
СИСТЕМНЫЙ АНАЛИЗ, УПРАВЛЕНИЕ И ОБРАБОТКА ИНФОРМАЦИИ

УДК 681.518

МЕТОД ПРОГНОЗИРОВАНИЯ ГРУПП ВРЕМЕННЫХ РЯДОВ С ПРИМЕНЕНИЕМ АЛГОРИТМОВ КЛАСТЕРНОГО АНАЛИЗА

Астахова Надежда Николаевна, аспирант, Рязанский государственный радиотехнический университет, 390005, Российская Федерация, г. Рязань, ул. Гагарина, 59/1, e-mail: asnadya@yandex.ru

Демидова Лилия Анатольевна, доктор технических наук, профессор, Рязанский государственный радиотехнический университет, 390005, Российская Федерация, г. Рязань, ул. Гагарина, 59/1, e-mail: liliya.demidova@rambler.ru

Предложен метод прогнозирования групп временных рядов (ВР) с применением алгоритмов кластерного анализа. Обосновано использование в качестве базового алгоритма кластеризации алгоритма четких *c*-средних. Введена метрика для оценки расстояния между ВР, позволяющая учесть в ходе кластеризации различную во времени актуальность элементов временных рядов. Предложено группировать ВР в кластеры (подгруппы), используя значения элементов этих рядов в качестве величин характеристик, на основе которых алгоритм кластеризации вырабатывает решение об отнесении ВР к тому или иному кластеру. Выполнено сопоставление координат центров кластеров и обобщающих ВР, т.е. центроидов кластеров. Реализовано описание ВР – центроидов кластеров с использованием моделей прогнозирования на основе строго бинарных деревьев и модифицированного алгоритма клонального отбора. Показана возможность формирования с применением вышеуказанных моделей прогнозирования аналитических зависимостей, наилучшим образом описывающих известные значения ВР и обеспечивающих получение минимальных значений средних относительных ошибок прогнозирования. Предложено использовать общую модель прогнозирования, построенную для ВР – центроида кластера, при прогнозировании частных (отдельных) ВР, входящих в кластер. Сделан вывод о возможности получения индивидуальных результатов прогнозирования для частных ВР за счет использования значений их элементов в качестве величин переменных в общих моделях прогнозирования. Продемонстрирована перспективность применения предлагаемого метода прогнозирования групп ВР.

Ключевые слова: временной ряд, кластеризация, центроид кластера, алгоритм четких *c*-средних, модель прогнозирования, средняя относительная ошибка прогнозирования, строго бинарное дерево, антитело, антиген, модифицированный алгоритм клонального отбора

FORECASTING METHOD FOR GROUPED TIME SERIES WITH THE USE OF ALGORITHMS FOR CLUSTER ANALYSIS

Astakhova Nadezhda N., post-graduate student, Ryazan State Radio Engineering University, 59/1 Gagarin Str., Ryazan, 390005, Russian Federation, e-mail: asnadya@yandex.ru

Demidova Liliya A., D.Sc. (Engineering), Professor, Ryazan State Radio Engineering University, 59/1 Gagarin Str., Ryazan, 390005, Russian Federation, e-mail: liliya.demidova@rambler.ru

The paper is focused on the forecasting method for time series groups with the use of algorithms for cluster analysis. K-means algorithm is suggested to be a basic one for clustering. In order to estimate the distance between time series a metric has been introduced. It allows to take into account the relevance of the time series elements in different time during the clustering. It is supposed to group time series into clusters (subgroups), using time series values of the elements as characteristic values. On the basis of such data the

clustering algorithm makes a decision to classify the time series to a particular cluster. The coordinates of the centers of clusters have been put in correspondence with summarizing time series data – the centroids of the clusters. A description of time series, the centroids of the clusters, is implemented with the use of predictive models. They are based on strict binary trees and a modified clonal selection algorithm. With the help of such predictive models, the possibility of forming analytic dependences is shown. The last-mentioned describes acquainted values of time series in the best way and provides minimum values of the average relative error of prediction. It is suggested to use a common prediction model, which is constructed for time series - the centroid of the cluster, in predicting the private (individual) time series in the cluster. The conclusion about the possibility of obtaining individual results of time series prediction is made by using the values of the elements of private time series as values of the variables in the general prediction models. The promising application of the suggested method for grouped time series forecasting is demonstrated.

Keywords: time series, clustering, cluster centroid, k-means algorithm, average forecasting error rate, strictly binary tree, antibody, antigen, modified clonal selection algorithm

Введение. В настоящее время управление производственными, социально-экономическими, медико-биологическими и многими другими процессами требует их объективного прогнозирования с учетом взаимосвязей различных факторов. Для этой цели могут применяться различные подходы, основанные, в том числе, на экспертных оценках и методах прогнозирования с использованием данных, представленных в виде взаимосвязанных временных рядов (ВР). Отдельным аспектам прогнозирования на основе взаимосвязанных ВР (ВВР) посвящен ряд работ отечественных и зарубежных авторов [1, 5–7, 26, 27]. Однако некоторые направления анализа/прогнозирования ВВР до сих пор остаются слабо разработанными. Одно из таких направлений связано с разработкой методов прогнозирования групп ВР с применением алгоритмов кластерного анализа, что позволит обеспечить получение прогноза с использованием относительно небольших объемов вычислений. Цель настоящей статьи – разработка метода прогнозирования, относящегося именно к этому направлению.

Общая характеристика проблематики работы. При разработке моделей прогнозирования следует учитывать не только тенденции прогнозируемого ВР, но также и тенденции других ВР, оказывающих на него прямое или косвенное влияние. Так, например, зависимыми являются такие описываемые с помощью ВР макропоказатели, характеризующие степень развития страны, как валовой внутренний продукт и уровень экспорта; средний уровень доходов и уровень образования населения; уровень медицинского обеспечения и средняя продолжительность жизни. Взаимосвязи между ВР могут быть «двунаправленными», при этом на один ВР могут оказывать влияние сразу несколько других ВР. При этом взаимосвязи обычно носят «статистический», а не строго детерминированный характер.

Многие ВР имеют сходные законы изменения значений своих элементов и могут быть объединены в подгруппы по критерию сходства этих законов. В этом случае целесообразно выполнение разработки модели прогнозирования, общей для всех ВР, входящих в подгруппу. Такая общая модель может быть использована непосредственно для прогнозирования всех частных ВР подгруппы. Кроме того, для отдельных частных ВР общая модель прогнозирования может быть уточнена (в случае недостаточной точности прогнозных значений, получаемых с ее применением).

Так как при решении задачи прогнозирования ВР именно процесс разработки модели характеризуется самой высокой вычислительной трудоемкостью, то использование одной общей модели прогнозирования для всех ВР, входящих в подгруппу, может быть значительным шагом в развитии подходов к анализу коррелирующих ВР. Разработка такого подхода к прогнозированию для групп ВР позволит выполнять прогнозирование частных ВР, входящих в подгруппы, с приемлемыми временными затратами.

Очевидно, что применение технологий кластерного анализа, а именно алгоритмов кластеризации, таких как алгоритм четких C -средних (k -means) [13, 16], алгоритм нечетких C -средних (fuzzy C -means – FCM) [13, 22, 24], ЕМ-алгоритм [8] позволит объединить сходные ВР в подгруппы (кластеры) и определить ВР-центроиды кластеров. В сущности, именно это и необходимо для решения задачи разработки общих моделей прогнозирования для подгрупп ВР.

Для разработки моделей прогнозирования ВР-центроидов кластеров авторами настоящей статьи предлагается использовать подход, основанный на применении строго бинарных деревьев (СБД) и модифицированного алгоритма клонального отбора (МАКО) [2, 11]. В этом случае удастся сформировать аналитические зависимости, наилучшим образом описывающие известные значения ВР и обеспечивающие получение минимальных значений аффинитета – средней относительной ошибки (СООП).

Теоретическое обоснование предлагаемого метода прогнозирования. Пусть имеется группа ВР T : t_i ($i=\overline{1,mF}$). Кроме того, каждый ВР t_i содержит n ($10 \leq n \leq 30$) элементов t_i^j ($j=\overline{1,n}$), измеренных в некоторые последовательные моменты времени (с постоянным шагом по времени). Далее будет по умолчанию полагаться, что относительные точности значений элементов во всех рядах одинаковы [6].

Существенный интерес представляет решение задачи разработки моделей прогнозирования всех ВР, входящих в группу T , с приемлемыми (для выполнения всех необходимых вычислений) временными затратами. Последнее требование приобретает особую актуальность, если необходимо непрерывно строить прогнозы в режиме реального времени (например, для оценки будущих состояний пациентов, находящихся в стадии наркоза и т.п.).

Временные ряды, как и другие объекты анализа данных, могут быть объединены в кластеры (подгруппы) с учетом значений тех или иных характеристик. Такими характеристиками могут быть, например, максимальное или минимальное значения элементов ВР, величины математического ожидания, дисперсии и т.п., вычисленные на основе значений элементов ВР [1, 21].

Обычно при разработке моделей прогнозирования ВР решается задача подбора некоторой аналитической зависимости, наилучшим образом описывающей закон изменения значений элементов ВР во времени. Поэтому представляется целесообразным использовать сходство математических законов изменения значений элементов ВР во времени для группирования ВР в кластеры.

Предположение о возможном наличии сходства математических законов изменения значений элементов ВР, а, следовательно, и о сходстве соответствующих видов аналитических зависимостей, может быть обосновано тем, что многие социально-экономические показатели, показатели здоровья населения и пр. взаимосвязаны между собой – поэтому изменение тенденций одного из них влечет изменение тенденций других [3, 4, 16].

К сожалению, нередко информация о том, какие ВР образуют относительно сильно связанные подгруппы в группе анализируемых ВР, отсутствует. Это приводит к необходимости привлечения дополнительных приемов анализа данных, в частности, алгоритмов кластерного анализа.

Использование таких алгоритмов, как алгоритм четких C -средних [13, 16] и алгоритм нечетких C -средних [13, 22, 24], ЕМ-алгоритм [8] позволяет в ходе итерационных вычислений выполнить разбиение группы объектов на заранее заданное количество кластеров « c » в соответствии с некоторым критерием оптимальности. При этом определяются координаты центроидов кластеров. В контексте решения задачи кластеризации ВР координаты центроидов могут быть использованы для формирования обобщающих ВР-центроидов, характеризующих частные (отдельные) ВР, входящие в соответствующие кластеры (рис. 1). При этом,

очевидно, что для ВР-центроида возможна разработка некоторой модели прогнозирования, которая в дальнейшем может быть использована либо непосредственно для прогнозирования частных ВР (отнесенных к данному кластеру), либо в качестве базовой модели – с целью ее дальнейшего уточнения и последующего применения для прогнозирования частных ВР.

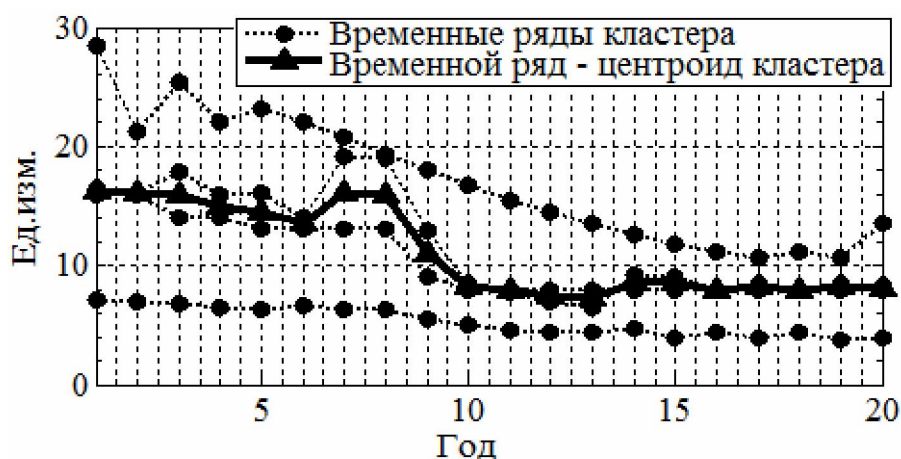


Рис. 1. Подгруппа временных рядов и ВР-центроид

С целью минимизации временных затрат на разработку моделей прогнозирования предлагается при решении задачи кластеризации ВР использовать алгоритм четких C -средних [13], который реализует разбиение группы объектов T на подгруппы T_r ($r = \overline{1, c}$) таким образом, что:

$$\bigcup_{r=1}^c T_r = T; \quad (1)$$

$$T_r \cap T_h = \emptyset; \quad r = \overline{1, c}; \quad h = \overline{1, c}; \quad r \neq h; \quad (2)$$

$$\emptyset \subset T_r \subset T, \quad r = \overline{1, c}. \quad (3)$$

В контексте решения задачи кластеризации ВР под объектом следует понимать именно ВР. Пусть $u_r(t_i)$ – характеристическая функция, которая может принимать два значения: «0», если объект t_i не принадлежит кластеру T_r , и «1», если объект t_i принадлежит кластеру. Тогда четкое C -разбиение группы объектов T на кластеры может быть описано матрицей $U = [u_r(t_i)]$ ($u_r(t_i) \in \{0, 1\}; r = \overline{1, c}; i = \overline{1, m}$).

Матрица $U = [u_r(t_i)]$ должна удовлетворять следующим требованиям:

$$\sum_{r=1}^c u_r(t_i) = 1 \quad (i = \overline{1, m}); \quad (4)$$

$$0 < \sum_{i=1}^n u_r(t_i) < n \quad (r = \overline{1, c}). \quad (5)$$

Алгоритм четких C -средних осуществляет минимизацию целевой функции

$$J(U, V) = \sum_{r=1}^c \sum_{i=1}^m u_r(t_i) \cdot d^2(v_r, t_i), \quad (6)$$

где $U = [u_r(t_i)]$ – четкое C -разбиение группы объектов T на основе характеристических функций $u_r(t_i)$, определяющих принадлежность объекта t_i кластеру T_r ; $V = (v_1, \dots, v_c)$ – цен-

троиды кластеров; $d(v_r, t_i)$ – расстояние между центром кластера v_r и объектом t_i ; c – количество кластеров T_r ; m – количество объектов; $r = \overline{1, c}$; $i = \overline{1, m}$. При реализации алгоритма четких C -средних выполняются следующие шаги [13] (рис. 2).

1. Инициализация случайным образом начального четкого C -разбиения $U = [u_r(t_i)]$, удовлетворяющего требованиям (4) и (5).

2. Вычисление координат центроидов кластеров:

$$v_r^j = (1/m_r) \cdot \sum_{i=1}^m u_r(t_i) \cdot t_i^j, \quad (7)$$

где m_r – количество объектов, отнесенных к r -му кластеру; $j = \overline{1, n}$.

3. Вычисление нового четкого C -разбиения $U = [u_r(t_i)]$, удовлетворяющего требованиям (4) и (5).

4. Шаги 2 и 3 повторяются до достижения наперед заданной точности ε : $|J(U, V) - J'(U, V)| \leq \varepsilon$, где $J(U, V)$, $J'(U, V)$ – значения целевой функции (6) на двух последовательных итерациях (или пока не будет выполнено наперед заданное количество итераций H).

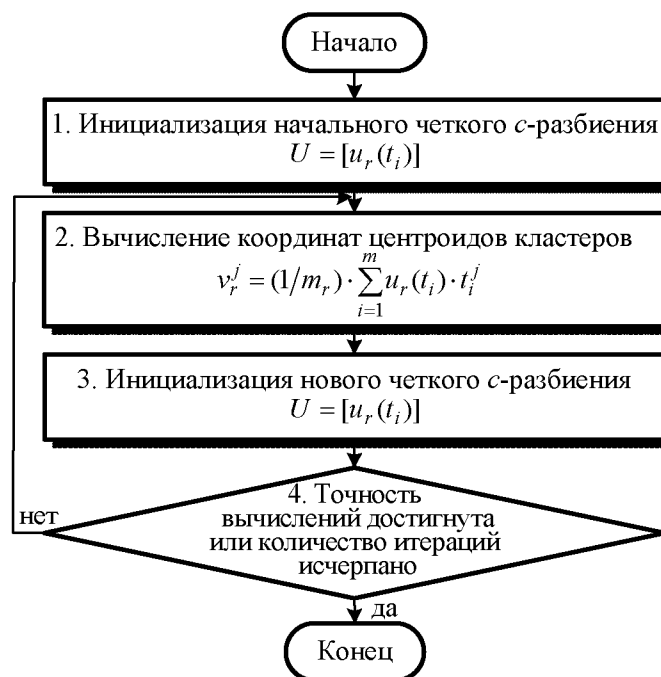


Рис. 2. Схема реализации алгоритма четких c -средних

Алгоритм четких C -средних обеспечивает нахождение локально-оптимального разбиения объектов на кластеры. Для принятия адекватного решения о результатах кластеризации зачастую необходимо многократное выполнение данного алгоритма при заданном количестве кластеров – для различных исходных разбиений объектов на кластеры [13]. Кроме того, целесообразно применение какого-либо эволюционного алгоритма (например, генетического алгоритма) [15, 19, 20, 23], позволяющего существенно сократить время поиска субоптимального разбиения объектов на кластеры. Искомое разбиение характеризуется мини-

мальным значением целевой функции алгоритма кластеризации (минимальным значением некоторого показателя качества кластеризации) [13].

Обычно для вычисления расстояния $d(v_r, t_i)$ между центром кластера v_r и объектом

t_i используется евклидова метрика: $d(v_r, t_i) = \sqrt{\sum_{j=1}^n (v_r^j - t_i^j)^2}$, где n – количество характери-

стик объекта. В контексте решения задачи кластеризации ВР на основе сходства математических законов изменения значений их элементов предлагается выполнить модификацию этой метрики так, чтобы обеспечить учет различной актуальности элементов ВР:

$$d(v_r, t_i) = \sqrt{\sum_{j=1}^n (j/n)(v_r^j - t_i^j)^2} \quad (8)$$

Целесообразность такой модификации объясняется тем, что с течением времени зависимости между теми или иными анализируемыми показателями изменяются. Поэтому при построении моделей прогнозирования большее предпочтение следует отдавать самым близким к моменту прогнозирования элементам ВР. Использование весовых коэффициентов вида « j/n » позволяет считать наиболее значимыми расхождения между значениями самых актуальных элементов ВР (например, при $j=n$ значение весового коэффициента равно «1», а при $j=1$ – равно « $1/n$ »). Использование формулы (8) позволяет не только удовлетворить требование учета актуальности элементов ВР, но и, ввиду высокой чувствительности к разнонаправленности тенденций ВР, обеспечит объединение в кластеры на основе сходства тенденций.

Следует отметить проблему, решение которой является принципиальным в случае применения алгоритмов кластерного анализа. Она связана с разной масштабностью анализируемых ВР, характеризующих те или иные показатели, имеющие различные единицы измерения, различные диапазоны изменения и соответствующие им статистические характеристики (математическое ожидание и т.п.). Для решения этой проблемы целесообразно использование алгоритмов нормализации, которые широко применяются в статистике, математической экономике и эконометрике. Суть их состоит в определении некоторого среднего уровня – медианы, относительно которой выравниваются все анализируемые ВР [1]. В качестве медианы может выступать некоторая условная прямая, один из ВР анализируемой группы ВР или же ВР-центроид S_i , элементы S_j которого определяются как:

$$S_j = (1/m) \cdot \sum_{i=1}^m t_i^j, \quad (9)$$

где t_i^j – j -й элемент i -го ВР; $i = \overline{1, m}$; $j = \overline{1, n}$; m – количество ВР; n – количество элементов ВР-центроида. Алгоритм нормализации i -го ВР ($i = \overline{1, m}$) на основе ВР-центроида S может быть представлен в виде следующей последовательности шагов (рис. 3).

Шаг 1.

1.1. Определяется средний шаг hS изменения элементов ВР-центроида:

$$hS = (\max_{j=\overline{1, n}}(S_j) - \min_{j=\overline{1, n}}(S_j)) / n, \quad (10)$$

где n – количество элементов ВР-центроида; S_j – j -й элемент ВР-центроида S .



Рис. 3. Схема алгоритма нормализации временных рядов

1.2. Определяется средний шаг ht_i изменения элементов i -го ВР ($i = \overline{1,m}$):

$$ht_i = (\max_{j=\overline{1,n}}(t_i^j) - \min_{j=\overline{1,n}}(t_i^j)) / n, \quad (11)$$

где n – количество элементов ВР; t_i^j – j -ый элемент i -го ВР.

Шаг 2.

2.1. Определяется средний уровень ВР-центроида:

$$\bar{S} = (1/n) \cdot \sum_{j=1}^n S_j, \quad (12)$$

где n – количество элементов ВР-центроида; S_j – j -ый элемент ВР-центроида.

2.2. Определяется средний уровень $\bar{t}_i$ i -го ВР ($i = \overline{1,m}$):

$$\bar{t}_i = (1/n) \cdot \sum_{j=1}^n t_i^j, \quad (13)$$

где n – количество элементов ВР; t_i^j – j -ый элемент i -го ВР.

Шаг 3.

Рассчитывается величина Δt_i^j , представляющая собой отношение разности между средним уровнем $\bar{t}_i$ i -го ВР и j -м элементом t_i^j i -го ВР к среднему шагу ht_i :

$$\Delta t_i^j = (\bar{t}_i - t_i^j) / ht_i, \quad (14)$$

где t_i^j – j -ый элемент i -го ВР; $\bar{t}_i$ – средний уровень i -го ВР; ht_i – средний шаг изменения элементов i -го ВР; $i = \overline{1, m}$; $j = \overline{1, n}$.

Величина Δt_i^j может быть интерпретирована как количество шагов (возможно, не целое) длиной ht_i , приходящихся на интервал длиной $|\bar{t}_i - t_i^j|$.

Шаг 4.

Каждый j -ый элемент t_i^j i -го ВР преобразуется к виду:

$$t_i^j = \bar{S} + \Delta t_i^j \cdot hS. \quad (15)$$

Преобразованные таким образом ВР в дальнейшем могут быть использованы для группирования рядов в кластеры с применением алгоритма четких C -средних. Так как центроиды кластеров выражают общие тенденции для подгрупп ВР, формирующих соответствующие кластеры, то целесообразна разработка моделей прогнозирования для ВР-центроидов.

В настоящее время известны различные подходы к построению моделей прогнозирования ВР [4, 7, 9–12, 17, 18, 23]. Одним из перспективных является подход, реализующий применение эволюционных алгоритмов (генетических алгоритмов [9, 10], алгоритмов клонального отбора [2, 11, 12, 14, 25] и т.п.). Они основаны на принципах имитации естественного отбора в биологических системах и обеспечивают (при приемлемых временных затратах) построение моделей прогнозирования ВР, наилучшим образом описывающих известные значения ВР и характеризующихся приемлемыми значениями показателей качества моделей.

В контексте решения задачи разработки моделей прогнозирования для ВР-центроидов кластеров целесообразно применение МАКО, моделирующего законы функционирования естественной иммунной системы и обеспечивающего формирование достаточно сложных аналитических зависимостей [2, 11]. Важное достоинство МАКО – обеспечение возможности поддержания разнообразия популяции решений (антител). Это позволяет предотвратить преждевременную «сходимость» популяции к локальному, а не к глобальному решению (антителу).

В [11, 12] рассмотрены принципы разработки моделей прогнозирования k -го порядка с применением МАКО. Этот алгоритм позволяет сформировать при приемлемых временных затратах на основе СБД аналитическую зависимость, наилучшим образом описывающую известные значения ВР и обеспечивающую получение минимального значения аффинитета (affinity) Aff. В качестве последнего используется средняя относительная ошибка прогнозирования (Average Forecasting Error Rate) $AFER$:

$$AFER = (100\% / (n - k)) \cdot \sum_{j=k+1}^n |(f^j - d^j) / d^j|, \quad (16)$$

где d^j и f^j – соответственно реальное и предсказанное значения для j -го элемента ВР (для j -го отсчета по времени); n – количество элементов ВР (количество отсчетов по времени).

В контексте решения задачи прогнозирования подгрупп ВР в качестве d^j следует использовать, например, j -ый элемент i -го ВР (t_i^j) или соответствующий j -ый элемент v_r^j ВР-центроида r -го кластера.

Возможные варианты аналитических зависимостей представляются в виде антител Ab , осуществляющих распознавание антигенов Ag (известных значений ВР). В качестве «лучшего» варианта выбирается антитело Ab , обеспечивающее минимальное значение аффинитета Aff [11, 12, 14]. Кодирование антитела Ab осуществляется посредством записи в строку символов, выбираемых из трёх символьных алфавитов [11]. (1) Алфавит арифметических операций (операций сложения, вычитания, умножения и деления) $Operation = \{ '+', '-', ' ', '/' \}$. (2) Алфавит функционалов $Functional = \{ 'S', 'C', 'Q', 'L', 'E', ' ' \}$, в котором символы $'S', 'C', 'Q', 'L', 'E'$ определяют математические функции «синус», «косинус», «квадратный корень», «натуральный логарифм», «экспонента», а символ $' '$ – отсутствие какой-либо математической функции. (3) Алфавит терминалов $Terminal = \{ 'a', 'b', ..., 'z', '?' \}$, в котором символы $'a', 'b', ..., 'z'$ определяют аргументы искомой аналитической зависимости, а символ $'?'$ – некоторую константу.

Применение трёх символьных алфавитов обеспечивает корректное преобразование в аналитические зависимости случайным образом формируемых антител, структура которых может быть описана с помощью СБД [2, 11, 12, 14].

Количество символов из алфавита терминалов $Terminal$ в антителе Ab определяет максимально возможный порядок K модели прогнозирования ($K \geq k$, где k – реальный порядок модели) [2, 11, 12, 14]. Иными словами при прогнозировании значения элемента d^j ВР на момент времени j могут использоваться K значений элементов ВР: $d^{j-K}, ..., d^{j-2}, d^{j-1}$.

Использование СБД вида, приведенного на рисунке 4, позволяет строить сложные аналитические зависимости и обеспечивает высокую точность прогнозирования ВР [2, 11, 12].

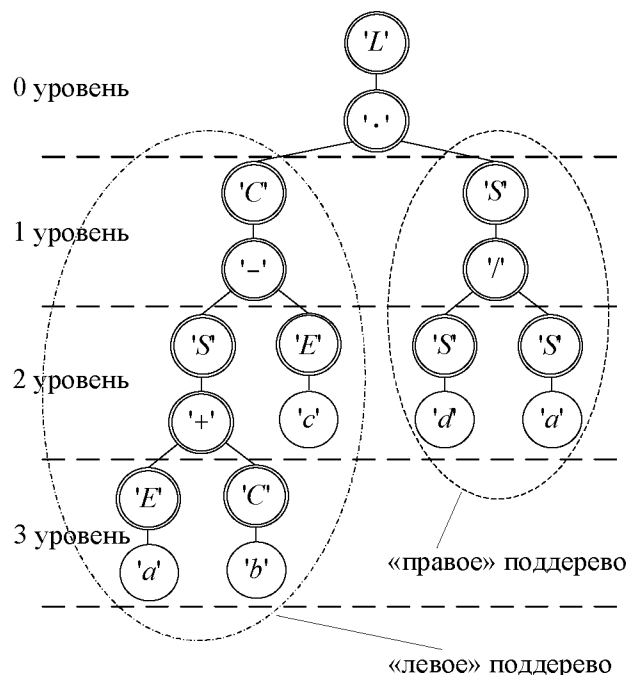


Рис. 4. Пример строго бинарного дерева, используемого для формирования антитела

Такое СБД может быть сформировано в результате композиции одного «левого» поддерева максимально возможного порядка $K = 3$ и некоторого количества «правых» поддерева максимально возможного порядка $K = 2$. При этом термины «левое»/«правое» поддерево используются для указания того, в какую ветвь (левую или правую) некоторого уровня СБД следует включать новое поддерево [2, 11, 12].

В этом случае при формировании антител целесообразно сначала реализовать разбиение СБД на поддеревья; далее выполнить обход вершин каждого поддерева с формированием упорядоченных списков символов, находящихся в его вершинах; затем – провести последовательное объединение этих списков [2].

При формировании упорядоченного списка символов на основе поддерева осуществляется последовательный двукратный обход его вершин. Сначала при движении по поддереву снизу-вверх и слева-направо попарно обходятся вершины, содержащие символы из алфавита терминалов *Terminal*, и соответствующие им расположенные сверху вершины, содержащие символы из алфавита функционалов *Functional*. Затем при движении в том же направлении попарно обходятся вершины, содержащие символы из алфавита арифметических операций *Operation*, и соответствующие им расположенные сверху вершины, содержащие символы из алфавита функционалов *Functional* [2].

В сформированном таким образом антителе первые два символа содержат пару символов нулевого уровня СБД из алфавита функционалов *Functional* и алфавита арифметических операций *Operation*. Затем следуют списки символов, соответствующие «правым» поддеревьям максимально возможного порядка $K = 2$ (при движении по СБД сверху вниз), и, наконец, список символов, входящих в «левое» поддерево максимально возможного порядка $K = 3$ [2].

При использовании такого способа формирования антитела обеспечивается наглядность представления структуры СБД в виде объединения поддерева, а само антитело легко «интерпретируется» в аналитическую зависимость. Например, антитело, сформированное на основе СБД, приведенного на рисунке 4, кодируется строкой символов:

$$L \cdot S / SeSdC - S + EaCbEa, \quad (17)$$

Эта строка может быть преобразована к аналитической зависимости:

$$f(a, b, c, d) = \ln(\cos(\sin(\exp(a) + \cos(b)) - \exp(c)) \cdot \sin(\sin(d) / \sin(a))). \quad (18)$$

В контексте задачи разработки модели прогнозирования k -го порядка при $k = 4$ с учетом упорядоченности символов a, b, c, d в алфавите терминалов *Terminal* аналитическая зависимость (18) может быть записана как:

$$f(d^{j-1}, d^{j-2}, d^{j-3}, d^{j-4}) = \ln(\cos(\sin(\exp(d^{j-1}) + \cos(d^{j-2})) - \exp(d^{j-3})) \cdot \sin(\sin(d^{j-4}) / \sin(d^{j-1}))). \quad (19)$$

При интерпретировании антител в виде аналитических зависимостей целесообразно использовать рекурсивную процедуру интерпретирования [11].

МАКО, применяемый для поиска «лучшего» антитела, определяющего «лучшую» аналитическую зависимость, включает в себя подготовительную часть (реализующую формирование начальной популяции антител) и итерационную часть, в которой выполняются следующие операции: упорядочение антител по возрастанию значений аффинитета Aff; отбор и клонирование части «лучших» антител, характеризующихся наименьшими значениями аффинитета Aff; гипермутация клонов антител; самоуничтожение клонов антител, «похожих» на другие клоны и антитела текущей популяции [два антитела (клона) длиной S символов считаются похожими, если количество совпадений символов при их попарном сравнении в одинаковых позициях кодов антител не меньше S_d , где S_d ($S_d \leq S$) – порог самоуничтожения антитела (клона)]; вычисление аффинитета клонов антител и формирова-

ние новой популяции антител; супрессия полученной популяции антител; генерация новых антител и добавление их к текущей популяции до получения ее исходного размера; проверка условия завершения работы МАКО (по ограничению на количество генераций).

Подробное описание МАКО дано в работе [11], а его укрупненная блок-схема приведена на рисунке 5.

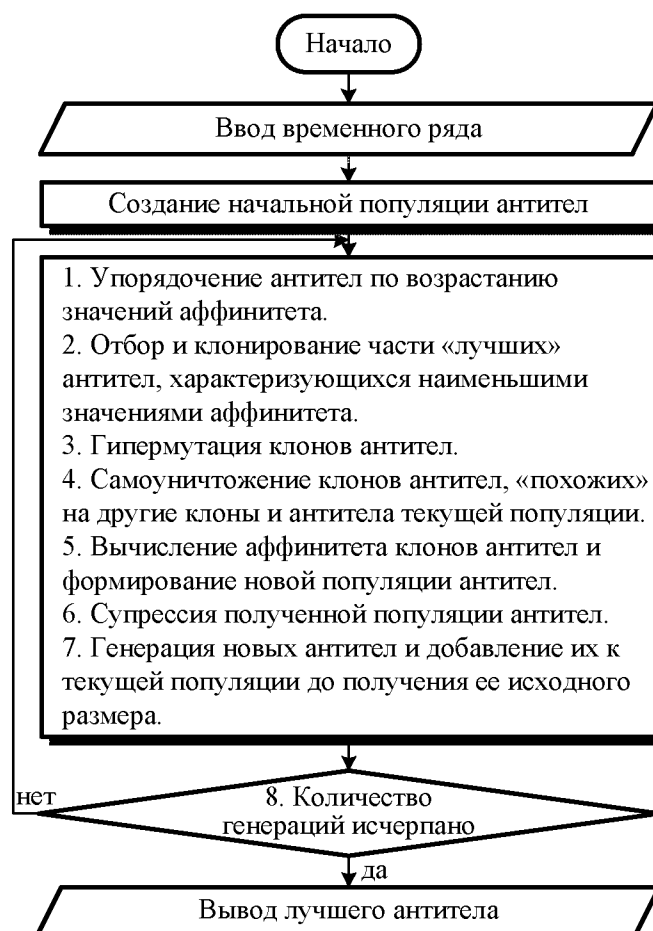


Рис. 5. Укрупненная блок-схема модифицированного алгоритма клонального отбора

Использование общих моделей прогнозирования, построенных для ВР-центроидов кластеров, позволяет обеспечить уникальность результатов прогнозирования для всех частных ВР, отнесенных к кластеру. Данный факт объясняется тем, что при вычислении прогнозных значений для частных ВР на основе аналитических зависимостей, определяющих модели прогнозирования для ВР-центроидов кластеров, в качестве значений переменных выступают значения элементов частных ВР.

С учетом вышеизложенного метод прогнозирования для групп ВР, обеспечивающий получение прогнозов с приемлемыми временными затратами, может быть реализован при выполнении этапов, приведенных на рисунке 6.

Прогнозирование ВР с использованием общих моделей не ведет к получению одинаковых для подгруппы (кластера) ВР прогнозов. Модель прогнозирования определяет лишь математический закон изменения значений элементов ВР посредством формируемой с при-

менением МАКО аналитической зависимости. Сами же прогнозные значения для каждого частного ВР будут уникальными, поскольку будут вычисляться при подстановке в общую модель прогнозирования известных значений элементов частного ВР.

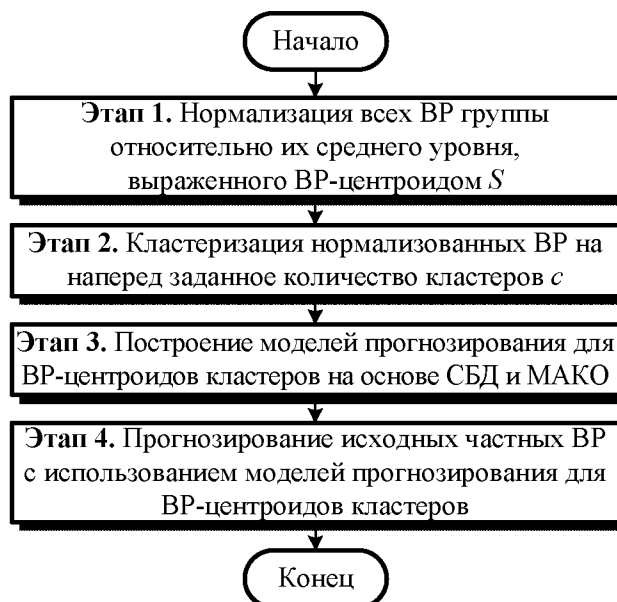


Рис. 6. Этапы прогнозирования групп временных рядов

Экспериментальные исследования. Апробация предлагаемого метода прогнозирования групп ВР была выполнена с использованием рядов для 23 макроэкономических показателей Российской Федерации, взятых с сайта World DataBank за период с 1999 по 2014 г. (<http://databank.worldbank.org/data/views/reports/tableview.aspx?isshared=true#>).

По результатам реализации алгоритма четких C -средних, в котором расстояния между ВР вычислялись в соответствии с (8), все показатели были разделены на 4 кластера (подгруппы ВР). Информация об их содержимом представлена в первом столбце таблицы 1. Данные за период с 1999 по 2011 г. были использованы для разработки общих моделей прогнозирования, а за период с 2011 по 2014 г. – для прогнозирования частных ВР на 3 шага по времени вперед. Данные прогнозов за эти три года сравнивались с фактическими значениями показателей (табл. 1). Кроме того, в табл. 1 приведены значения $AFER$ по формуле (16) при $n = 13$ за период с 1999 по 2011 г.; значения $Error$ средних относительных ошибок прогнозирования на 3 шага вперед за период с 2011 по 2014 г.; средние значения $AFER$ и $Error$ по каждому кластеру и по всем кластерам в целом. Значение $COOP Error$ на 3 шага вперед вычислялось по формуле:

$$Error = (100\% / 3) \cdot \sum_{j=n+1}^{n+3} |(f^j - d^j) / d^j|, \quad (20)$$

в которой при $n = 13$ величины f^j и d^j с индексами $j = 14$, $j = 15$ и $j = 16$ соответствовали предсказанному (прогнозному) и реальному значениям элементов ВР за 2012, 2013 и 2014 гг.

Для сравнения в таблице 2 приведены аналогичные результаты прогнозирования макропоказателей, полученные с применением индивидуальных моделей (при этом для наглядности показатели сгруппированы в кластеры также, как и в таблице 1, хотя в данном случае кластеризация ВР не производилась).

Таблица 1

Результаты прогнозирования макропоказателей с использованием общих моделей

Наименование показателя	Единица измерения	AFER, %	2012		2013		2014		Error, %
			факт	прогноз	факт	прогноз	факт	прогноз	
1-й кластер									
Запуск процедур для регистрации бизнеса (количество)	Кол.	2,24	7,95	8	7	7,95	7,82	7	5,86
Экспорт высоких технологий	%	1,24	8,27	8,4	8,95	8,27	8,57	8,95	5,00
Коэффициент рождаемости		3,24	1,83	1,83	1,08	1,07	1,21	1,63	6,09
Добавленная стоимость в с/х	% от ВВП	2,05	13,6	13,9	13,47	13,6	11,9	13,47	3,51
Военные расходы	% от ВВП	0,25	4,09	3,99	3,07	3,06	3,97	3,97	10,67
Среднее значение AFER по кластеру		1,80	Среднее значение Error по кластеру						6,20
2-й кластер									
Коэффициент подростковой фертильности (в возрасте 15-19 лет)	Число рождений/1000 женщин	0,79	59,59	60,2	59,82	60,48	59,52	59,91	3,97
Добавленная стоимость в промышленности	% от ВВП	0,91	28,97	26,65	29	29,51	28,98	28,36	4,55
Экспорт товаров и услуг	% от ВВП	0,46	34,15	36,51	35,12	33,37	36,16	35,52	2,31
Импорт товаров и услуг	% от ВВП	0,31	29,04	28,7	29,03	28,07	29,05	28,34	1,57
Валовое накопление капитала	% от ВВП	0,33	21,9	22,59	21,87	21,6	21,89	21,96	1,66
Доходы, за исключением грантов	% от ВВП	0,52	24,95	24,37	24,91	24,36	24,84	24,73	2,58
Добавленная стоимость в сфере услуг	% от ВВП	0,19	29,04	29,68	29,14	28,04	28,96	28,45	0,93
Среднее значение AFER по кластеру		0,5	Среднее значение Error по кластеру						2,51
3-й кластер									
Валовой национальный доход на душу населения по методу Атласа	\$	2,36	9642	11037,91	10406	11082	11740	11160,96	8,64
Валовой национальный доход на душу населения по паритету покупательной способности	\$	0,93	19373	20588,31	20861	20627,68	22279	21655,73	3,40
Потребление энергии	Кг. нефт. экв.	0,72	4740	5039,61	5022	5075,21	5114	5083,71	2,66
Потребление электроэнергии	кВт*ч/чел.	0,55	6279	6523,76	6457	6538,07	6485	6540,05	2,00
Среднее значение AFER по кластеру		1,14	Среднее значение Error по кластеру						4,17
4-й кластер									
Количество деторождений при помощи квалифицированного медицинского персонала	%	0,07	99,54	99,85	99,49	99,59	99,82	99,51	0,24
Процент иммунизации против кори	%	0,17	97,68	98,31	97,68	98,28	97,72	98,28	0,61

Процент населения, имеющее начальное образование, всего	%	0,57	96,4	98,27	97,09	98,15	95,19	98,27	2,09
Соотношение девочек и мальчиков в системе начального и среднего образования	%	0,13	98,28	98,7	98,79	98,39	99,29	98,69	0,48
Улучшенные источники воды	%	0,46	96,21	98,26	96,74	98,02	96,66	98,23	1,69
Ожидаемая продолжительность жизни при рождении	Число лет	0,23	72,02	71,45	72,83	72,94	73,13	71,95	0,85
Улучшенные средства санитарии	%	0,62	74,96	71,49	74,86	74,03	74,65	75,48	2,28
Среднее значение <i>AFER</i> по кластеру		0,32	Среднее значение <i>Error</i> по кластеру						1,18
Среднее значение <i>AFER</i> по всем кластерам		0,84	Среднее значение <i>Error</i> по всем кластерам						3,20

Таблица 2

**Результаты прогнозирования макропоказателей
с использованием индивидуальных моделей**

Наименование показателя	Единица измерения	AFER, %	2012		2013		2014		Error, %
			факт	прогноз	факт	прогноз	факт	прогноз	
1-й кластер									
Запуск процедур для регистрации бизнеса (количество)	кол.	1,25	7,95	7,33	7	7,33	7,82	7,72	4,60
Экспорт высоких технологий	%	2,25	8,27	7,56	8,95	8,73	8,57	9,74	8,23
Коэффициент рождаемости		0,82	1,83	1,84	1,08	1,06	1,21	1,13	3,00
Добавленная стоимость в с/х	% от ВВП	0,36	13,6	13,78	13,47	13,64	11,9	11,74	1,31
Военные расходы	% от ВВП	0,74	4,09	3,78	3,07	3,08	3,97	3,96	2,72
Среднее по кластеру		1,16	Среднее значение <i>Error</i> по кластеру						3,97
2-й кластер									
Коэффициент подростковой фертильности (в возрасте 15-19 лет)	Число рождений/1000 женщин	0,05	59,59	59,67	59,82	59,69	59,52	59,69	0,21
Добавленная стоимость в промышленности	% от ВВП	0,19	28,97	25,7	29	28,96	28,98	28,99	3,82
Экспорт товаров и услуг	% от ВВП	0,31	34,15	38,77	35,12	35,02	36,16	36,13	4,63
Импорт товаров и услуг	% от ВВП	0,14	29,04	30,17	29,03	29	29,05	29,86	2,26
Валовое накопление капитала	% от ВВП	0,13	21,9	22,3	21,87	21,9	21,89	21,87	0,69
Доходы, за исключением грантов	% от ВВП	0,96	24,95	24,3	24,91	24,33	24,84	24,33	2,33
Добавленная стоимость в сфере услуг	% от ВВП	0,35	29,04	29,81	29,14	29,19	28,96	28,65	1,30
Среднее по кластеру		0,31	Среднее значение <i>Error</i> по кластеру						1,12

3-й кластер									
Валовой националь- ный доход на душу населения по методу Атласа	\$	1,95	9642	10989,15	10406	10181	11740	11116,68	7,15
Валовой националь- ный доход на душу населения по парите- ту покупательной способности	\$	1,92	19373	22111,39	20861	22241,23	22279	22186,1	7,06
Потребление энергии	Кг. нефт. экв.	1,76	4740	4489,77	5022	5409,37	5114	5441,85	6,47
Потребление электро- энергии	кВт*ч/чел.	0,59	6279	6440,99	6457	6412,21	6485	6279,38	2,15
Среднее по кластеру		1,56	Среднее значение <i>Error</i> по кластеру						5,70
4-й кластер									
Количество деторож- дений при помощи квалифицированного медицинского персо- нала	%	0,05	99,54	99,7	99,49	99,54	99,82	99,49	0,18
Процент иммуниза- ции против кори	%	0,19	97,68	98,57	97,68	98,29	97,72	98,29	0,71
Процент населения, имеющего начальное образование	%	0,31	96,4	97,1	97,09	96,4	95,19	97,09	1,14
Соотношение девочек и мальчиков в систе- ме начального и среднего образования	%	0,14	98,28	98,8	98,79	98,28	99,29	98,79	0,52
Улучшенные источ- ники воды	%	0,13	96,21	97	96,74	96,21	96,66	96,74	0,48
Ожидаемая продол- жительность жизни при рождении	Число лет	0,96	72,02	70,08	72,83	70,09	73,13	70,1	3,53
Улучшенные средства санитарии	%	0,35	74,96	74,48	74,86	72,72	74,65	74,94	1,30
Среднее по кластеру		0,31	Среднее значение <i>Error</i> по кластеру						1,12
Среднее по всем ВР		0,78	Среднее значение <i>Error</i> по всем кластерам						2,86

Общие модели прогнозирования для частных ВР каждого кластера были определены на основе антител, сопоставленных ВР-центроидам кластеров:

$$\begin{aligned} & _+L+E-C/C/C-L\cdot E+_iCeEbSfQhQeEdE?Sd;; \\ & L-S\cdot S/_-C\cdot S-_S-EdLcChEaQg_hQbSeLe; \\ & L+S/_/E+L-S\cdot E\cdot L+_iShSbEeLgChSbEcEc; \\ & Q-_/L-E-Q\cdot E-E\cdot E+QgLfQgQh_fQdS?QcEf; \end{aligned}$$

в форме соответственно аналитических зависимостей вида:

$$\begin{aligned} f_1(d^{j-1}, d^{j-2}, d^{j-3}, d^{j-4}, d^{j-5}, d^{j-6}) = & \ln(\exp(\exp(\sin(\sin(\sqrt{d^{j-3}}) - \\ & \ln(d^{j-4})) \cdot \ln(d^{j-5})) \cdot \cos(\ln(d^{j-5}) - \cos(d^{j-2}))) \cdot \exp(\sin(d^{j-6}) - \\ & \exp(d^{j-1}))) - \exp(d^{j-5}) \cdot \sin(d^{j-1})); \end{aligned}$$

$$\begin{aligned}
 f_2(d^{j-1}, d^{j-2}, d^{j-3}, d^{j-4}, d^{j-5}, d^{j-6}) &= \exp(\sqrt{\cos(\sin(\exp(\sin(d^{j-2})/\sin(d^{j-6}))) - \\
 &\sin(d^{j-5})) \cdot \cos(\sin(d^{j-3}) + 12,25)) \cdot \sin(\ln(d^{j-5}) \cdot \sin(d^{j-2}))) / \cos(\ln(d^{j-1}) + \cos(d^{j-4}))); \\
 f_3(d^{j-1}, d^{j-2}, d^{j-3}, d^{j-4}, d^{j-5}) &= \ln(\sqrt{\sqrt{\sin(\sin(\sqrt{d^{j-3}}) - \exp(d^{j-4})) \cdot \exp(d^{j-2})) - \\
 &\sqrt{\sqrt{d^{j-2}} \cdot \sqrt{d^{j-1}}}}) \cdot \sqrt{\exp(d^{j-5}) - d^{j-1}}) - \sin(\sin(d^{j-2}) - \cos(d^{j-3}))); \\
 f_4(d^{j-1}, d^{j-2}, d^{j-3}, d^{j-4}, d^{j-5}, d^{j-6}) &= \exp(\sin(\cos(\cos(\sin(d^{j-6}) \cdot \\
 &\ln(d^{j-5})) \cdot 56,21) + \sin(d^{j-4}) - \sin(d^{j-2})) + \ln(\ln(d^{j-1}) \cdot \sqrt{d^{j-3}}) \cdot \\
 &\sin(\sin(d^{j-3}) - \sin(d^{j-1}))).
 \end{aligned}$$

Порядок модели прогнозирования определяет сам МАКО, минимизируя СООП *AFER* (16).

Самое существенное влияние на время разработки модели прогнозирования на основе СБД и МАКО оказывают такие параметры МАКО, как количество итераций, размер популяции антител, коэффициент клонирования антител и коэффициент размножения клонов. В рассматриваемом примере было выполнено 400 итераций МАКО для популяции из 20 антител при коэффициентах клонирования антител и размножения клонов, равных соответственно 0,3 и 0,8. Была использована ПЭВМ, работающая под 64-разрядной версией Windows 7, с оперативной памятью 2 Гб и двухядерным процессором Pentium 4 с тактовой частотой каждого ядра 3,4 ГГц. В этом случае на построение одной модели прогнозирования потребовалось 77 секунд. Таким образом, для построения 4-х общих моделей необходимо 308 секунд (5 минут 8 секунд), а для построения 23 индивидуальных моделей потребовалась бы 1 771 секунда (29 минут 31 секунда), что больше в 5,75 раза.

При реализации предлагаемого метода необходимы дополнительные временные затраты, обусловленные необходимостью выполнения процедуры кластеризации ВР. Однако, время, затрачиваемое на процедуру кластеризации, в рассматриваемом примере составляет всего 0,254 секунды. Это гораздо меньше времени, которое необходимо потратить на построение еще 19 моделей прогнозирования (это время составляет 1463 секунды, т.е. 24 минуты 23 секунды).

При применении данного метода к новой группе ВР, то есть впервые, может потребоваться выполнение процедуры кластеризации ВР при разном количестве кластеров «С» для нахождения оптимального разбиения, определяемого по минимуму значения целевой функции (6) алгоритма кластеризации. Однако и в этом случае имеет место существенное снижение временных затрат на прогнозирование для групп ВР.

При использовании общих моделей прогнозирования в рассматриваемом примере усредненное по всем частным ВР значение СООП *AFER* по формуле (16) составило 0,84 %, а усредненное по всем частным ВР значение СООП *Error* на 3 шага вперед по формуле (20) оказалось равным 3,20 % (табл. 1). В то же время применение индивидуальных моделей прогнозирования обеспечивает получение усредненного по всем частным ВР значения СООП *AFER* по формуле (16), равного 0,78 %, и усредненного по всем частным ВР значения СООП *Error* на 3 шага вперед по формуле (20), равного 2,86 % (табл. 2). Таким образом, значения ошибок прогнозирования, полученные в случае применения общих моделей прогнозирования оказались лишь незначительно больше, чем при использовании индивидуальных моделей прогнозирования.

Использование предлагаемого метода прогнозирования групп ВР позволило с приемлемой точностью определить значения прогноза для частных ВР, принадлежащих одному кластеру, но имеющих существенные различия относительно среднего уровня значений своих элементов. При этом за счет использования формулы (8) для вычисления расстояния между ВР была снижена актуальность значений элементов ВР, наиболее удаленных от момента разработки моделей прогнозирования, приходящегося на 2012 г. В частности, с точностью 99,81 %

было оценено будущее значение добавленной стоимости в сфере услуг (в % от ВВП) и с точностью 98 % – потребление электроэнергии (в кВт*ч/чел.) за период с 2011 г. по 2014 г.

На рисунке 7 представлены графические зависимости для значений элементов частных ВР, относящихся соответственно к кластерам 1, 2, 3 и 4 (в том числе отображены значения прогноза на 3 шага по времени вперед).

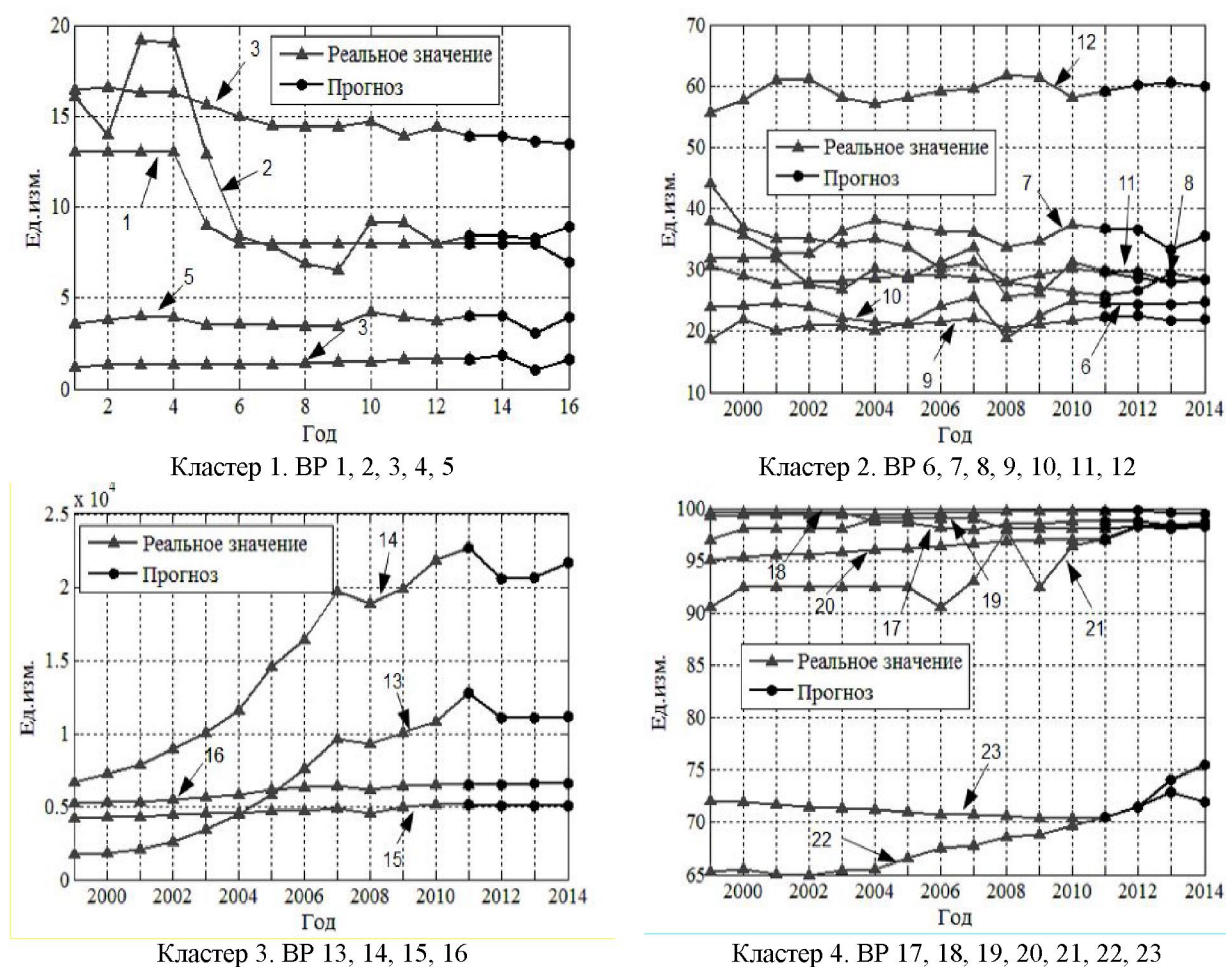


Рис. 7. Результаты прогнозирования для групп временных рядов

Все ВР обозначены в соответствии с нумерацией, приведенной в таблице 1, а под обозначением «Ед. изм.» по вертикальной оси на рисунках следует понимать единицу измерения соответствующего частного ВР (третий столбец табл. 1).

Заключение. Предлагаемый метод прогнозирования групп ВР реализует совместное применение алгоритма четких s -средних и моделей прогнозирования на основе СБД и МА-КО. Это обеспечит получение индивидуальных (частных) прогнозных значений для всех ВР группы с приемлемыми временными затратами. При этом очевидна возможность реализации распараллеливания вычислений, что позволит увеличить скорость выполнения расчетов.

Результаты вычислительных экспериментов, полученные в ходе прогнозирования макроэкономических показателей Российской Федерации, подтверждают перспективность применения и дальнейшего развития предлагаемого метода прогнозирования групп ВР.

С учетом полученных результатов можно сделать вывод о том, что использование алгоритмов кластерного анализа позволяет образовывать кластеры (подгруппы) связанных ВР, имеющих сходные законы изменения значений своих элементов, и обеспечивает увеличение скорости прогнозирования ВР. Применение общих моделей прогнозирования (моделей прогнозирования для ВР-центроидов кластеров) для отдельных (частных) ВР, входящих в соответствующие подгруппы, не приводит к существенному снижению точности прогнозирования. При этом требуемая точность прогноза для частного ВР может быть достигнута в процессе уточнения общей модели прогнозирования с применением МАКО.

Дальнейшие исследования могут быть связаны с анализом применимости при реализации предлагаемого метода прогнозирования других алгоритмов кластерного анализа и подходов к построению моделей прогнозирования (при соблюдении требования о минимизации временных затрат на получение прогнозных значений для всей группы ВР).

Список литературы

1. Андерсен Т. Статистический анализ временных рядов / Т. Андерсен. – Москва : Мир, 1976. – 756 с.
2. Астахова Н. Н. Использование почти полных строго бинарных деревьев и модифицированного алгоритма клонального отбора при разработке моделей прогнозирования временных рядов с короткой актуальной частью / Н. Н. Астахова, Л. А. Демидова // Вестник Рязанского государственного радиотехнического университета. – 2013. – № 4–2 (46). – С. 89–96.
3. Белов В. В. Проблемы факторного прогнозирования социально-экономических показателей / В. В. Белов // Вестник Московского государственного университета приборостроения и информатики. – 2005. – № 2. – С. 116–122.
4. Борисов В. И. Применение мультифрактального анализа кратковременных рядов вариабельности сердечного ритма при оценке результатов лечения артериальной гипертензии / В. И. Борисов, В. С. Кубланов // Прикаспийский журнал: управление и высокие технологии. – 2014. – № 3. – С. 134–143.
5. Брейман А. Д. Многомерное хранение журналов событий для извлечения и анализа процессов / А. Д. Брейман, Е. М. Богословский // Прикаспийский журнал: управление и высокие технологии. – 2014. – № 2. – С. 127–128.
6. Брумштейн Ю. М. Одно- и многомерные временные ряды: анализ возможных методов оптимизации отсчетов и оценки характеристик / Ю. М. Брумштейн, М. В. Иванова // Прикаспийский журнал: управление и высокие технологии. – 2012. – № 4. – С. 34–43.
7. Брумштейн Ю. М. Анализ методов исследования процессов, описываемых взаимосвязанными временными рядами / Ю. М. Брумштейн, М. В. Иванова // Известия Волгоградского государственного технического университета. Серия. Актуальные проблемы управления, вычислительной техники и информатики в технических системах. – 2011. – Вып. 10, №3 (76). – С. 45–51.
8. Воронцов К. В. Модификации ЕМ-алгоритма для вероятностного тематического моделирования / К. В. Воронцов, А. А. Потапенко // Машинное обучение и анализ данных. – 2013. – Т. 1, № 6. – С. 657–686.
9. Демидова Л. А. Разработка однофакторных нечетких моделей для анализа тенденций временных рядов с использованием генетического алгоритма / Л. А. Демидова // Научно-технические ведомости СПбГПУ. – 2007. – № 52–2. – С. 156–164.
10. Демидова Л. А. Прогнозирование тенденций рынка труда на основе однофакторных нечетких временных рядов / Л. А. Демидова // Системы управления и информационные технологии. – 2007. – Т. 29, № 3.2. – С. 241–246.
11. Демидова Л. А. Модели прогнозирования временных рядов с короткой актуальной частью на основе модифицированного алгоритма клонального отбора / Л. А. Демидова // Вестник Рязанского государственного радиотехнического университета. – 2012. – № 39–2. – С. 64–71.
12. Демидова Л. А. Оценка качества моделей прогнозирования на основе строго бинарных деревьев и модифицированного алгоритма клонального отбора / Л. А. Демидова // Cloud of Science. – 2014. – Т. 1, № 2. – С. 202–222.
13. Демидова Л. А. Принятие решений в условиях неопределенности / Л. А. Демидова, В. В. Кираковский, А. Н. Пылькин. – Москва : Горячая линия-Телеком, 2012. – 288 с.

14. Демидова Л. А. Модифицированный алгоритм клонального отбора для анализа временных рядов с короткой длиной актуальной части / Л. А. Демидова, А. В. Корячко, Т. С. Скворцова // Системы управления и информационные технологии. – 2010. – Т. 42, № 4.1. – С. 131–136.
15. Кныш Д. С. Параллельные генетические алгоритмы: обзор и состояние проблемы / Д. С. Кныш, В. М. Курейчик // Известия Российской академии наук. Теория и системы управления. – 2010. – № 4. – С. 72–82.
16. Паклин Н. Б. Бизнес-аналитика от данных к знаниям / Н.Б. Паклин, В.И. Орешков. – Санкт-Петербург : Питер, 2013. – 704 с.
17. Пилюгина А. А. Опыт использования аппарата нечетких множеств в прогнозировании валютного курса / А. А. Пилюгина, А. А. Бойко // Прикаспийский журнал: управление и высокие технологии. – 2014. – № 3. – С. 143–157.
18. Петрушин В. Н. Формализация временного ряда методом двойного сглаживания / В. Н. Петрушин, Г. О. Рытиков // Cloud of Science. – 2014. – Т. 1, № 2. – С. 230–238.
19. Полковникова Н. А. Нейросетевые технологии, нечёткая кластеризация и генетические алгоритмы в экспертной системе / Н. А. Полковникова, В. М. Курейчик // Известия Южного федерального университета. Технические науки. – 2014. – № 7 (156). – С. 7–15.
20. Рутковская Д. Нейронные сети, генетические алгоритмы, нечеткие системы / Д. Рутковская, М. Пилиньский, Л. Рутковский. – Москва : Горячая линия-Телеком, 2004. – 452 с.
21. Терехов А. А. Идентификация статистического материала и консолидация временных рядов / А. А. Терехов // Вестник Рязанского государственного радиотехнического университета. – 2009. – № 27. – С. 62–70.
22. Чубукова И. А. Data Mining : учебное пособие / И. А. Чубукова. – Москва : Интернет-университет информационных технологий : БИНОМ : Лаборатория знаний, 2006. – 382 с.
23. Щербаков М. В. Применение эволюционных коннективистских моделей идентификации состояний динамики систем для решения задачи прогнозирования потребления электроэнергии в автоматическом режиме / М.В. Щербаков, И. П. Козлов, Н. Л. Щербакова // Прикаспийский журнал: управление и высокие технологии. – 2011. – № 4. – С. 70–75.
24. Штовба С. Д. Проектирование нечетких систем средствами Matlab / С. Д. Штовба. – Москва : Горячая линия – Телеком, 2007. – 288 с.
25. Demidova L. A. Time series forecasting models on the base of modified clonal selection algorithm / L. A. Demidova // 2014 International conference on computer technologies in physical and engineering applications (ICCTPEA). – Санкт-Петербургский государственный университет, 2014. – С. 33–34.
26. Lee L. W. Handling forecasting problems based on two-factors high-order fuzzy time series / L. W. Lee, L. H. Wang, S. M. Chen, Y. H. Leu // IEEE Transactions on fuzzy systems. – 2006. – Vol. 14, № 3. – P. 468–477.
27. Tahseen A. J. Multivariate high order fuzzy time series forecasting for car road accidents / A. J. Tahseen, S. M. Aqil Burney, C. Ardil // International journal of computational intelligence. – 2007. – Vol. 4, № 1. – P. 15–20.

References

1. Andersen T. *Statisticheskiiy analiz vremennykh ryadov* [Statistical time series' analysis]. Moscow, Mir Publ., 1976. 756 p.
2. Astakhova N. N., Demidova L. A. Ispolzovanie pochni polnykh strogo binarnykh derev i modifitsirovannogo algoritma klonalnogo otbora pri razrabotke modeley prognozirovaniya vremennykh ryadov s korotkoy aktualnoy chastyu [Use of almost full strictly binary trees and modified clonal selection algorithm for developing forecasting models of time series with short actual part]. *Vestnik Ryazanskogo gosudarstvennogo radiotekhnicheskogo universiteta* [Bulletin of the Ryazan State Radio Engineering University], 2013, no. 4–2 (46), pp. 89–96.
3. Belov V. V. Problemy faktornogo prognozirovaniya sotsialno-ekonomiche-skikh pokazateley [Factorial forecasting problems of socio-economic indexes]. *Vestnik Moskovskogo gosudarstvennogo universiteta priborostroeniya i informatiki* [Bulletin of Moscow State University of Instrument Making and Informatics], 2005, no 2, pp. 116–122.

4. Borisov V. I., Kublanov V. S. Primenenie multifraktalnogo analiza kratkovremennykh ryadov variabelnosti serdechnogo ritma pri otsenke rezultatov lecheniya arterialnoy gipertenzii [Multifractal analysis application of short-term time series of warm rhythm variability at an assessment of results of arterial hypertension treatment]. *Prikaspiyskiy zhurnal: upravlenie i vysokie tekhnologii* [Caspian Journal: Management and High Technologies], 2014, no. 3, pp. 134–143.

5. Breyman A. D., Bogoslovskiy Ye. M. Mnogomernoe khranenie zhurnalov sobytiy dlya izvlecheniya i analiza protsessov [Multidimensional storage of logs for processes' extraction and analysis]. *Prikaspiyskiy zhurnal: upravlenie i vysokie tekhnologii* [Caspian Journal: Management and High Technologies], 2014, no. 2, pp. 127–128.

6. Brumshteyn Yu. M., Ivanova M. V. Odn- i mnogomernye vremennye ryady: analiz vozmozhnykh metodov optimizatsii otschetov i otsenki kharakteristik [One and multidimensional time series: analysis of optimization possible methods of counting and assessment of characteristics]. *Prikaspiyskiy zhurnal: upravlenie i vysokie tekhnologii* [Caspian Journal: Management and High Technologies], 2012, no. 4, pp. 34–43.

7. Brumshteyn Yu. M., Ivanova M. V. Analiz metodov issledovaniya protsessov, opisyyaemykh vzaimosvyazannymi vremennymi ryadami [Research methods' analysis of processes, described by the interconnected time series]. *Izvestiya Volgogradskogo gosudarstvennogo tekhnicheskogo universiteta. Seriya. Aktualnye problemy upravleniya, vychislitel'noy tekhniki i informatiki v tekhnicheskikh sistemakh* [News of Volgograd State Technical University. Series. Actual Problems of Management, Computer Facilities and Informatics in Technical Systems], 2011, vol. 10, no. 3 (76), pp. 45–51.

8. Vorontsov K. V., Potapenko A. A. Modifikatsii EM-algoritma dlya veroyatnostnogo tematicheskogo modelirovaniya [Modifications of EM-algorithm for probabilistic thematic modeling]. *Mashinnoe obucheniye i analiz dannykh* [Machine training and data analysis], 2013, vol. 1, no. 6, pp. 657–686.

9. Demidova L. A. Razrabotka odnofaktornykh nechetkikh modeley dlya analiza tendentsiy vremennykh ryadov s ispolzovaniem geneticheskogo algoritma [Development of one-factorial fuzzy models for the analysis of tendencies of temporary ranks with use of genetic algorithm]. *Nauchno-tekhnicheskie vedomosti SPbGPU* [Scientific and Technical Sheets of SPbGPU], 2007, no. 52-2, pp. 156–164.

10. Demidova L. A. Prognozirovaniye tendentsiy rynka truda na osnove odnofaktornykh nechetkikh vremennykh ryadov [Tendencies' forecasting of labor market on the base of one-factorial fuzzy time series]. *Sistemy upravleniya i informatsionnye tekhnologii* [Control Systems and Information Technologies], 2007, vol. 29, no. 3.2, pp. 241–246.

11. Demidova L. A. Modeli prognozirovaniya vremennykh ryadov s korotkoy aktualnoy chastyu na osnove modifitsirovannogo algoritma klonalnogo otbora [Time series' forecasting models with short actual part on the base of modified clonal selection algorithm]. *Vestnik Ryazanskogo gosudarstvennogo radio-tekhnicheskogo universiteta* [Bulletin of the Ryazan State Radio Engineering University], 2012, no. 39–2, pp. 64–71.

12. Demidova L. A. Otsenka kachestva modeley prognozirovaniya na osnove strogo binarnykh derev'ev i modifitsirovannogo algoritma klonalnogo otbora [Forecasting models' quality assessment on the base of strictly binary trees and modified clonal selection algorithm]. *Cloud of Science*, 2014, vol. 1, no. 2, pp. 202–222.

13. Demidova L. A., Kirakovskiy V. V., Pylkin A. N. *Prinyatie resheniy v usloviyakh neopredelennosti* [Making-decision in the conditions of uncertainty.], Moscow, Goryachaya liniya-Telekom Publ., 2012. 288 p.

14. Demidova L. A., Koryachko A. V., Skvortsova T. S. Modifitsirovannyy algoritm klonalnogo otbora dlya analiza vremennykh ryadov s korotkoy dlinoy aktualnoy chasti [Modified clonal selection algorithm for analysis of time series with a short length of actual part]. *Sistemy upravleniya i informatsionnye tekhnologii* [Control System and Information Technologies.], 2010, vol. 42, no. 4.1, pp. 131–136.

15. Knysh D. S., Kureychik V. M. Parallelnye geneticheskie algoritmy: obzor i sostoyaniye problema [Parallel genetic algorithms: review and condition of a problem]. *Izvestiya Rossiyskoy akademii nauk. Teoriya i sistemy upravleniya* [News of Russian Academy of Sciences. Theory and Control Systems], 2010, no. 4, pp. 72–82.

16. Paklin N. B., Oreshkov V. I. *Biznes-analitika ot dannykh k znaniya* [Business-analitik's outlets: from data to knowledge], Saint Petersburg, Piter Publ., 2013. 704 p.

17. Pilyugina A. A., Boyko A. A. Opyt ispolzovaniya apparata nechetkikh mnozhestv v prognozirovanii valyutnogo kursa [Use experience of fuzzy set theory in an exchange rate forecasting]. *Prikaspiyskiy zhurnal: upravlenie i vysokie tekhnologii* [Caspian Journal: Management and High Technologies], 2014, no. 3, pp. 143–157.
18. Petrushin V. N., Rytikov G. O. Formalizatsiya vremennogo ryada metodom dvoynogo sglazhivaniya [Formalization of time series by method of double smoothing]. *Cloud of Science*, 2014, vol. 1, no. 2, pp. 230–238.
19. Polkovnikova N. A., Kureychik V. M. Neyrosetevye tekhnologii, nechetkaya klasterizatsiya i geneticheskie algoritmy v ekspertnoy sisteme [Neural network technologies, fuzzy clustering and genetic algorithms in expert system]. *Izvestiya Yuzhnogo federalnogo universiteta. Tekhnicheskie nauki* [News of the Southern federal university. Technical science], 2014, no. 7 (156), pp. 7–15.
20. Rutkovskaya D., Pilinskiy M., Rutkovskiy L. Neyronnye seti, gene-ticheskie algoritmy, nechetkie sistemy [Neural networks, genetic algorithms, fuzzy systems], Moscow, Goryachaya liniya-Telekom Publ., 2004. 452 p.
21. Terekhov A. A. Identifikatsiya statisticheskogo materiala i konsolidatsiya vremennykh ryadov [Statistical material identification and time series' consolidation]. *Vestnik Ryazanskogo gosudarstvennogo radiotekhnicheskogo universiteta* [Bulletin of the Ryazan State Radio Engineering University], 2009, no. 27, pp. 62–70.
22. Chubukova I. A. *Data Mining* [Data Mining: manual], Moscow, Internet University of Information Technologies Publ. House, BINOM Publ., Laboratory of knowledge Publ., 2006. 382 p.
23. Shcherbakov M. V., Kozlov I. P., Shcherbakova N. L. Primenenie evolyutsionnykh konnektivistskikh modeley identifikatsii sostoyaniy dinamiki sistem dlya resheniya zadachi prognozirovaniya potrebleniya elektroenergii v avtomaticheskom rezhime [Application of evolutionary connective models of systems' dynamics conditions' identification for the solution of a forecasting problem of electricity consumption in the automatic mode]. *Prikaspiyskiy zhurnal: upravlenie i vysokie tekhnologii* [Caspian Journal: Management and High Technologies], 2011, no. 4, pp. 70–75.
24. Shtovba S. D. *Proektirovanie nechetkikh sistem sredstvami Matlab* [Fuzzy systems' design by means of Matlab], Moscow, Goryachaya liniya – Telekom Publ., 2007. 288 p.
25. Demidova L. A. Time series forecasting models on the base of modified clonal selection algorithm [Time series forecasting models on the base of modified clonal selection algorithm]. *2014 International conference on computer technologies in physical and engineering applications (ICCTPEA)*, Saint Petersburg State University Publ. House, 2014, pp. 33–34.
26. Lee L. W., Wang L. H., Chen S. M., Leu Y. H. Handling forecasting problems based on two-factors high-order fuzzy time series. *IEEE Transactions on fuzzy systems*, 2006, vol. 14, no. 3, pp. 468–477.
27. Tahseen A. J., Aqil Burney S. M., Ardil C. Multivariate high order fuzzy time series forecasting for car road accidents. *International journal of computational intelligence*, 2007, vol. 4, no. 1, pp. 15–20.

УДК 004.4:[616.2+616-073.75]

**АНАЛИЗ ПРИНЦИПОВ РАЗРАБОТКИ СИСТЕМЫ
ПОПУЛЯЦИОННОГО СКРИНИНГА ОНКОЛОГИЧЕСКИХ ЗАБОЛЕВАНИЙ
ЛЕГКИХ НА ОСНОВЕ ИСПОЛЬЗОВАНИЯ
РЕНТГЕНОВСКОЙ КОМПЬЮТЕРНОЙ ТОМОГРАФИИ**

Марусина Мария Яковлевна, доктор технических наук, профессор, заведующая кафедрой, Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики (Университет ИТМО), 197101, Российская Федерация, г. Санкт-Петербург, Кронверкский пр., 49, e-mail: marusina_m@mail.ru

Волгарева Александра Павловна, магистрант, Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики (Универ-