УДК 62-50

МЕТОДЫ СТРУКТУРНОГО АНАЛИЗА В ПРИКЛАДНЫХ ЗАДАЧАХ ИССЛЕДОВАНИЯ ВРЕМЕННЫХ РЯДОВ[1]

М.Д. Гольдовская, Ю.А. Дорофеюк, Н.Е. Киселева

Описаны методы и алгоритмы структурного анализа временных рядов, базирующиеся на методологии классификационного анализа данных. Рассмотрен специальный случай одномерных временных рядов, наиболее часто встречающийся при решении практических задач. Для этого случая разработаны алгоритмы глобальной оптимизации соответствующих критериев качества структуризации. Предложенные методы реализованы при решении ряда прикладных задач исследования временных рядов.

Ключевые слова: структурный анализ данных, алгоритмы анализа временных рядов, одномерный временной ряд.

ВВЕДЕНИЕ

Для многих технических, социально-экономических и медико-биологических объектов управления исходные для анализа данные задаются в виде значений параметров, изменяющихся во времени, т. е. временных рядов. Именно поэтому создание эффективных методов анализа временных рядов является важной и актуальной задачей. Довольно часто при решении этой задачи необходимо не исследование точных значений рассматриваемых характеристик, а выделение некоторой структуры из имеющегося массива данных. В некоторых задачах временные ряды не поддаются классическому статистическому анализу ввиду своей сложности, пропусках в данных и др. В таких случаях для их исследования предлагается применять методы классификационного анализа данных.

В настоящей работе на базе методологии классификационного анализа данных [1, 2] разработаны методы и алгоритмы структурного анализа временных рядов, которые носят универсальный характер и могут использоваться при исследовании, идентификации, диагностике и совершенствовании методов принятия решений для широкого класса социально-экономических, организационно-административных, инженерно-технических и медико-биологических объектов. Разработанные алгоритмы были реализованы при решении ряда прикладных задач.

[1] Работа выполнена при частичной финансовой поддержке РФФИ, проекты 13-07-00992-а, 11-07-00178-а, 11-07-13137-офи-м-РЖД.

1. АНАЛИЗ ВРЕМЕННЫХ РЯДОВ

В настоящей работе рассматриваются два типа временных рядов. Первый тип — временные ряды, являющиеся естественным обобщением объектов классификационного анализа на случай классификации траекторий. Второй тип — одномерный временной ряд значений некоторого параметра исследуемого объекта в дискретные моменты времени. Для этого типа объектов задача структуризации сводится к классификационному анализу самих значений временного ряда.

1.1. Анализ временных рядов первого типа

Рассмотрим общий случай многомерных рядов первого типа. Формальная постановка задачи классификационного анализа подразумевает определение: множества объектов, подлежащих структуризации (классификации); множества решающих правил и критерия качества структуризации [1].

Постановка задачи. Пусть в момент времени t каждый объект $x_j(t)$ из исследуемого набора n объ-

ектов описывается набором значений k параметров $\{x_j^{(l)}(t),\ l = 1,\ ...,\ k\}$, т. е. является точкой в k-мерном пространстве X. Многомерный временной ряд $\tilde{x}_j = (x_j(1),\ ...,\ x_j(m))$ характеризует динамику (траекторию в X) состояния j-го объекта для m моментов времени. Тогда *множеством объектов, подлежащих структуризации,* является множество n таких временных рядов (динамических объектов) длины m, т. е. множество $X_n = \{\tilde{x}_1,\ ...,\ \tilde{x}_n\}$.

Для структуризации множества X_n в работе применяются методы размытой автоматической классификации (кластерного анализа) на r классов с фоновым классом [1], в котором *множество решающих правил* $H(X_n)$ — это n вектор-функций размерности $(r + 1)$:

$$H(X_n) = \{H(\tilde{x}_j),\ j = 1,\ ...,\ n\},$$

$$H(\tilde{x}_j) = (h_0(\tilde{x}_j),\ h_1(\tilde{x}_j),\ ...,\ h_r(\tilde{x}_j)), \qquad (1)$$

где $h_i(\tilde{x}_j)$ — функция принадлежности $\tilde{x}_j$ к i-му классу, а $h_0(\tilde{x}_j)$ — функция принадлежности $\tilde{x}_j$ к фоновому классу [1]. Для любого $\tilde{x}_j$ вектор-функция $H(\tilde{x}_j)$ должна принадлежать некоторому ограниченному замкнутому множеству V в $(r + 1)$-мерном евклидовом пространстве, т. е. $H(\tilde{x}_j) \in V \subseteq R^{r+1}$. Множество V определяет тип размытости для так поставленной задачи автоматической классификации.

В рамках общего вариационного подхода *критерий качества классификации* выбирается в соответствии с методом обобщенного среднего [2], а именно так, чтобы траектории объектов из одного класса хорошо описывались моделью (эталоном) этого класса. *Обобщенным средним* или эталоном множества, заданного функцией принадлежности $h(\tilde{x})$, называется модель $\tilde{\alpha}_h = \arg\max_{\tilde{\alpha} \in \Lambda} K(h(\tilde{x}),\ \tilde{\alpha})$. Введем в рассмотрение множество Λ возможных эталонов классов. Между элементами множества X_n и элементами $\tilde{\alpha}_i \in \Lambda$ вводится мера близости $K(\tilde{x}_j,\ \tilde{\alpha}_i)$. Величина $K(X_n,\ H(\tilde{x}_j),\ \tilde{\alpha}_i) = \sum_{j=1}^{n} K(\tilde{x}_j,\ \tilde{\alpha}_i)h_i(\tilde{x}_j)$ отражает меру того, насколько хорошо эталон $\tilde{\alpha}_i = \arg\max_{\tilde{\alpha} \in \Lambda} \sum_{j=1}^{n} K(\tilde{x}_j,\ \tilde{\alpha}_l)h_l(\tilde{x}_j)$ описывает множество точек i-го класса, заданное функциями принадлежности $h_i(\tilde{x}_j)$. В работе используется следующий *критерий качества классификации* траекторий:

$$J(H) = \sum_{i=1}^{r} \sum_{j=1}^{n} K(\tilde{x}_j,\ \tilde{\alpha}_i)h_i(\tilde{x}_j) + B \sum_{j=1}^{n} h_0(\tilde{x}_j), \quad (2)$$

где $\tilde{\alpha}_i$ — эталон i-го класса, а B — «вес» фонового класса. Тогда задача структурного анализа множества многомерных временных рядов X_n состоит в максимизации функционала (2) по классификациям $H(X_n)$ с учетом вида эталонов классов $\tilde{\alpha}_i$. Для выявления конкретного вида классификации, максимизирующей функционал (2), исследован вид его субдифференциала в данной точке (классификации) $H(X_n)$ [3].

Центральным для вариационного подхода в рамках поставленной задачи является понятие *эталонной классификации* [1]. Рассмотрим некоторый вектор моделей $A = (\tilde{\alpha}_1,\ ...,\ \tilde{\alpha}_r) \in \Lambda$. Назовем классификацию $H_A(\tilde{x}) = (h_0(\tilde{x}),\ h_1(\tilde{x}),\ ...,\ h_\gamma(\tilde{x}))$ *эталонной* с вектором A, если она удовлетворяет соотношению $H_A(\tilde{x}) = \arg\max_{(h_0,\ h_1,\ ...,\ h_r) \in V} \left[\sum_{i=1}^{r} K(\tilde{x},\ \tilde{\alpha}_i)h_i + Bh_0 \right]$.

Была доказана следующая теорема о виде оптимальной классификации временных рядов [3].

Теорема 1. *Если классификация H^* доставляет максимум функционалу* (2), *то тот же максимум достигается на некоторой эталонной классификации $H_A(\tilde{x})$ с вектором эталонов $A = (\tilde{\alpha}_1,\ ...,\ \tilde{\alpha}_r) \in \Lambda$, компоненты которого $\tilde{\alpha}_i$ являются обобщенными средними классов для классификации $H_A(\tilde{x})$.* ♦

Алгоритм структурного анализа многомерных рядов описывается следующей итерационной процедурой. Задается начальная классификация $H_0(X_n)$ (для выбора начальной классификации разработаны специальные алгоритмы [4]). На l-м шаге для классификации $H_l(X_n)$ в каждом нефоновом классе находится его эталон $\tilde{\alpha}_i^l$, по вектору эталонов $A^l = (\tilde{\alpha}_1^l,\ ...,\ \tilde{\alpha}_r^l)$ строится $(l + 1)$-е приближение оптимальной классификации $H_{l+1} = H_{A^l}$, где H_{A^l} — эталонная классификация с вектором эталонов A^l, и т. д. Доказана сходимость алгоритма к локальному экстремуму $J(H)$ [1]. Алгоритм конкретизируется как по виду критерия (2), так и по типам размытости.

Рассмотрим случай, когда *множество эталонов классов является множеством возможных траекторий исследуемого объекта*. Так же, как и при обычной классификации, наиболее простым способом введения множества эталонов классов служит множество всех возможных траекторий объектов, т. е. $\Lambda_1 = R^{n\times k}$. Другими словами, эталоны классов представляют собой в пространстве параметров траектории такой же длины, что и траектории объектов. Для множества эталонов Λ_1 рассматриваются следующие типы меры близости.

Мера близости типа l_1. Эта мера близости совпадает с евклидовым расстоянием в пространстве $R^{n\times k}$: $K_1(\tilde{x},\ \tilde{\alpha}) = -\sum_{i=1}^{n}\sum_{j=1}^{k}(x_i^{(j)} - \alpha_i^{(j)})^2$. Знак «минус» поставлен для удобства интерпретации, так как максимизация меры близости эквивалентна минимизации соответствующего расстояния. Использование такой меры приводит к тому, что траектории объектов классифицируются как точки евклидова пространства $R^{n\times k}$ с критерием средневзвешенного квадратичного отклонения. В результате обобщенные средние классов будут совпадать с центрами классов в этом пространстве. Следовательно, в окончательной классификации эталон каждого класса является средним траекторий объектов в соответствующем классе.

Мера близости типа l_1. Эта мера близости совпадает с расстоянием суммы модулей в пространстве $R^{n\times k}$: $K_2(\tilde{x},\ \tilde{\alpha}) = -\sum_{i=1}^{n}\sum_{j=1}^{k}|x_i^{(j)} - \alpha_i^{(j)}|$. Использование данной меры приводит к тому, что обобщенные средние будут совпадать с медианами классов в пространстве $R^{n\times k}$.

Мера близости с учетом приращения параметров. При анализе динамики часто нужно разбить объекты на группы схожих не по абсолютным значениям параметров, а по их приращениям. Для этого в меру близости добавляется расстояние между приращениями параметров:

$$K_3(\tilde{x},\ \tilde{\alpha}) = -D_0\sum_{i=1}^{n}\sum_{j=1}^{k}(x_i^{(j)} - \alpha_i^{(j)})^2 -$$
$$- D_1\sum_{i=2}^{n}\sum_{j=1}^{k}[(x_i^{(j)} - x_{i-1}^{(j)}) - (\alpha_i^{(j)} - \alpha_{i-1}^{(j)}]^2,$$

где D_0 и D_1 — некоторые весовые коэффициенты, выбираемые экспертным путем.

Эталоны классов — представители классов. Другая возможность ввода множества эталонов заключается в использовании самого исходного множества объектов, т. е. $\Lambda_2 = X_n$, а в качестве меры близости — одной из мер $K_1(\tilde{x},\ \tilde{\alpha})$, $K_2(\tilde{x},\ \tilde{\alpha})$ или $K_3(\tilde{x},\ \tilde{\alpha})$. В данном случае эталонная траектория каждого из классов будет совпадать с траекторией одного из объектов.

Качественное описание эталонов классов. Во многих случаях классы необходимо описывать не количественно, а качественно. Например, можно описать один из классов в следующих терминах: по первому параметру траектории объектов класса вначале принимают в основном низкие значения, затем увеличиваются, а в конце опять уменьшаются; по второму параметру траектории объектов класса все время принимают высокие значения. Таких качественных описаний может быть достаточно много.

Качественное описание эталонов классов с учетом приращений. Более сложный вариант качественного описания моделей траекторий классов дается в терминах приращений. В подобных случаях диапазон значений приращения каждого параметра $x^{(j)}$ разбивается на q градаций. Иначе говоря, на q градаций разбивается диапазон значений параметра $y^{(j)}$, который в i-й момент времени принимает значение $y_i^{(j)} = x_i^{(j)} - x_{i-1}^{(j)}$. Для этого случая мера близости $K_3(\tilde{x},\ \tilde{\alpha})$ имеет вид:

$$K_4(\tilde{x},\ \tilde{\alpha}) = -D_0\sum_{i=1}^{n}\sum_{j=1}^{k}(x_i^{(j)} - \alpha_i^{(j)})^2 -$$
$$- D_1\sum_{i=2}^{n}\sum_{j=1}^{k}[(x_i^{(j)} - x_{i-1}^{(j)}) - \beta_i^{(j)}]^2.$$

Отметим также *схемы порождения данных* в задачах структурного анализа временных рядов. Исходная информация о функционировании системы динамических объектов представляет собой трехмерную таблицу (куб данных) «объект — параметр — время» [5]. Если число моментов времени в кубе данных достаточно велико, то можно предположить, что временной ряд каждого объекта состоит из коротких отрезков стандартного вида, например, соответствующих становлению, росту, реорганизации объекта или сезонным колебаниям в его работе и т. д. Задача состоит в нахождении полного набора таких стандартных отрезков временного ряда, достаточных для построения содержательно адекватного описания функционирования объекта в этих терминах. Для этого строится таблица значений «параметр — момент времени»,

в качестве объектов классификации рассматриваются все возможные отрезки временного ряда длины m, получаемые из него с помощью «окошка» ширины m, которое сдвигается вдоль этого временного ряда. Для полученной в результате классификации строится содержательное описание классов (в основном путем анализа эталонов классов). Тогда для каждого объекта формируется последовательность номеров классов, к которым он принадлежит в разные моменты времени. Последовательность номеров преобразуется в последовательность содержательных описаний классов, что и позволяет описать динамику «жизненного цикла объекта» за весь наблюдаемый период времени в качественных терминах.

1.2. Анализ временных рядов второго типа

Одномерный случай классификационного анализа обладает уникальным свойством, существенно упрощающим процедуру целенаправленного перебора, используемую при автоматической классификации. 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). Для анализа временных рядов такого типа в настоящей статье используется одномерный вариант алгоритма m-локальной оптимизации [6]. Поскольку работа этого алгоритма является определяющей для получения эффективной классификации временных рядов, дадим краткое описание работы его детерминированного варианта.

Пусть задано начальное разбиение R_0 всех точек классифицируемой выборки $x_1, ..., x_n$ на r классов. Ввиду упорядоченности классов на оси единственного параметра, на каждом конкретном шаге алгоритма достаточно рассматривать только пару соседних классов. Для определенности будем обозначать через A_1 левый из этой пары классов, а через A_2 — правый. Алгоритм содержит m циклов, на s-м цикле ($s = 1, ..., m$) производится локальная оптимизация классификации, полученной на предыдущем цикле, с использованием процедуры «переброски» s точек из одного класса в другой для каждой пары соседних классов.

На первом цикле осуществляется «переброска» по одной точке. Здесь классификация, полученная на предыдущем цикле, — это начальная классификация R_0. Поясним эту процедуру для первого этапа этого цикла, когда рассматривается пара классов, расположенная в самой левой части диапазона значений x_j. Обозначим через A_1 и A_2 первый и второй классы начального разбиения R_0 соответственно (классы нумеруются слева направо). В классе A_1 находится точка $x_j^{1,1,1}$ (индексы сверху — номера цикла, этапа и класса соответственно), ближайшая к границе рассматриваемой пары классов. Обозначим через $\rho_0(x_j^{1,1,1})$ индекс этой точки (для аналогичной точки на s-м цикле это обозначение будет иметь вид $\rho_{s-1}(x_j^{s,1,1})$). По построению $\rho_0(x_j^{1,1,1}) = 1$. Затем «перебросим» эту точку в класс A_2 и подсчитаем ее индекс на первом цикле: $\rho_1(x_j^{1,1,1}) = \text{sign}[J(\rho_0(x_j^{1,1,1} \in A_1) - J(\rho_0(x_j^{1,1,1}) \in A_2)]$, — где $J(x_j^{1,1,1} \in A_1)$ — значение критерия качества классификации J, подсчитанное только для точек классов A_1 и A_2 при условии, что точка $x_j^{1,1,1}$ принадлежит классу A_1, аналогично определяется $J(x_j^{1,1,1} \in A_2)$. Точка $x_j^{1,1,1}$ остается в первом классе (т. е. $\rho_1(x_j^{1,1,1}) = \rho_0(x_j^{1,1,1}) = 1$), если $J(x_j^{1,1,1} \in A_1) \geqslant$ $\geqslant J(x_j^{1,1,1} \in A_2)$, и переходит во второй класс ($\rho_1(x_j^{1,1,1}) = -1$) в противном случае. Если точка $x_j^{1,1,1}$ перешла во второй класс, то аналогичная процедура проделывается с точкой $x_{j-1}^{1,1,1}$, которая является ближайшей к новой границе между классами A_1 и A_2 среди всех точек первого класса (в данном случае — это предыдущая точка классифицируемой последовательности). И так продолжается до тех пор, пока точка $x_{j-1}^{1,1,1}$ не останется в первом классе. Это означает, что на первом этапе первого цикла из первого класса во второй будут «переброшены» l ближайших к границе точек. Если точка $x_j^{1,1,1}$ осталась в первом классе, то аналогичная процедура проводится с точками второго класса, начиная с точки $x_j^{1,1,2}$, которая является ближайшей к границе рассматриваемой пары классов. После того, как закончится «перебрасывание» точек из второго класса в первый (если это будет иметь место) либо не произойдет «перебрасывания» точки $x_j^{1,1,2}$, выполняется переход на второй этап первого цикла.

На втором этапе вся последовательность процедур первого этапа повторяется, только через A_1

обозначаются точки, входящие во второй класс после завершения первого этапа первого цикла, а через A_2 — третий класс начального разбиения R_0. И так далее до тех пор, пока не будут пройдены все $(r - 1)$ этапов первого цикла.

На всех этапах s-го цикла описанные процедуры повторяются с точностью до числа «перебрасываемых» точек — «перебрасывается» не по одной, а по s точек, ближайших к границе текущей пары классов. Очевидно, что процедура не может применяться для классов A_i, число точек n_i в которых меньше, чем $(s + c)$. В настоящем алгоритме $c = 2$.

Это правило используется в алгоритме для автоматического выбора максимально возможной глубины перебора $m_{\max}$. А именно, значение m (глубина перебора) выбирается из условия: в классификации, полученной после $(m - 1)$-го цикла, должен быть хотя бы один класс, число точек в котором не меньше $(m + 2)$.

Завершение m-го цикла является окончанием первой итерации. На второй итерации повторяются все процедуры первой, только на первом цикле вместо начального разбиения R_0 используется результирующая классификация первой итерации.

Алгоритм прекращает работу, если в пределах одной итерации не произойдет ни одного «перебрасывания» точек из класса в класс. Была доказана следующая теорема [6].

Теорема 2. *Алгоритм одномерной m-локальной оптимизации для $m = m_{\max}$ сходится за конечное число шагов к глобальному максимуму критерия качества классификации.* ♦

В случае необходимости применяется экспертно-классификационный алгоритм выбора «оптимального» числа классов, входящий в комплексный алгоритм автоматической классификации [4].

2. ПРИМЕНЕНИЕ РАЗРАБОТАННЫХ МЕТОДОВ В ПРИКЛАДНЫХ ЗАДАЧАХ

Разработанные алгоритмы были применены для:

— исследования социально-экономического развития субъектов РФ;

— корректировки оценок показателей экономической активности по субъектам РФ в условиях малых выборок;

— структурно-классификационного анализа пульсового сигнала лучевой артерии в задачах медицинской диагностики.

Первые две из этих задач относятся к задачам анализа временных рядов первого типа, а последняя — к задачам анализа временных рядов второго типа.

2.1. Задача исследования социально-экономического развития субъектов РФ

Отбор и предобработка исходных данных. В качестве объектов исследования рассматривались короткие временные ряды значений 47 показателей социального развития для 79 регионов РФ за 3 года. Множество исходных показателей разбивалось на шесть тематических групп: доходы населения (13 показателей); расходы и сбережения (14 показателей); потребление продуктов питания (8 показателей); демографические характеристики (4 показателя); характеристики социальной напряженности (6 показателей); объем финансовой помощи из межрегиональных фондов (2 показателя). Для обеспечения сопоставимости данных за разные годы все стоимостные показатели были пересчитаны в сопоставимых ценах (проведено дисконтирование). Предобработка исходного материала включала в себя статистическую фильтрацию исходных параметров и заполнение пропущенных наблюдений [4].

Отбор основных показателей, характеризующих регионы. При помощи методики формирования информативных показателей (факторов) [4] из 47 исходных показателей социально-экономической ситуации в регионах для последующей классификации и формирования рейтинга регионов были отобраны шесть основных показателей [7]: среднедушевой доход, доля оплаты труда в среднедушевом доходе, превышение доходов над расходами, число пенсионеров на 1000 чел. населения, уровень безработицы, общий объем финансовой помощи на душу населения.

Классификация регионов. Для классификации траекторий регионов применялся комплексный алгоритм автоматической классификации [4]. Оказалось, что для отобранных информативных параметров траектории 79 объектов (за три года) были разбиты на семь классов. Для этой цели применялся экспертно-классификационный алгоритм выбора «оптимального» числа классов [4].

Построение линейно-упорядоченного рейтинга регионов. Исходными данными для составления рейтингов служат результаты одномерных автоматических классификаций регионов по каждому из шести информативных показателей. Классы занумерованы так, что «лучшие» по данному показателю регионы находятся в первом классе, «худшие» — в последнем. Таким образом, номер класса служит рейтингом объекта. Пользуясь результатами классификации, можно дать качественную характеристику изменений социально-экономической ситуации в регионах за рассматриваемый период.

При построении рейтингов по двум и более показателям одновременно возникает проблема многокритериальности: как упорядочить два объекта, один из которых имеет более высокий рейтинг по одному показателю, а второй — по другому показателю. В данном случае применялось следующее простое правило: при низких значениях первого параметра упорядочение производится по второму параметру; при низких значениях второго параметра упорядочение ведется по первому. Полученные результаты свидетельствуют о том, что примененная методология структурно-классификационного анализа позволила свести плохо обозримую совокупность большого числа исходных показателей к небольшому числу наиболее информативных, а затем, используя эти показатели, удалось разбить множество траекторий регионов на классы регионов, близких между собой по уровню социального развития. Исходная информация структурировалась как по множеству показателей, так и по множеству траекторий регионов и представляется в сжатом, обозримом виде, удобном для принятия управленческих решений.

Полученные результаты сыграли серьезную роль при анализе и оценке эффективности работы государственных органов власти (не только регионального, но и федерального уровня) по управлению социальным развитием регионов РФ.

2.2. Задача корректировки (сглаживания) оценок показателей экономической активности по субъектам РФ в условиях малых выборок

В настоящее время по вопросам экономической активности, занятости и безработицы в рамках специальной программы обследования населения, утвержденной Правительством РФ, ежемесячно опрашивается около 69 тыс. чел. (представляющих около 33 тыс. домашних хозяйств) в возрасте 15—72 года, или около 0,06 % населения данного возраста. Однако из-за недостаточного финансирования этой программы не было возможности обеспечить необходимую достоверность помесячных данных по этим показателям в разрезе большинства субъектов РФ. Объем месячной выборки обеспечивает представительные данные только в целом по РФ и некоторым крупным (по численности населения) регионам. Отметим, что рассматриваемые показатели в рамках этой программы определяются по данным только этого обследования. Так, например, для оценки уровня безработицы не привлекаются данные службы занятости или другие данные подразделений Росстата. А именно, уровень безработицы — это отношение численности безработных к численности экономически активного населения, полученные в результате выборочного обследования.

Простейшим методом сглаживания служит метод скользящего среднего. Экспериментальные расчеты показали, что выборка, построенная путем объединения выборок для трех последовательных месяцев, достаточно представительна, и построенная этим методом кривая уровня исследуемого показателя оказывается достаточно гладкой. Однако метод скользящего среднего имеет один существенный недостаток: чтобы рассчитать значение скользящего среднего за текущий месяц, необходимы данные выборочного обследования за следующий месяц. Задача состоит в разработке такого метода сглаживания, который был бы свободен от указанного недостатка.

Идея предлагаемого метода структурной группировки регионов состоит в том, что для повышения надежности оценки показателя (сглаживания) в одну выборку объединяются не выборки за разные месяцы, полученные в одном и том же регионе, а выборки, полученные в одном и том же месяце, но в нескольких регионах, близких по динамике исследуемого показателя [8]. Далее метод группировки регионов описан как метод оценки показателя s в i-м регионе в k-м месяце текущего года. Этот регион и этот месяц называются расчетными. Метод включает в себя четыре этапа.

Этап 1. Производится сглаживание помесячных данных обследования, для чего применяется процедура скользящего среднего.

Этап 2. При помощи методов автоматической классификации производится структуризация множества траекторий показателя s регионов РФ на два класса — эталонный и фоновый, далее для расчетов используется только эталонный класс. Для классификации применяется специально разработанный алгоритм формирования виртуального региона [8], в качестве меры близости в котором используется коэффициент корреляции между траекториями показателя s различных регионов. Таким образом, формируется группа регионов (эталонный класс), близких в заданном виде к расчетному региону по динамике показателя s; выборки вошедших в эту группу регионов объединяются. Полученная группа регионов рассматривается как один виртуальный регион, ассоциируемый с расчетным регионом.

Этап 3. На базе объединенной выборки виртуального региона с помощью процедуры масштабирования находится искомая оценка показателя s для расчетного региона по состоянию на расчетный месяц. Хотя получаемые в процессе работы

алгоритма временные ряды по форме могут почти не отличаться друг от друга, их средние значения и масштаб могут отличаться значительно. Это объясняется тем, что в качестве меры близости временных рядов при формировании виртуального объекта используется значение коэффициента корреляции. Для того чтобы устранить полученное в результате этого смещение и изменение масштаба и применяется процедура масштабирования, т. е. при помощи линейной регрессии производится линейное преобразование полученного временного ряда оценок.

Этап 4. Производится сезонное сглаживание (выделение линейного тренда и сезонной составляющей) временного ряда оценок показателя *s*, полученного на третьем этапе. В результате определяются «трендовое» значение и сезонная составляющая показателя *s* в расчетном регионе по состоянию на расчетный месяц.

Для проверки эффективности метода структурной группировки траекторий были рассчитаны оценки показателей безработицы, экономической активности и занятости в период с сентября 2010 г. по октябрь 2011 г. по всем регионам РФ и проведено сравнение этих оценок с оценками, полученными методом скользящего среднего. Результаты расчетов позволяют сделать следующие выводы [8].

- Оценки соответствующих показателей, полученные методом скользящего среднего и методом группировки регионов, очень близки (например, даже по регионам «проблемного» Северо-Кавказского федерального округа разница между полученными оценками составляет в среднем всего около 2 % от уровня оцениваемого параметра).
- Ошибки метода скользящего среднего — это ошибки интерполяции (среднее значение оцениваемого показателя за три последовательных месяца приписывается среднему месяцу). Ошибки метода структурной группировки объектов связаны с неоднородностью выборки (группируемые регионы хотя и близки по динамике оцениваемого показателя, но не идентичны). Тот факт, что при разных источниках ошибок результаты получаются достаточно близкими, говорит о том, что метод структурной группировки объектов может эффективно применяться для достоверной оценки уровня соответствующего параметра.
- Метод структурной группировки объектов обладает решающим преимуществом: он позволяет формировать оценки уровня анализируемого параметра сразу же после получения данных выборочного обследования.

2.3. Задача структурно-классификационного анализа пульсового сигнала лучевой артерии в задачах медицинской диагностики

Специфика задачи анализа пульсового сигнала в задачах медицинской диагностики состоит в том, что многие его характеристики одномерны. Именно поэтому в работе [9] использовалась одномерная модификация алгоритма *m*-локальной оптимизации для выделения как амплитудных, так и временных параметров основного квазипериода пульсового сигнала лучевой артерии.

Основные трудности выбора информативных параметров для описания квазипериодического пульсового сигнала связаны с надежным выделением основного (базового) квазипериода, поскольку его вариабельность даже для одного и того же сеанса записи весьма значительна из-за наличия большого числа влияющих на сигнал факторов [9]. Рассмотрим подробнее процедуру выделения основного квазипериода пульсового сигнала. На анализируемой записи сигнала выделяются все локальные максимумы, последовательность которых и является временным рядом второго типа. Затем с помощью одномерного варианта алгоритма *m*-локальной оптимизации строится автоматическая классификация этого ряда на r_{opt} классов. Для этого применялся экспертно-классификационный алгоритм выбора «оптимального» числа классов, входящий в комплексный алгоритм автоматической классификации [4]. Обычно при обработке реальных пульсограмм значение r_{opt} находилось в диапазоне 3—5. Самый правый на оси значений класс (большие значения амплитуды) в большинстве случаев соответствует максимумам основного квазипериода анализируемого сигнала. Далее на реализации сигнала выделяются максимумы, попавшие в крайний правый класс, тогда отрезки сигнала между смежными выделенными максимумами и служат претендентами на искомые квазипериоды. К сожалению, на реальных пульсограммах часто наблюдаются существенные колебания значений амплитуд. В связи с этим далее анализируется распределение выделенных максимумов на временно́й шкале, т. е. фактически проводится анализ динамического ряда этих максимумов. Если расстояние между соседними максимумами этого динамического ряда больше $T_c K_a$, где T_c — средняя длительность квазипериода, а K_a — коэффициент аритмии (выбирается экспертным путем в диапазоне 1,5—2,5), то для этой непериодизированной области выполняется коррекция. Процедура коррекции организована итеративным путем. А именно: в непериодизированной области

ищется абсолютный максимум амплитуды и его включают в исследуемый динамический ряд. Затем вновь анализируется распределение максимумов на временной шкале и т. д. Остается проблема, связанная с неопределенностью коэффициента аритмии K_a. Как показали тесты, влияние выбора его значения сказывается на качестве периодизации лишь в экзотических случаях. Предложенная процедура периодизации выделила практически все периоды на всей экспериментальной выборке пульсограмм. Исключения составили отдельные квазипериоды на нескольких пульсограммах, производящие весьма неоднозначное впечатление даже на специалиста-прикладника.

По аналогичной схеме находились и другие квазипериодические составляющие сигнала в рамках основного квазипериода.

Работа проводилась на базе обширного экспериментального материала, полученного в ходе исследования детей и подростков в клинике функциональной патологии НЦ здоровья детей РАМН. В исследовании принимали участие 417 чел. в возрасте от 9 до 16 лет. Все пациенты по основному диагнозу были разделены на два класса: первичная артериальная гипертензия и различные виды психосоматической функциональной патологии при нормальном артериальном давлении. Подчеркнем, что во второй класс входили не здоровые люди, а пациенты с другими, нередко достаточно серьезными заболеваниями, что существенно осложняло задачу диагностики артериальной гипертензии. Разработанное программно-алгоритмическое обеспечение позволило получить существенно более эффективные диагностические правила определения ранней гипертензии у детей, чем применяемые в настоящее время в медицинской практике.

ЗАКЛЮЧЕНИЕ

Предложена общая постановка задачи динамического структурного анализа временных рядов, когда каждый объект по каждому параметру характеризуется набором значений для некоторой последовательности моментов времени (траекторией). В рамках вариационного подхода разработаны соответствующие алгоритмы структурного анализа временных рядов. Проведен теоретический анализ этих алгоритмов, а также результаты их применения при решении ряда прикладных задач. Предложенные алгоритмы реализованы в составе компьютерного программно-алгоритмического комплекса, предназначенного для классификационного анализа сложно организованных данных при решении широкого класса прикладных задач [5].

ЛИТЕРАТУРА

1. *Бауман Е.В., Дорофеюк А.А.* Классификационный анализ данных // Тр. Междунар. конф. по проблемам управления / ИПУ РАН. — М., 1999. — Т. 1. — С. 62—67.
2. *Bezdek J.C.* Pattern recognition with fuzzy objective function algorithms. — N.-Y.: Plenum press, 1981. — 260 p.
3. *Дорофеюк А.А., Бауман Е.В., Покровская И.В.* Методы структуризации многомерных динамических объектов / Интеллектуализация обработки информации (ИОИ-2010): 8-я Междунар. конф., Пафос, Республика Кипр: Сб. докл. — М.: МАКС Пресс, 2010. — С. 125—128.
4. *Дорофеюк Ю.А.* Комплекс алгоритмов экспертно-классификационного анализа для решения прикладных задач / Четвертая междунар. конф. по проблемам управления (МКПУ-IV): Сб. тр. / ИПУ РАН. — М., 2009. — С. 373—379.
5. *Бауман Е.В., Дорофеюк А.А., Дорофеюк Ю.А., Киселева Н.Е.* Программно-алгоритмический комплекс структурно-классификационного анализа сложно организованных данных // Таврический вестник информатики и математики. — 2008. — № 1. — С. 66—72.
6. *Дорофеюк Ю.А.* Алгоритм *m*-локальной оптимизации в задачах структуризации // Управление развитием крупномасштабных систем (MLSD'2010): Тр. четвертой Междунар. конф. / ИПУ РАН. — М., 2010. — С. 248—256.
7. *Дорофеюк Ю.А., Гольдовская М.Д., Покровская И.В.* Методы структурно-экспертного анализа данных в задаче оценки эффективности функционирования региональных систем управления // Теория активных систем: Тр. Междунар. науч.-практ. конф. / ИПУ РАН. — М., 2010. — С. 139—142.
8. *Лайкам К.Э., Дорофеюк А.А., Дорофеюк Ю.А., Чернявский А.Л.* Классификационные методы коррекции результатов мониторинга социально-экономических показателей в условиях нерепрезентативных выборок // Вопросы статистики. — 2011. — № 5. — С. 13—18.
9. *Процедуры* классификационного анализа в задаче формирования информативных признаков при исследовании ритмической структуры биосигнала / А.А. Десова, А.А. Дорофеюк, В.В. Гучук и др. // Автоматика и телемеханика. — 2008. — № 6. — С. 143—152.

Статья представлена к публикации членом редколлегии А.С. Манделем.

Марина Дмитриевна Гольдовская — науч. сотрудник, ☎(495) 334-90-70, ✉ m_goldovskaya@mail.ru,

Юлия Александровна Дорофеюк — канд. техн. наук, науч. сотрудник, ☎ (495) 334-75-40, ✉ dorofeyuk_julia@mail.ru,

Нелли Евсеевна Киселева — науч. сотрудник, ☎(495) 334-90-70, ✉ nekisel-eva@mail.ru,

Институт проблем управления им. В.А. Трапезникова РАН, г. Москва.