x}) = (h_0(\tilde{x}), h_1(\tilde{x}), ..., h_r(\tilde{x}))$ [1]. Здесь $h_0(\tilde{x})$ – функция принадлежности $\tilde{x}$ к фоновому классу, а $h_i(\tilde{x})$ – функция принадлежности его к i-му классу. Для любого $\tilde{x}$ значение $H(\tilde{x})$ должна принадлежать некоторому ограниченному замкнутому множеству V пространства значений вектор-функции H, т.е. $H(\tilde{x}) \in V \subseteq \mathrm{R}^{r+1}$. Множество V определяет тип размытости для данной задачи. Итак, рассматривается следующий класс размытых классификаций: $\Xi(V) = \{H : \forall \tilde{x} \in X\, H(\tilde{x}) \in V\}$.

3) Критерий качества классификации

В рамках общего вариационного подхода всё рассмотрение проводится для произвольного выпуклого функционала Ф от вектор-функции H. Для ДКА предлагается строить критерий качества классификации в соответствии с методом обобщенного среднего [1].

Считается, что объекты одного и того же класса искомой классификации должны хорошо описываться некоторой моделью траектории этого класса, а объекты, чьи траектории плохо описываются всеми моделями классов, должны попасть в фоновый класс. Поэтому критерий качества должен отражать, во-первых, близость траекторий объектов внутри нефоновых классов и, во-вторых, отнесение к фоновому классу объектов, чьи траектории достаточно удалены от моделей нефоновых классов. Далее, вводится в рассмотрение множество Λ возможных моделей траекторий классов. Между элементами множества объектов X и элементами множества моделей Λ вводится мера близости $K(\tilde{x}, \tilde{\alpha})$ $(\tilde{x} \in X; \tilde{\alpha} \in \Lambda)$.

В таком случае величина $K(h(\tilde{x}), \tilde{\alpha}) = \sum_{j=1}^{n} K(\tilde{x}_j, \tilde{\alpha}) h(\tilde{x}_j)$ отражает меру того, насколько хорошо модель $\tilde{\alpha}$ описывает точки множества, заданного через свою функцию принадлежности $h(\tilde{x})$.

Обобщенным средним или эталоном множества, заданного функцией принадлежности $h(\tilde{x})$, называется модель [1]:

$$\tilde{\alpha}_h = \arg\max_{\tilde{\alpha}\in\Lambda} K(h\,(\tilde{x})\,,\tilde{\alpha}) \tag{3.1}$$

В соответствии с этим вводится следующий критерий качества классификации

$$J\,(H) = \sum_{i=1}^{r} K(h_i\,(\tilde{x})\,,\tilde{\alpha}_{h\,i}) + B\sum_{j=1}^{n} h_0\,(\tilde{x}_j), \tag{3.2}$$

где $\tilde{\alpha}_{h\,i}$ – эталон i-го класса (3.1), а B – некоторая константа, регулирующая отнесение объектов к фоновому классу.

Задача классификации состоит в максимизации функционала (3.2) по вектор-функциям принадлежностей объектов к классам H. Одновременно, в каждом классе выстраивается эталонная траектория, отражающая общую тенденцию изменения значений показателей для объектов данного класса.

Заключение

Описанная методика экспертно-классификационного анализа сложно-организованных данных использовалась при решении широкого круга задач анализа и оценки эффективности функционирования крупномасштабных систем управления, в том числе при оценке эффективности управления социально-экономическим развитием субъектов РФ, при анализе и совершенствовании управления: региональным здравоохранением (на примере Новгородской области), региональными пассажирскими автоперевозками (на примере Московской области), жилищно-коммунальным хозяйством крупного города (на примере Москвы) и ряда других. Полученные результаты свидетельствуют о высокой эффективности разработанных методов, алгоритмов и процедур.

Список литературы

1. *Бауман Е.В., Дорофеюк А.А.* Классификационный анализ данных. // Труды Международной конференции по проблемам управления. Том 1. – М.: СИНТЕГ, 1999. – С. 62-77.
2. *Браверман Э.М., Мучник И.Б.* Структурные методы обработки эмпирических данных. – М.: Наука, 1983. – 464 с.
3. *Дорофеюк А.А., Чернявский А.Л.* Алгоритмы построения хорошо интерпретируемых классификаций. / Проблемы управления. №2, 2007. – С. 83-84.
4. *Дорофеюк Ю.А.* Комплексный алгоритм автоматической классификации и его применение для анализа и принятия решений в больших системах управления. / Теория активных систем. Труды международной научно-практической конференции. / – М.: ИПУ РАН. 2007. – С. 39 -42.
5. *Чернявский А.Л., Бауман Е.В., Дорофеюк А.А.* Методы динамического классификационного анализа данных. Искусственный интеллект, № 2, 2002, с. 290-298.

Статья поступила в редакцию 27.04.2008