УДК 62-50

МЕТОДОЛОГИЯ ЭКСПЕРТНО-КЛАССИФИКАЦИОННОГО АНАЛИЗА В ЗАДАЧАХ УПРАВЛЕНИЯ И ОБРАБОТКИ СЛОЖНООРГАНИЗОВАННЫХ ДАННЫХ (история и перспективы развития)

А.А. Дорофеюк

Рассмотрена история и перспективы развития методологии структурно-классификационного анализа сложноорганизованных данных — стремительно развивающееся как в России, так и за рубежом направления, «выросшего» из статистических методов обработки данных и распознавания образов. Описаны как теоретические, так и прикладные полученные результаты.

Ключевые слова: экспертно-классификационный анализ; автоматическая классификация, экстремальная группировка параметров; кусочная аппроксимация сложных зависимостей; структурное прогнозирование; многовариантная экспертиза.

ВВЕДЕНИЕ

Основу методологии структурно-классификационного анализа данных заложили работа Э.М. Бравермана 1960 г. и последующие работы Э.М. Бравермана и А.А. Дорофеюка 1963—1966 гг., которые выполнялись в лаборатории Института автоматики и телемеханики (ныне Институт проблем управления им. В.А. Трапезникова РАН), руководимой М.А. Айзерманом. Зарождение и пути развития этой тематики подробно рассмотрены и проанализированы в статье [1] из сборника, посвящённого памяти выдающегося советского и российского учёного — Марка Ароновича Айзермана. Отметим только, что именно М.А. Айзерману принадлежит идея формирования нового научного направления, связанного с моделированием способности многих биологических объектов к самообучению, к формированию новых понятий, способности автономно группировать наблюдаемые объекты (или события) в классы «подобных». Сейчас это стремительно развивающееся направление, идеологически близкое к распознаванию образов, называют «интеллектуальным анализом данных», «структурно-классификационным анализом», «разведочным анализом», на Западе оно известно, в основном, под названием «DATA MINING».

Основная идея этого направления состоит в следующем. Пусть имеется некоторый массив данных, описывающий состояние исследуемой группы из n объектов. Обычно под этим подразумевается, что имеется k параметров (числовых, качественных или номинальных), значения которых $x_j^{(i)}$ ($i = 1, ..., k, j = 1,..., n$), и определяют этот массив как матрицу данных. В динамических случаях значения параметров изменяются со временем $x_j^{(i)}(t)$, поэтому массив становится трёхмерным (куб данных). Необходимо выявить структуру этого массива для построения сжатого, содержательно хорошо интерпретируемого описания исследуемых объектов с целью: идентификации основных характеристик их функционирования, выявления и прогнозирования интегральных показателей поведения объектов во времени, поиска закономерностей их взаимодействия и т. д. При этом выявление структуры производится по всем трём «направлениям» куба данных — структуре объектов в различных подпространствах параметров, структуре взаимосвязи параметров по данным имеющего-

ся массива, структуре динамических характеристик объектов (например, структуре траекторий параметров или интегральных показателей исследуемых объектов).

1. СТРУКТУРИЗАЦИЯ ОБЪЕКТОВ

Для выявления структуры объектов широко применяются методы структурного анализа данных (другие названия: автоматическая классификация, кластер-анализ, самообучение, распознавание образов без учителя, численная таксономия, стратификация и др.). Основы теории и алгоритмической базы этого направления были заложены Э.М. Браверманом и А.А. Дорофеюком в 1961—1966 гг. в ИПУ РАН (в то время — ИАТ), а затем продолжены А.А. Дорофеюком и сотрудниками его лаборатории — Е.В. Бауманом, А.Л. Чернявским, И.Б. Мучником, Н.Е. Киселёвой, В.Г. Мовсумовым, А.Г. Дмитриевым, Ю.А. Дорофеюк. В Институте этой же проблематикой занимались также Я.З. Цыпкин, Г.К. Кельманс, Р.П. Агаев, В.Н. Вапник, А.Я. Червоненкис, А.Р. Стефанюк, А.А. Журавель, Касавин А.Д., В.Я. Лумельский. Из других научных центров, активно занимавшихся задачами этого направления, отметим ВЦ РАН (Ю.И. Журавлёв, К.В. Рудаков, В.В. Моттль, К.В. Воронцов и др.), ЦЭМИ РАН (С.А. Айвазян, З.И. Бежаева, И.С. Енюков, В.С. Мхитарян, О.В. Староверов, и др.), ЦНИИКА (И.Ш. Торговицкий, И.П. Баумштейн и др.), Институт математики СО АН СССР (Н.Г. Загоруйко, Г.С. Лбов, В.Н. Елкина и др.), ИЭОПП СО АН СССР (Б.Г. Миркин, В.Л. Купершток и др.), Горьковский (ныне Нижегородский) Государственный университет (Ю.И. Неймарк и его сотрудники), Институт кибернетики АН УССР (М.И. Шлезингер, В.А. Ковалевский, А.В. Миленький, А.Г. Ивахненко и др.). Из зарубежных центров упомянем университеты в Мэриленде, Лос-Анжелесе, Энн-Арборе, Пэдью (США); Риме, Солерно, Удине, Милане, Национальную лабораторию кибернетики в Неаполе (Италия); Марселе, Экс-ин-Провансе, Бордо, Гренобле, университет Париж-IX, ИНРИА (Франция); технические университеты в Аахене и Брауншвейге (Германия), Международный институт прикладного системного анализа (Австрия). С большинством из них Институт долгие годы поддерживал плодотворное сотрудничество в рамках взаимного командирования специалистов, организации и проведения конференций и симпозиумов, межправительственных соглашений (Италия, Франция, Австрия), а также прямых договоров и соглашений между Академией наук или Институтом и зарубежными научными центрами — университеты в Удине, Солерно, Риме (Италия), ИНРИА (Франция), Мичиганский университет (Энн-Арбор, США).

1.1. Содержательная постановка задачи автоматической классификации

Пусть исследуется некоторое множество n объектов. Предположим, что все объекты разделены по своим свойствам на r классов, причем объекты с близкими свойствами попадают в один и тот же класс, а с существенно различными — в разные классы. Кроме того, будем предполагать, что каждый такой объект характеризуется значениями некоторого заранее выбранного набора из k параметров $\{x^{(1)}, ..., x^{(k)}\}$, причём предполагается, что этот набор достаточно полно характеризует свойства исследуемых объектов, которые необходимо учитывать при их классификации. Введем в рассмотрение k-мерное пространство параметров X, в котором i-й оси соответствуют значения параметра $x^{(i)}$, т. е. j-му объекту в пространстве X соответствует точка $x_j = (x_j^{(1)}, ..., x_j^{(k)})$. В соответствии со сделанными предположениями, близким в пространстве X точкам будут соответствовать объекты с близкими свойствами. Тогда задачу выявления структуры объектов можно поставить как задачу разбиения пространства X на такие r областей, чтобы близкие точки исходной выборки, как правило, попадали в одну и ту же область (задача автоматической классификации). Для случая конечной выборки задача сводится к выделению r изолированных, «компактных» групп точек (классов) исходной выборки в k-мерном пространстве параметров X.

За 10 лет (1960—1970 гг.) было опубликовано огромное число работ, в которых предлагались различные эвристические алгоритмы автоматической классификации, опирающиеся на содержательную постановку этой задачи, когда либо не формализован критерий качества классификации, либо не доказано, что алгоритм экстремизирует какой-либо формальный критерий. Разнообразие и особенности таких алгоритмов достаточно полно отражены в обзоре [2].

1.2. Формальная постановка задачи автоматической классификации

Впервые формальная постановка задачи автоматической классификации была сделана М.И. Шлезингером в 1963 г. (опубликована в 1965 г.). Он сформулировал критерий качества разбиения следующего общего вида:

$$R = \sum_{j=1}^{r} p_j \int_{A_j} \int_{A_j} S(x, y) P(x/j) P(y/j) dx dy, \qquad (1)$$

где $S(x, y)$ — потери от отнесения точек x и y к одному и тому же классу A_j, $P(x/j)$ — условная плотность распределения вероятностей в классе A_j, p_j —

априорная вероятность класса A_j. Однако в дальнейшем он рассматривал только случай конечного классифицируемого множества точек с квадратичной функцией потерь $S(x, y) = (x - y)^2$. В этом случае критерий (1) принимает вид средневзвешенной дисперсии точек в классах $R(r) = \sum_{j=1}^{r} \sum_{x_i \in A_j} (x_i - c_j)^2$, где c_j — центр тяжести точек в классе A_j.

В это же время (1963—1966 гг.) в ИАТе были введены в рассмотрение формальные критерии качества классификации и разработаны алгоритмы их экстремизации. Здесь следует выделить два случая — конечное или бесконечное множество классифицируемых точек. Первые работы относились к существенно более простому случаю конечного числа классифицируемых точек. Был предложен целый ряд критериев качества классификации, базирующихся на характеристиках средней близости точек в классах и средней близости (удаленности) самих классов. В статье [3] был введён класс критериев $I = f(I_1, I_2)$, причем I должно увеличиваться с увеличением I_1 и уменьшаться с уменьшением I_2. К такому классу относятся, например, критерии

$$I_3 = I_1 - qI_2, \quad I_4 = \frac{I_1}{I_2}, \quad I_5 = \frac{I_1 - qI_2}{I_1 + qI_2}, \qquad (2)$$

где q — некоторая константа, корректирующая разномасштабность величин I_1 и I_2. В выражениях (2) критерий I_1 — это средняя по классам мера близости точек в классах $I_1 = \frac{1}{r} \sum_{i=1}^{r} K(A_i, A_i)$, где $K(A_i, A_i) = \frac{2}{n_i(n_i - 1)} \sum_{i=1}^{n_i} \sum_{j>i} K(x_i, x_j)$, а критерий I_2 — это средняя мера близости (удаленности) классов друг от друга $I_2 = \frac{2}{r(r-1)} \sum_{i=1}^{r} \sum_{j>i} K(A_i, A_j)$, где

$$K(A_i, A_j) = \frac{1}{n_i n_j} \sum_{x_l \in A_i} \sum_{x_s \in A_j} K(x_l, x_s). \qquad (3)$$

Здесь $K(x, y) = \frac{1}{1 + \alpha R^p(x, y)}$ — потенциальная функция, n_i — число точек в классе A_i, α и p —настраиваемые параметры. Ясно, что классификация тем лучше, чем больше I_1 и чем меньше I_2.

Был разработан набор алгоритмов экстремизации введённых критериев для конечного n [3]. В последние годы для решения этой задачи был разработан высокоэффективный алгоритм m-локальной оптимизации (в составе комплекса алгоритмов структурно-классификационного анализа данных) [4].

Существенно более сложным является случай бесконечной классифицируемой последовательности объектов. Здесь можно использовать только рекуррентные алгоритмы. Первый рекуррентный алгоритм для такого случая был предложен и теоретически исследован Э.М. Браверманом [5], там же была доказана его сходимость. Критерий качества классификации (экстремизируемый функционал) и сам алгоритм формулируются на языке метода потенциальных функций [6], а именно: в спрямляющем пространстве Z (см. книгу [6]) критерий качества, предложенный в работе [5], является частным случаем критерия (1) для квадратичной функции потерь и может быть записан в виде (для наглядности рассмотрен случай $r = 2$)

$$K_1 = \int_A (z - z_A)^2 P(z)dz + \int_B (z - z_B)^2 P(z)dz = \frac{(M^A)^2}{p^A} + \frac{(M^B)^2}{p^B}, \qquad (4)$$

где z_A и z_B — центры классов A и B соответственно, $P(z)$ — функция плотности распределения вероятностей появления точек классифицируемой последовательности, $M^A = \int_A zP(z)dz$ — первый ненормированный момент класса A, а $p^A = \int_A zP(z)dz$ — априорная вероятность класса A (нулевой ненормированный момент). Аналогично определяются соответствующие величины для класса B.

В работе [5] была доказана важная теорема, позволяющая по виду аддитивного экстремизируемого функционала выбирать класс разделяющих поверхностей. В частности, для функционалов вида (4), зависящих только от ненормированных моментов не выше первого, можно брать линейные разделяющие функции $f(z) = (c, z) - a$. Для предложенного в работе [5] алгоритма впервые для бесконечного случая была доказана его сходимость, обеспечивающая стационарное значение (4).

В работе [7] предложены рекуррентные алгоритмы, являющиеся непосредственным обобщением алгоритмов, разработанных в работе [3] для конечного n, на случай бесконечной последовательности. В этом случае аналоги критериев I_1, I_2 и I_3 могут быть записаны как выпуклые функционалы от нулевых и первых ненормированных моментов. В работе [8] была доказана теорема (обобщение теоремы Э.М. Бравермана [5]), в соответствии с которой оптимальные разделяющие функции для таких критериев можно также искать в классе линейных.

Отметим, что, в отличие от распознавания образов с учителем, теоретическое исследование сходимости рекуррентных алгоритмов автоматической классификации (распознавания образов без учителя) невозможно проводить классическими методами стохастической аппроксимации ввиду невыпуклости экстремизируемого функционала [2].

1.3. Вариационный подход

Следующий период исследований задач автоматической классификации был связан с так называемым вариационным подходом, т. е. рассмотрением уравнений, следующих из необходимых условий экстремума функционала качества классификации (равенства нулю первой его вариации). Теоретическую базу таких исследований заложил Э.М. Браверман [5], реализовав вариационный подход для конкретного критерия качества классификации (4). Эта работа была далее обобщена на существенно более широкий класс функционалов Е.В. Бауманом и А.А. Дорофеюком [9].

1.4. Размытая автоматическая классификация

Начиная с работы [8], задачи автоматической классификации в общей постановке исследуются для случая размытой классификации, когда вместо характеристических функций классов вводятся функции принадлежности к классу. Другими словами, размытая классификация задается r-мерной вектор-функцией $H(x) = (h_1(x), ..., h_r(x))$, где $h_i(x)$ — функция принадлежности x к i-му классу. Функция $H(x)$ удовлетворяет следующим условиям: $H(\cdot) \in L_2(X, P)$, и для любого x значение $H(x)$ принадлежит некоторому ограниченному множеству V пространства значений вектор-функции $H(x)$, т. е. $H(x) \in V \subseteq R^k$. Путем выбора ограничивающего множества V можно получить различные типы размытости, а именно — чёткую классификацию, размытую классификацию и классификацию с размытыми границами.

В работе [8] был рассмотрен критерий качества классификации достаточно общего вида:

$$\Phi = \Phi_2(\mu(H)), \qquad (5)$$

где Φ — выпуклый функционал, $\mu(H) = (p_i, M_i; i = 1, ..., r)$. Значительная часть известных критериев качества классификации точек евклидова пространства является частным случаем функционала (5). В работе [8] предложен алгоритм максимизации критерия (5), доказана его сходимость к стационарному значению для случая строго выпуклого, дважды непрерывно дифференцируемого функционала (5).

В последующем был рассмотрен ещё более широкий класс критериев качества классификации — произвольный выпуклый функционал $\Phi_3 = \Phi_3(H)$ от вектор-функции $H(x)$. Было показано, что к этому классу относится не только подавляющее большинство известных критериев качества классификации (в том числе функционалы в неметрических шкалах), но и широкий класс функционалов, используемых в других задачах анализа данных (кусочная аппроксимация сложных зависимостей, экстремальная группировка параметров, диагонализация матрицы связи и др.). С этого времени область применения методов автоматической классификации расширилась настолько, что появилось новое направление, получившее весьма общее название «анализ данных». Это направление, в отличие от традиционных статистических методов, требующих для своего применения некоторой вероятностной модели (построение которой достаточно трудная, а иногда и принципиально неразрешимая задача) предназначено для «разведочного» анализа многомерных массивов сложноорганизованных данных [10]. Соответствующие алгоритмы стали называться алгоритмами классификационного (структурно-классификационного) анализа данных.

Для исследования вида оптимальной размытой классификации важно понятие *опорной размытой классификации* $H_F(x)$ для произвольного линейного функционала $F(H)$: $H_F(x) = \arg\max_{H \in V}(F(x), H)$. Доказана теорема о том, что оптимальная размытая классификация принадлежит классу опорных классификаций.

Этот результат позволяет построить итерационный алгоритм максимизации функционала $\Phi = \Phi_3$. Основу алгоритма составляют два правила: правило нахождения опорной классификации по данному линейному функционалу $F(H)$ и правило нахождения по результатам классификации такого функционала, который был бы субградиентом исходного функционала [10]. Доказана теорема о сходимости этого алгоритма.

1.4.1. Размытая классификация с фоновым классом

Во многих задачах классификационного анализа приходится классифицировать объекты одинаково далёкие от всех классов, например, при грубых ошибках наблюдений или при неправильно выбранном числе классов (заниженном по отношению к истинному). Был введён в рассмотрение специальный класс, в пределах которого не учитывается близость объектов друг к другу, который был назван фоновым [10]. При наличии фонового класса размытая классификация задается вектор-функцией $H(x) = (h_0(x), h_1(x), ..., h_r(x))$, где $h_0(x)$ — функция принадлежности x к фоновому классу. При исследовании размытой классифика-

ции с фоновым классом в дополнение к уже рассмотренным выше трём типам размытости появляются ещё дополнительные варианты, например, размытая классификация с чётким фоновым классом.

2. СТРУКТУРИЗАЦИЯ ПАРАМЕТРОВ

При решении задач автоматической классификации объектов достаточно часто возникала проблема размерности пространства параметров X и выбора набора информативных (в смысле получения качественной классификации) параметров. Именно тогда возникла задача структуризации параметров. Формально задача ставится как задача нахождения такой классификации (группировки) параметров и таких эталонов классов (в этой задаче они называются факторами), обеспечивающих экстремальное значение некоторого заданного критерия качества такой группировки, имеющего интуитивно понятный содержательный смысл. В работе [11] была поставлена задача *экстремальной группировки параметров*, которая, в определённом смысле, является обобщением задачи факторного анализа. Были разработаны и теоретически изучены два алгоритма экстремальной группировки параметров, отличающиеся видом критерия качества группировки [11]. В обоих критериях в качестве меры связи (или «близости») параметров используется коэффициент корреляции (ковариации). Как эти алгоритмы, так и их многочисленные модификации широко применяются до сих пор, как независимо — в задачах анализа структуры конкретных наборов параметров, так и в составе программно-алгоритмических комплексов, предназначенных для решения крупномасштабных задач структурного анализа больших массивов сложноорганизованных данных [4, 12].

Одна из важнейших задач структурно-классификационного анализа — задача выделения (иногда — построения) так называемых «информативных» параметров. Дело в том, что практически все алгоритмы структуризации достаточно чувствительны к присутствию в исходном наборе «шумящих» или «малоинформативных» параметров, т. е. параметров слабо связанных с основными характеристиками исследуемой системы (по сравнению с другими, «информативными» параметрами). Наличие таких параметров приводит к «размыванию» исследуемой структуры, а при их значимом числе — к серьёзному её искажению.

Алгоритмы экстремальной группировки параметров дают хороший инструмент получения набора информативных параметров. А именно, в качестве информативных предлагается выбирать либо синтетические параметры — факторы, которые являются аналогами центров классов в задаче классификации объектов, либо по 1—2 параметра из каждой группы, ближайших к соответствующему фактору [4].

Несколько особняком стоит задача структуризации номинальных признаков, которая стала весьма актуальной в последнее время в связи с решением прикладных задач структуризации для крупномасштабных слабо формализованных систем управления [13]. В таких задачах приходится рассматривать десятки, а иногда и сотни классификаций объектов, входящих в исследуемую систему (для различных — пространств признаков, видов выбранной метрики в этом пространстве, значений свободных параметров применяемого алгоритма, различных типов алгоритмов и т. д.). Задача исследования такого множества классификаций, как правило, неподъёмная для эксперта-прикладника. Учитывая, что каждая классификация — это n-позиционный, r-градационный номинальный признак (n — число объектов, r — число классов), то задача структуризации множества классификаций эквивалентна задаче структуризации соответствующих номинальных признаков [14].

Интересные результаты получены в задаче структуризации параметров долевого типа, широко используемых в демографии, медицинской статистике, социологии, при обработке результатов переписи населения и др. [15].

3. МЕТОДЫ СТРУКТУРНОЙ АППРОКСИМАЦИИ СЛОЖНЫХ ЗАВИСИМОСТЕЙ

В конце 1960-х гг. интенсивно велись работы по применению разработанных алгоритмов автоматической классификации для решения целого ряда прикладных задач. В процессе решения некоторых из них появились новые постановки задач структурного анализа данных. Самой интересной как с теоретической, так и с прикладной точки зрения оказалась задача *кусочной аппроксимации сложных зависимостей*. Исторически она вначале формулировалась как задача идентификации статической характеристики некоторого технологического объекта (процесса), функционирующего в нескольких режимах. Дадим более подробно содержательную постановку этой задачи.

Рассмотрим технологический объект, состояние которого достаточно точно описывается вектором значений контролируемых входных параметров $x = \{x^{(1)}, ..., x^{(k)}\}$, $x \in X$, где X — пространство входных параметров. Эффективность работы объекта определяется значениями выходного параметра y. Необходимо идентифицировать статическую характеристику объекта, другими словами, моделью объекта служит функциональный преобразователь $y = F(x)$, где $F(x)$ — неизвестная функ-

ция. Обычно для такой идентификации по известным значениям векторов входных параметров $x_1, ..., x_n$ и соответствующих значений выходного параметра $y_1, ..., y_n$ строится аппроксимация $y = \tilde{F}(x)$, для которой заданный критерий качества аппроксимации J принимает экстремальное значение. Обычно таким критерием служит остаточная дисперсия y относительно аппроксимирующей функции $\tilde{F}(x)$, т. е. функционал вида

$$J = \int_X [y - \tilde{F}(x)]^2 dP(x). \qquad (6)$$

Существует ряд методов решения этой задачи — метод наименьших квадратов, метод максимального правдоподобия, классические алгоритмы регрессионного и корреляционного анализа, процедуры типа стохастической аппроксимации и др. Все эти методы предполагают априорный выбор класса аппроксимирующих функций $\tilde{F}(x, \alpha)$, который обычно задаётся параметрически — с помощью векторного параметра α.

В практических задачах объем имеющегося статистического материала жёстко ограничивает число оцениваемых параметров в смысле статистической достоверности получаемых результатов. А это означает, что для сложных функций $\tilde{F}(x, \alpha)$ требуется оценивать слишком большое число параметров, что невозможно на ограниченном материале. Однако в процессе анализа реальных объектов было замечено, что во многих случаях статическая характеристика, хотя и является сложной функцией во всей допустимой области изменения вектора входных параметров, но может быть представлена как совокупность достаточно простых функций $F_j(x)$ в пределах отдельных областей B_j пространства входов X, соответствующих различным режимам функционирования объекта. Другими словами, статическая характеристика сложного вида может быть представлена как совокупность достаточно простых «кусков».

Это означает, что аппроксимируемая функция $y = F(x)$ может быть представлена в виде $y = \sum_{j=1}^{r} h_j(x) F_j(x)$, где $h_j(x)$ — функции принадлежности x областям B_j, на которые разбивается пространство X (или область определения функции $F(x)$). Как уже говорилось в п. 1.4, вид $h_j(x)$ определяется выбранным типом размытости.

Аналогично определяется вид аппроксимирующей функции

$$\tilde{F}(x, \alpha) = \sum_{j=1}^{r} h_j(x) \tilde{F}_j(x, \alpha). \qquad (7)$$

В этом случае функционал (6) имеет вид

$$J = \sum_{j=1}^{r} \int_X h_j(x)[y - \tilde{F}_j(x, \alpha)]^2 dP(x). \qquad (8)$$

Для нахождения по статистическим данным аппроксимирующей функции (7), минимизирующей значение функционала (8), были разработаны специальные *методы кусочной аппроксимации*, существенно использующие алгоритмы автоматической классификации. Первые публикации на эту тему касались задач контроля качества сложных изделий (кусочно-постоянная или ступенчатая аппроксимация) [16] и идентификации статической характеристики промышленного объекта (кусочно-линейная и кусочно-полиномиальная аппроксимация) [17]. Исчерпывающее описание алгоритмов решения последней задачи содержится в брошюре [18].

Алгоритмы кусочной аппроксимации можно условно разделить на одноэтапные и двухэтапные. В одноэтапных алгоритмах поиск оптимального в смысле критерия (8) разбиения $\{B_j\}$, $j = 1 \div r$, и соответствующих локальных аппроксимаций $\tilde{F}_j(x, \alpha)$, $j = 1 \div r$, производится одновременно. В двухэтапных — предполагается, что область значений входных параметров в пространстве X, соответствующих одному и тому же режиму функционирования объекта, является достаточно компактным кластером. Поэтому вначале производится автоматическая классификация выборочных значений входных параметров, которая порождает разбиение пространства X на области B_j, соответствующие различным режимам функционирования объекта. На втором этапе для этого разбиения находятся оптимальные локальные регрессии $\tilde{F}_j(x, \alpha)$, $j = 1 \div r$.

Наибольший интерес в смысле приложений представляют рекуррентные алгоритмы кусочной аппроксимации, поскольку они позволяют проводить идентификацию объекта в реальном времени (в режиме нормальной эксплуатации). Кроме того, на базе рекуррентных алгоритмов достаточно просто реализовать адаптивные схемы идентификации, позволяющие отслеживать медленные изменения статической характеристики объекта (например, в нефтехимии это происходит из-за старения катализатора). Такие алгоритмы были разработаны на базе вариационного подхода [18]. Однако теоретический анализ сходимости таких алгоритмов сопряжён с существенными трудностями, которые удалось преодолеть только после доказательства того, что задача кусочно-линейной аппроксимации является частным случаем задачи автоматической классификации [19]. Впослед-

ствии были разработаны оптимальные алгоритмы кусочно-линейной аппроксимации [20].

Для задач кусочной аппроксимации была предложена оригинальная иерархическая схема одновременного поиска наборов информативных переменных и локальных аппроксимаций, названная методом *иерархической кусочной аппроксимации* [21]. Идея этого метода состоит в следующем. Разбиение пространства входов X на области B_j подразумевает, что эти области соответствуют различным режимам функционирования объекта. А это, в свою очередь, может означать, что для каждого режима может быть свой набор информативных входных переменных. Другими словами, в таком анизотропном случае информативные переменные необходимо искать для каждой области B_j независимо. Подобное рассуждение справедливо не только для всего пространства X, но и для каждой области B_j (каждый режим функционирования объекта может распадаться на подрежимы) и т. д.

В процессе решения прикладных задач, связанных с аппроксимацией сложных зависимостей, было замечено, что для многих объектов в промышленности, экономике, геологии и других областях искомая зависимость $y = F(x)$ имеет следующую структуру: на фоне некоторой, как правило простой, зависимости $y = f(x)$ (основная закономерность, тренд, тенденция и т. д.) в отдельных (аномальных) областях B_j^* пространства X (но не обязательно во всех) наблюдаются существенные отклонения от $f(x)$. Другими словами, искомую функцию $F(x)$ в таких случаях целесообразно представлять как композицию двух функций — глобальной составляющей $f(x)$ и локальных функций отклонения от неё $F_j^*(x)$ в аномальных областях B_j^*. В этом случае аппроксимирующую функцию $\tilde{F}(x)$ следует искать в виде: $\tilde{F}(x, \alpha) = f(x, \alpha) + \sum_{j=1}^{r} \varepsilon_j^*(x)\tilde{F}_j^*(x, \alpha)$, где $\varepsilon_j^*(x)$ — характеристическая функция аномальной области B_j^* (принимает значение 1 только для точек этой области). Задача нахождения такой функции была названа *задачей комбинированной кусочной аппроксимации*, были разработаны алгоритмы решения этой задачи, существенно использующие процедуры кусочно-линейной аппроксимации [22].

Отметим ещё один интересный алгоритм кусочно-линейной аппроксимации второго типа, в котором при построении аппроксимации анализируется не только близость областей B_j, но и, в определённом смысле, близость локальных регрессий $\tilde{F}_j(x, \alpha_j)$ в этих областях [23]. Идея этого алгоритма состоит в следующем. Вначале по выборочным значениям входных параметров с помощью одного из алгоритмов автоматической классификации пространства X разбивается на $r_{\text{нач}}$ областей, где $r_{\text{нач}} \gg r$ (r — экспертная оценка, вообще говоря, неизвестного числа различных режимов работы исследуемого объекта). Единственное ограничение на $r_{\text{нач}}$ — это возможность построения статистически значимой оценки локальной линейной регрессии $\tilde{F}_j(x, \alpha_j)$ для большинства областей. Области, для которых это невозможно сделать, объединяются, исходя из ранее введённой меры близости (3) между группами точек A_i и A_j, соответствующих областям B_i и B_j. Такое объединение продолжается до тех пор, пока в каждой области не будет построена статистически значимая оценка локальной регрессии $\tilde{F}_j(x, \alpha_j)$. На втором этапе полученные области объединяются с помощью следующего алгоритма. На каждом шаге ищутся ближайшие в смысле меры близости (3) области B_i и B_j, затем рассматривается гипотеза: «аппроксимации локальных регрессий $\tilde{F}_i(x, \alpha_i)$ и $\tilde{F}_j(x, \alpha_j)$ статистически не различимы (эквивалентны)». Если гипотеза подтверждается, то области B_i и B_j объединяются, и для объединённой области B_{ij} строится аппроксимация локальной регрессии $\tilde{F}_{ij}(x, \alpha_{ij})$, в противном случае рассматривается следующая пара ближайших областей и т. д. Для проверки этой гипотезы в работе [23] используется статистика G. Chou, для которой необходимо знать аппроксимации локальных регрессий $\tilde{F}_i(x, \alpha_i)$, $\tilde{F}_j(x, \alpha_j)$ и $\tilde{F}_{ij}(x, \alpha_{ij})$, а также выбрать уровень значимости F_0. Важная особенность описанного алгоритма — автоматическое определение числа r областей B_j, причём в определённом смысле оптимальным образом.

4. МЕТОДЫ СТРУКТУРНОГО ПРОГНОЗИРОВАНИЯ

Многие крупномасштабные системы управления, в первую очередь — организационно-административные, функционируют в условиях большой информационной размытости и неопределённости. Именно поэтому в последнее время для исследования таких систем стали широко применяться не только методы структурного анализа данных, но и *методы структурного прогнозирования*, основу которых составляют процедуры клас-

сификационного анализа. Основная идея методов структурного прогнозирования состоит в том, что исследуются не точные значения параметров, описывающих состояние каждого объекта (например, траектории состояний), а лишь класс, к которому принадлежит каждый объект в рамках некоторой структуры (классификации) множества объектов, входящих в исследуемую систему [24]. Такое интегральное описание объектов позволяет существенно повысить эффективность анализа поведения системы, а также устойчивость и робастность процедур принятия управленческих решений и прогнозов.

Опишем вкратце общую схему работы одного из алгоритмов структурного прогнозирования [25]. Пусть исследуемая система состоит из n объектов, каждый из которых характеризуется набором из k параметров, измеряемых в дискретные моменты времени. В k-мерном пространстве параметров X j-ый объект в момент времени t представляется точкой $x_j(t) = \{x_j^{(1)}(t), x_j^{(2)}(t), ..., x_j^{(k)}(t)\}$. Упорядоченная совокупность точек $x_j(t_1), ..., x_j(t_m)$ является известной частью траектории, характеризующей динамику j-го объекта. Как уже говорилось, для многих прикладных задач для j-го объекта требуется прогнозировать не точные значения параметров-характеристик $x_j(t)$ в момент времени t_{m+1}, а лишь класс, к которому будет принадлежать объект в этот момент времени в рамках некоторой структуры (классификации) множества объектов изучаемой системы. Таким образом, основу предложенного алгоритма составляет процедура выявления структуры объектов, входящих в исследуемую систему. Для этой цели в работе [25] применяется комплексный алгоритм автоматической классификации, специально разработанный для решения таких задач [4]. С его помощью в момент времени t_1 производится структуризация n точек в пространстве X на r классов, каждый из которых и характеризует определённый тип объекта. Число классов r выбирается с помощью человеко-машинной процедуры, входящей в комплексный алгоритм [4]. Вводится понятие модели (эталона) класса $a_i(t)$, $i = 1, ..., r$, (обычно это центр класса) [10]. Для каждого объекта вычисляются расстояния до эталонов $R_{ij}(t)$, $i = 1, ..., r$, $j = 1, ..., n$.

В момент времени t_2 каждая точка $x_j(t_2)$ с помощью одного из алгоритмов распознавания образов с учителем относится к тому или иному классу в рамках классификации, полученной на первом шаге. В работе [25] для этого применяется алгоритм метода потенциальных функций, который в спрямляющем пространстве эквивалентен алгоритму ближайшего среднего [26]. После этого производится пересчёт эталонов $a_i(t_2)$, $i = 1, ..., r$, а также пересчёт (для точек $x_j(t_1)$) или подсчёт (для точек $x_j(t_2)$) расстояний $R(x_j(t_2), a_i(t_2))$ до новых эталонов $i = 1, ..., r$, $j = 1, ..., n$. Такая процедура выполняется для всех m моментов времени. В итоге для каждого объекта получается последовательность (траектория) из m позиций. В каждой позиции находится $r + 1$ число, первое из которых — это номер класса, к которому относился этот объект в соответствующий момент времени, а последующие числа — это значения расстояний до центров классов в тот же момент времени. Требуется спрогнозировать номер класса (тип объекта), к которому будет относиться каждый объект в момент времени t_{m+1}.

В качестве прогнозной модели для каждого объекта используется марковская цепь с r состояниями, т. е. на каждом шаге рассчитываются элементы матрицы переходных вероятностей $P = \|p_{ji}\|$, $j = 1, ..., n$, $i = 1, ..., r$. В работе [25] разработан специальный алгоритм пересчёта на каждом шаге соответствующих переходных вероятностей p_{ji} с использованием информации о значениях расстояний до центров классов и условия нормировки $\sum_{i=1}^{r} p_{ji} = 1$ для всех $j = 1, ..., n$. Построенная матрица переходных вероятностей используется для прогнозирования принадлежности объекта тому или иному классу. На практике обычно применяется не рандомизированная, а байесовская схема, когда объект относится к тому классу i_0, для которого $p_{ji_0} = \max_{i=1,...,r} p_{ji}$. Возможны различные модификации описанной выше схемы [25]: классификация объектов задаётся заранее (например, экспертным путём) и в последующем остаётся неизменной; используются данные только об s прошлых состояниях множества объектов (алгоритм с «памятью», s — глубина памяти); для структуризации применяются алгоритмы размытой классификации, в том числе с фоновым классом [10].

5. ЭКСПЕРТНЫЕ МЕТОДЫ В ЗАДАЧАХ СТРУКТУРНОГО АНАЛИЗА

Экспертные методы применялись в задачах структурного анализа достаточно давно, в большинстве своём при выборе свободных параметров алгоритмов структуризации (см., например, обзор [2]). Затем появились специальные корректирующие экспертные процедуры в алгоритмах выбора «оптимального» числа классов, выбора информативных параметров, заполнения пропущенных наблюдений [4], построения хорошо интерпретируемых классификаций [27]. Но наибольшее распространение экспертные методы получили при

решении задач исследования слабо формализованных социально-экономических и организационных систем управления [28].

Наиболее востребованными оказалась *коллективная бесконфликтная многовариантная экспертиза*, впервые предложенная в работе [29], концепция которой базируется на следующих основных принципах [28]:

— экспертиза проводится в экспертных комиссиях, число которых не меньше числа различных точек зрения на исследуемую проблему;

— в одну и ту же комиссию должны включаться эксперты, имеющие близкие точки зрения на исследуемую проблему;

— в каждой комиссии могут работать только эксперты, не имеющие конфликтных взаимоотношений;

— для коллективной экспертизы отбираются условно компетентные эксперты (те, которые считаются компетентными для экспертов из одной и той же комиссии);

— организация и проведение экспертизы, обработка экспертных оценок, формирование результатов экспертизы должны проводиться специальной консалтинговой группой, приглашённой, для большей объективности, со стороны, независимой и незаинтересованной в результатах экспертизы.

Концепция была реализована в рамках специальной методики формирования экспертных комиссий [28]. Методика состоит из пяти основных разделов (этапов): выявление кандидатов для работы в экспертных комиссиях; выявление существенно различных точек зрения; определение групп неконфликтующих экспертов; оценка условной компетентности экспертов; формирование экспертных комиссий. Были разработаны также варианты *структурной, структурно-иерархической и заочной многовариантной экспертизы*.

ЗАКЛЮЧЕНИЕ

Разработанные алгоритмы структурно-классификационного анализа сложноорганизованных данных широко применяются для решения разнообразных прикладных задач. Достаточно обширный набор примеров решения таких задач содержится в работах [2, 4, 7, 18, 22, 23, 25, 26, 28, 30]. К ним надо добавить работы по медицинской диагностике и анализу медицинской информации с применением структурно-классификационных алгоритмов, например, [31, 32]; решению естественнонаучных задач, например, [33]; проектированию профессиональных и образовательных стандартов, например, [34]; созданию процедур оценки эффективности функционирования крупномасштабных организационно-административных систем, например, [35].

В ближайшей перспективе центр тяжести исследований в этой области будет перемещаться в сторону создания оптимальных и квазиоптимальных алгоритмов структуризации большой и сверхбольшой размерности для разнородных параметров, при существенных пропусках в исходной информации, значимом уровнем помех (в том числе целенаправленного свойства), которые значительно больше будут ориентированы на экспертную информацию, в том числе: при выборе общей стратегии и конкретного набора процедур обработки, для выбора свободных (настраиваемых) параметров, содержательной интерпретации и коррекции получаемых результатов. В прикладных работах основное внимание будет уделяться крупномасштабным, слабо формализованным объектам (финансовые, социально-экономические и организационно-административные системы) и комплексам взаимосвязанных технологических процессов (в машиностроении; приборостроении, в том числе медицинском; химии, нефтехимии и нефтепереработке; в производстве современного вооружения и др.).

ЛИТЕРАТУРА

1. *Дорофеюк А.А., Мучник И.Б.* Работа М.А. Айзермана в области распознавания образов и анализа данных / В кн.: «Марк Аронович Айзерман 1913 — 1992». — М.: Физматлит, 2002. — С. 115—159.
2. *Дорофеюк А.А.* Алгоритмы автоматической классификации // Автоматика и телемеханика. — 1971. — № 12.
3. *Дорофеюк А.А.* Алгоритмы обучения машины распознаванию образов без учителя, основанные на методе потенциальных функций // Автоматика и телемеханика. — 1966. — № 10.
4. *Дорофеюк Ю.А.* Комплексный алгоритм автоматической классификации и его использование в задачах анализа и принятия решений // Таврический вестник информатики и математики. — 2008. — № 1. — С. 171—177.
5. *Браверман Э.М.* Метод потенциальных функций в задаче обучения машины распознаванию образов без учителя // Автоматика и телемеханика. — 1966. — № 10.
6. *Айзерман М.А., Браверман Э.М., Розоноэр Л.И.* Метод потенциальных функций в теории обучения машин. — М.: Наука, 1970.
7. *Дорофеюк А.А.* Алгоритмы автоматической классификации, основанные на методе потенциальных функций, и их практическое использование // Вопросы технической кибернетики. — М.: Наука, 1968.
8. *Бауман Е.В., Дорофеюк А.А.* Рекуррентные алгоритмы автоматической классификации // Автоматика и телемеханика. — 1982. — № 3.
9. *Бауман Е.В., Дорофеюк А.А.* Вариационный подход к задаче автоматической классификации для одного класса аддитивных функционалов // Автоматика и телемеханика. — 1978. — № 8.
10. *Бауман Е.В., Дорофеюк А.А.* Классификационный анализ данных // Избранные труды Междунар. конф. по проблемам управления. — М.: СИНТЕГ, 1999. — Т. 1.
11. *Браверман, Э.М.* Методы экстремальной группировки параметров и задача выявления существенных факторов // Автоматика и телемеханика. — 1970. — № 1.

12. *Программно-алгоритмический* комплекс структурно-классификационного анализа сложноорганизованных данных / Е.В. Бауман, А.А. Дорофеюк, Ю.А. Дорофеюк, Н.Е. Киселёва // Таврический вестник информатики и математики. — 2008. — № 1. — С. 66—72.
13. *Дорофеюк А.А., Гольдовская М.Д., Покровская И.В.* Когнитивные методы структурного анализа в задаче оценки эффективности слабо формализованных региональных систем // Когнитивный анализ и управление развитием ситуаций / Тр. VII Междунар. конф. — М.: ИПУ, 2007. — С. 33—36.
14. *Бауман Е.В., Москаленко Н.Е.* Структуризация результатов размытого кластер-анализа // Искусственный интеллект. — 2004. — № 2. — С. 355—359.
15. *Бауман Е.В., Москаленко Н.Е.* Методы экстремальной группировки параметров долевого типа // Автоматика и телемеханика. — 2008. — № 11. — С. 133—142.
16. *Дорофеюк А.А., Торговицкий И.Ш.* Применение методов автоматической классификации данных в задаче контроля качества изделий // Стандарты и качество. — 1967. — № 4.
17. *Дорофеюк А.А., Касавин А.Д., Торговицкий И.Ш.* Применение методов автоматической классификации для построения статической модели объекта / Автоматика и телемеханика. — 1970. — № 2.
18. *Райбман Н.С., Дорофеюк А.А., Касавин А.Д.* Идентификация технологических объектов методами кусочной аппроксимации. — М.: Институт проблем управления, 1977. — 70 с.
19. *Бауман Е.В.* Сведение задачи кусочно-линейной аппроксимации к задаче автоматической классификации // Моделирование и оптимизация сложных систем управления. — М.: Наука, 1981.
20. *Бауман Е.В., Дорофеюк А.А., Корнилов Г.В.* Алгоритмы оптимальной кусочно-линейной аппроксимации сложных зависимостей // Автоматика и телемеханика. — 2004. — № 10. — С. 163—171.
21. *Dorofeyuk A., Kasavin A.* Hierarchical piecewise approximation method in identification of complex plants // Identification and System Parameter Estimation. Part 3. — Amsterdam: North-Holland PC. 1978. — P. 1727—1736.
22. *Алиев С.А., Дорофеюк А.А., Мовсумов В.Г.* Методы комбинированной кусочной аппроксимации и их приложения // Анализ данных и экспертные оценки в организационных системах. — М.: ИПУ, 1985. — С. 45—50.
23. *Дорофеюк А.А. Ибрагимли Ш.Д., Мовсумов В.Г.* Использование критерия статистической эквивалентности моделей в задаче кусочной аппроксимации // Автоматика и телемеханика. — 1976. — № 7. — С. 109—113.
24. *Дорофеюк А.А., Дорофеюк Ю.А.* Методы структурно-классификационного прогнозирования многомерных динамических объектов // Искусственный интеллект. — 2006. — № 2. — С. 138—141.
25. *Дорофеюк Ю.А.* Структурно-классификационные методы анализа и прогнозирования в крупномасштабных системах управления // Проблемы управления. — 2008. — № 4. — С. 78—83.
26. *Браверман Э.М., Мучник И.Б.* Структурные методы обработки эмпирических данных. — М.: Наука, 1983.
27. *Дорофеюк А.А., Чернявский А.Л.* Алгоритмы построения хорошо интерпретируемых классификаций // Проблемы управления. — 2007. — № 2. — С. 83—84.
28. *Дорофеюк А.А., Покровская И.В., Чернявский А.Л.* Экспертные методы анализа и совершенствования систем управления // Автоматика и телемеханика. — 2004. — № 10. — С. 172—188.
29. *Дорофеюк А.А.* Методы автоматической классификации в задачах получения экспертной информации // Статистика. Вероятность. Экономика. Учёные записки по статистике. — М: Наука, 1985. — Т. 49. — С. 137—145.
30. *Дорофеюк А.А., Покровская И.В., Шипилов Ю.В.* Процедуры структурно-классификационной экспертизы и их практическое использование // Третья Междунар. конф. по проблемам управления. Пленарные доклады и избранные труды. — М.: ИПУ, 2006. — С. 372—375.
31. *Классификационный* анализ характеристик пульсового сигнала в задачах диагностики сердечно-сосудистых заболеваний / А.А. Дорофеюк, В.В. Гучук, А.А. Десова и др. // Таврический вестник информатики и математики. — 2008. — № 1. — С. 152—158.
32. *Дорофеюк А.А., Дмитриев А.Г.* Методы кусочной аппроксимации многомерных кривых // Автоматика и телемеханика. — 1984. — № 12.
33. *Браверман Э.М., Дорофеюк А.А., Лумельский В.Я.* Применение методов распознавания образов без учителя в естественнонаучных исследованиях // Адаптивные системы. Распознавание образов. Тр. Междунар. симпозиума ИФАК по техническим и биологическим аспектам управления, Ереван, 1968. — М.: Наука, 1971.
34. *Классификация* объектов профессиональной деятельности специалиста при проектировании профессиональных и образовательных стандартов / В.В. Никитин, С.В. Мальцева, А.А. Дорофеюк и др. // Проблемы управления. — 2007. — № 4. — С. 51—55.
35. *Лифшиц Д.В., Дорофеюк Ю.А.*. Методология оценки эффективности управления жилищно-коммунальным хозяйством крупного города на базе экспертно-классификационных методов анализа и моделирования ситуаций // Управление развитием крупномасштабных систем (MLSD'2008). Материалы второй междунар. конф. — М.: ИПУ РАН, 2008. — Т. I. — С. 63—66.

Дорофеюк Александр Александрович — д-р техн. наук, профессор, зав. лабораторией обработки больших массивов информации в иерархических системах. Председатель секции «Управление социально-экономическими, медико-биологическими и организационными структурами» Учёного совета ИПУ, член Научного совета РАН по теории управляемых процессов и автоматизации, член НТС Минтранса МО, член Экспертного совета Фонда «Социальное развитие» при Правительстве РФ, член Экспертного совета РФФИ, член редколлегии журнала «Проблемы управления». Опубликовал более 200 научных трудов, в том числе 14 монографий. Под его руководством защищено 16 кандидатских и 3 докторские диссертации. Основные научные интересы — структурно-классификационный анализ сложноорганизованных данных; коллективная многовариантная экспертиза; системные методы поддержки принятия решений в слабо формализованных системах управления; методы анализа, совершенствования и прогнозирования в социально-экономических и организационных системах управления. ☎(495) 334-75-40, ✉ adorof@ipu.ru.