УДК 681.518

Л.А. Демидова, Т.С. Скворцова

ПРИМЕНЕНИЕ ГЕНЕТИЧЕСКИХ АЛГОРИТМОВ ДЛЯ ПРОГНОЗИРОВАНИЯ НЕЧЕТКИХ ВРЕМЕННЫХ РЯДОВ

Рассмотрена задача прогнозирования значений временных рядов на основе аппарата теории нечетких множеств. Обосновано применение генетического алгоритма для выбора оптимальных параметров двухфакторной модели прогнозирования и предложен вариант его реализации.

Ключевые слова: *генетический алгоритм, нечеткие множества, модель прогнозирования.*

Введение. Генетические алгоритмы (ГА) – это процедуры поиска оптимальных решений прикладных задач, основанные на механизмах естественного отбора и наследования. В них используется эволюционный принцип выживания наиболее приспособленных особей [1, 2].

Существуют два главных преимущества ГА перед классическими оптимизационными методами [1]: не предъявляется каких-либо существенных требований к видам целевых функций и ограничений; поиск экстремума целевой функции осуществляется одновременно по многим направлениям путем использования популяции возможных решений, а переход от одной популяции к другой позволяет избежать попадания в локальный оптимум, при этом ГА характеризуется полиномиальной сложностью вычислений.

Цель работы. Целью работы является исследование возможности применения ГА к решению задачи поиска оптимальных свободных параметров модели краткосрочного прогнозирования значений временного ряда (на один шаг вперед) при минимальных временных затратах.

Постановка задачи. Модель прогнозирования строится на основе исторических (накопленных) данных двух факторов: главного $d_1, d_2, \ldots, d_n$ и вспомогательного $e_1, e_2, \ldots, e_n$, где n – длина актуальной части ряда (количество наблюдений временного ряда), составляющая 20-30 значений.

Представим эти данные как нечеткие временные ряды $F_1(t)$ и $F_2(t)$, где $F_1(t)$ соответствует главному, а $F_2(t)$ – вспомогательному факторам прогнозирования. Тогда зависимость вида

$$(F_1(t-k), F_2(t-k)), \ldots, (F_1(t-2), F_2(t-2)),$$
$$(F_1(t-1), F_2(t-1)) \to F_1(t) \qquad (1)$$

называется двухфакторной моделью прогнозирования k-го порядка на основе нечетких временных рядов [3].

Универсум U для значений главного фактора определяется как $U = [D_{min} - D_1,\ D_{max} + D_2]$, где D_{min} и D_{max} – минимальное и максимальное значения из $d_1, d_2, \ldots, d_n$ ($i = \overline{1,n}$), а D_1 и D_2 – два действительных числа, позволяющие разбить универсум U на p интервалов $u_1, u_2, \ldots, u_p$ равной длины. Универсум V для значений вспомогательного фактора определяется аналогично как $V = [E_{min} - E_1,\ E_{max} + E_2]$ и разбивается на q интервалов $v_1, v_2, \ldots, v_q$ равной длины.

Лингвистические термы A_i ($i = \overline{1,p}$), представленные нечеткими множествами значений главного фактора, имеют следующий вид:

$$A_1 = 1/u_1 + 0{,}5/u_2 + 0/u_3 + \ldots + 0/u_{p-1} + 0/u_p;$$
$$A_2 = 0{,}5/u_1 + 1/u_2 + 0{,}5/u_3 + 0/u_4 + \ldots + 0/u_p;$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$A_p = 0/u_1 + 0/u_2 + \ldots + 0{,}5/u_{p-1} + 1/u_p.$$

Выбор значений функции принадлежности, равных 0, 0,5 и 1, позволяет упростить дальнейшие вычисления [3, 4].

При фаззификации данных для каждого значения главного фактора определяется интервал u_i, которому оно принадлежит, а также соответствующее нечеткое значение этого фактора:

$$X_i = \begin{cases} 1/A_1 + 0{,}5/A_2, \text{ если } i = 1; \\ 0{,}5/A_{i-1} + 1/A_i + 0{,}5/A_{i+1}, i = \overline{2, p-1}; \\ 0{,}5/A_{p-1} + 1/A_p \text{ для } i = p. \end{cases}$$

Лингвистические термы B_j и соответствующие нечеткие значения Y_j ($j = \overline{1,q}$) для вспомо-

гательного фактора определяются аналогично.

По полученным данным для всех значений d_i ($i = \overline{k+1,n}$) временного ряда строятся нечеткие логические зависимости:

$$(X_{i-k}, Y_{i-k}),...(X_{i-2}, Y_{i-2}),(X_{i-1}, Y_{i-1}) \to X_i, \quad (2)$$

которые непосредственно используются для выполнения прогноза [3, 4].

В качестве критерия оптимальности целесообразно использовать минимальное значение средней относительной ошибки прогнозирования *AFER* (Average Forecasting Error Rate) [4]:

$$AFER = \frac{\sum_{i=k+1}^{n} \left| (d_i^* - d_i)/d_i \right|}{n-k} \cdot 100\ \%, \quad (3)$$

где d_i^* и d_i – предсказанное и реальное значения для *i*-го периода прогноза ($i = \overline{k+1,n}$).

Прогнозируемое значение для *i*-го периода прогноза вычисляется по формуле [3, 4]:

$$d_i^* = \frac{w_{j_1} \cdot t_{j_1} + w_{j_2} \cdot t_{j_2} + ... + w_{j_p} \cdot t_{j_p}}{w_{j_1} + w_{j_2} + ... + w_{j_p}}.$$

Здесь w_j – коэффициенты, учитывающие повторы в нечетких логических зависимостях, а переменные t_j определяются по формуле:

$$t_j = \begin{cases} (\tilde{u}_1 + 0{,}5 \cdot \tilde{u}_2)/(1 + 0{,}5), & j = 1 \\ \dfrac{(0{,}5 \cdot \tilde{u}_{j-1} + \tilde{u}_j + 0{,}5 \cdot \tilde{u}_{j+1})}{(0{,}5 + 1 + 0{,}5)}, & 2 \le j \le n-1, \\ (0{,}5 \cdot \tilde{u}_{n-1} + \tilde{u}_n)/(0{,}5 + 1), & j = n \end{cases}$$

где $\tilde{u}_{j-1}, \tilde{u}_j, \tilde{u}_{j+1}$ - средние точки интервалов u_{j-1}, u_j, u_{j+1} соответственно [3, 4].

При использовании изложенного подхода особое внимание должно уделяться выбору свободных параметров модели прогнозирования, к которым относятся ее порядок *k*, числа коррекции диапазонов D_1, D_2, E_1, E_2, а также количества интервалов *p* и *q*, на которые разбиваются универсумы *U* и *V* соответственно.

Оптимальные значения этих параметров обеспечивают минимум показателя *AFER*, причем в базовой модели [3] они выбираются вручную. В данной работе для сокращения времени поиска и повышения точности результатов предлагается использовать классический ГА [1], где структура хромосомы будет иметь вид:

$$S = (D_1, D_2, E_1, E_2, p, q, k). \quad (4)$$

Если для некоторой хромосомы S_i все правые части групп нечетких логических зависимостей определены (не пусты) [4], то значение функции приспособленности вычисляется как $J(S_i) = AFER$. В том случае, когда не определена правая часть хотя бы одной такой зависимости, невозможно гарантировать вычисление прогнозируемого значения для нового члена временного ряда. Такую хромосому необходимо исключить из популяции как наихудшую, вычислив ее функцию приспособленности $J(S_i)$ как сумму средней относительной ошибки прогноза *AFER* и числа 100. Окончательно функция приспособленности принимает вид [4]:

$$J(S_i) = \begin{cases} AFER, \text{ если определены все правые} \\ \quad \text{части логических зависимостей;} \\ AFER + 100 - \text{в противном случае.} \end{cases} \quad (5)$$

Хромосома, для которой в результате выполнения заданного числа итераций ГА достигается минимум функции приспособленности (5), определяет оптимальный набор параметров модели прогнозирования.

Следует заметить, что поиск минимального значения функции (5) классическими методами численной оптимизации является весьма затруднительным, так как область допустимых решений задачи определяется неизвестными заранее пределами изменения параметров модели прогнозирования. Кроме того, указанная область будет невыпуклой и несвязной, поскольку переменные D_1, D_2, E_1, E_2 являются непрерывными, а переменные *p*, *q*, *k* – дискретными (целыми).

Реализация ГА выполняется по классической схеме [2] и включает выполнение следующих шагов.

Шаг 1. Создается начальная популяция размера *N* из случайным образом выбранных хромосом S_i ($i = \overline{1,N}$) вида (4).

Генерация выполняется случайным выбором аллелей для каждого гена. Если первоначальная популяция окажется неконкурентоспособной, то ГА переведет её в жизнеспособную популяцию. Для этого достаточно, чтобы значения генов находились в таких пределах, которые позволяют вычислить функцию приспособленности (5). Поэтому параметры D_1, D_2, E_1, E_2 представляются действительными числами, а количество интервалов *p* и *q* находятся в пределах от 2 до *n*. Порядок модели ограничен значениями от 1 до 5, что объясняется достаточно короткой длиной анализируемых временных рядов.

Шаг 2. При $g < G$, где *g* и *G* – текущее и максимальное количество генераций, вычисляется значение функции приспособленности (5) для каждой хромосомы. Затем создаются пары хромосом-родителей и осуществляется переход к следующему шагу 3. После выполнения *G*

генераций осуществляется переход к шагу 5.

Выбор родителей основан на комбинации методов рулетки и ранговой селекции [2] и имитирует естественный отбор: в родительскую пару включаются хромосомы с лучшими значениями функции приспособленности $J(S_i)$. Вероятность P_i выбора родителя с хромосомой S_i можно рассчитать по формуле:

$$P_i = [J_{max} - J(S_i)] / \sum_{j=1}^{N} [J_{max} - J(S_j)], \qquad (6)$$

где J_{max} – наихудшее значение функции $J(S_i)$ среди хромосом текущего поколения. Тогда выбор родителя будет состоять в определении хромосомы с меньшим значением функции (5) из двух случайно отобранных с учетом вероятностей P_i ($i = \overline{1,N}$). Пара найденных таким образом хромосом-родителей будет использоваться для скрещивания.

Шаг 3. Выполняются операции скрещивания и мутации для созданных пар хромосом-родителей текущей популяции.

Скрещивание заключается в передаче участков генов от родителей к потомкам. Для этого формируется случайное число N_c, равномерно распределенное на отрезке [0, 1], которое сравнивается с заданным коэффициентом скрещивания R_c. Если $R_c > N_c$, то операция скрещивания выполняется в точке z (номер гена в хромосоме), также выбранной случайным образом.

Мутация выполняется с некоторой вероятностью P_m, при которой происходит замена аллеля случайным значением. Оно выбирается с равной вероятностью в области определения гена. Заметим, что именно благодаря мутации расширяется область генетического поиска.

При выполнении мутации задается коэффициент мутации R_m и генерируется случайное число N_m, равномерно распределенное на отрезке [0, 1]. Если $R_m > N_m$, то случайным образом выбирается точка мутации z.

Шаг 4. Создается новая популяция, дополненная хромосомами-потомками. Затем хромосомы с худшими значениями функции приспособленности (5) отбрасываются. В итоге размер популяции сокращается до первоначального.

Шаг 5. Из полученной популяции выбирается хромосома с минимальным значением функции приспособленности (5).

Алгоритм заканчивает работу после выполнения заданного количества генераций G, причем можно показать, что вычислительная сложность предлагаемого генетического алгоритма пропорциональной величине $O(Gn^2N + GN^2)$, т.е. является *полиномиальной* и наиболее заметно зависит от размера популяции N.

Заключение. Эффективность разработанного ГА можно показать на следующем примере.

Модель прогнозирования строится по данным температуры с 1 по 24 июня 2007 года в г. Камбарка Удмуртской Республики. В качестве вспомогательного фактора рассматривается влажность воздуха. Для случая $N = 30$ и $G = 1000$ получены следующие оптимальные параметры модели: $D_1 = -1{,}148$; $D_2 = 0{,}485$; $E_1 = -0{,}908$; $E_2 = 1{,}187$; $p = 14$; $q = 14$; $k = 2$.

Средняя относительная ошибка прогнозирования по известным значениям ряда составляет $AFER = 2{,}168$ %. Прогнозируемое и реальное значения температуры 25 июня равны 14,6 °С и 14,8 °С соответственно. Относительная ошибка прогноза составляет 1,374 %. Графическая зависимость результатов прогнозирования по обучающей выборке приведена на рисунке.

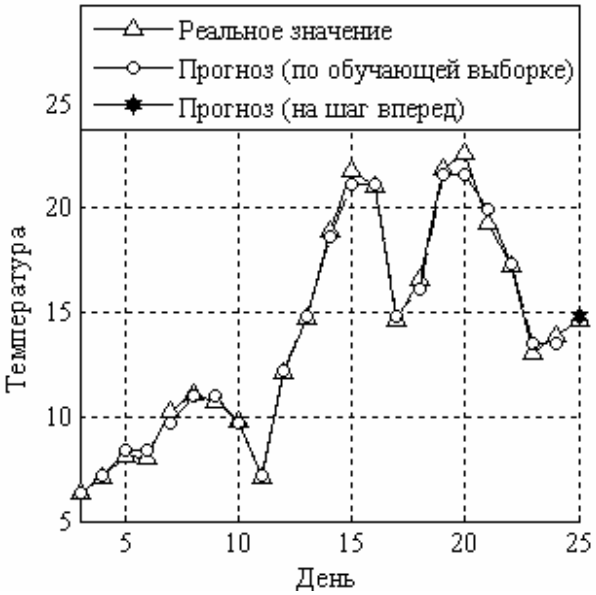

Рисунок – Графическая зависимость реальных и прогнозируемых значений

Проведенные дополнительные эксперименты позволяют подтвердить вывод о том, что сочетание параметров модели, полученное ГА, является достаточно близким к оптимальному. В частности, изменение найденных чисел корректировки границ диапазонов в небольших пределах приводит к увеличению ошибки *AFER*.

Например, для $D_1 = -1$, $D_2 = 1$, $E_1 = -1$ и $E_2 = 1$ получим $AFER = 2{,}704$ %, прогнозируемое значение на 25 июня 14,28 °С (относительная ошибка 2,204 %); для $D_1 = -0{,}5$, $D_2 = 0{,}1$, $E_1 = -0{,}5$, $E_2 = 1{,}5$ будем иметь $AFER = 3{,}238$ % и прогнозируемое значение на 25 июня 14,09°С (относительная ошибка 3,459 %).

Таким образом, разработанный ГА поиска оптимальных параметров двухфакторной модели прогнозирования на основе аппарата теории нечетких множеств позволяет повысить качество прогноза за счет минимизации функции приспособленности – средней относительной ошибки прогнозирования *AFER* .

Библиографический список

1. *Дьяконов В.П., Круглов В.В.* MATLAB 6.5 SP1/7/7 SP1/7 SP2 + Simulink 5/6. Инструменты искусственного интеллекта и биоинформатики. Серия «Библиотека профессионала». – М.: СОЛОН-ПРЕСС, 2006. – 456 с.: ил.

2. *Рутковская Д., Пилиньский М., Рутковский Л.* Нейронные сети, генетические алгоритмы и нечеткие системы. М.: Горячая линия – Телеком, 2006. 452 с.

3. *Lee L. W., Wang L. H., Chen S. M.* Handling forecasting problems based on two-factors high-order fuzzy time series // *IEEE Transactions on fuzzy Systems*. 2006, V.14. № 3. P. 468-477.

4. *Демидова Л.А., Скворцова Т.С.* Разработка двухфакторной модели прогнозирования временных рядов с использованием генетического алгоритма // Математическое и программное обеспечение вычислительных систем: межвуз. сб. М.: Горячая линия – Телеком, 2008. С. 99-108.