ГЕНЕТИЧЕСКИЙ ПОДХОД К ЗАДАЧАМ ПРОГНОЗИРОВАНИЯ

Скобцов Ю.А., Хмелевой С.В.
Донецкий национальный технический университет, г. Донецк
кафедра автоматизированных систем управления
E-mail: hmelevoy_sergey@ukr.net

Abstract

Skobtsov Y.A. Khmilovyy S.V. Evolutionary approach to prognosing problems.
A survey of timeseries prognosing methods using genetical algorithms and genetical programming are presented in the article. Different forms of individuals` coding – potential problem solving, problem-oriented genetic operators and main kinds of fitness-functions are considered.

Введение.

В связи с развитием рыночной экономики и перестройкой общественных отношений на Украине с одной стороны классические методы прогнозирования становятся неэффективными, и с другой стороны современные методы прогнозировании поведения участников рынка, апробированные в странах с развитой экономикой, требуют значительной доработки и адаптации к условиям переходной экономики Украины.

Поэтому все более актуальными становятся проблемы прогнозирования экономических процессов при создания автоматизированных систем управления. Поскольку показатели, характеризующие экономические процессы, описываются временными рядами, то в качестве методов решения таких задач могут использоваться статистические методы прогнозирования временных рядов. К сожалению, такие методы не всегда удовлетворяют конкретным практическим требованиям и малоприменимы к конкретным практическим ситуациям. В то же время современные прикладные проблемы требуют усовершенствования методов идентификации и построения прогнозных моделей, в частности, повышения гибкости, адаптивности, интеллектуальности моделей. Для этого применяются методы искусственного интеллекта.

Одним из наиболее перспективных подходов среди методов искусственного интеллекта в настоящее время является генетический подход. Он базируется на моделировании процессов эволюции, происходящих в природе. Применению генетического подхода к задаче прогнозирования и посвящена данная статья.

Постановка задачи прогнозирования.

В данной работе рассматривается задача одношагового прогнозирования, что подразумевает предсказание значения прогнозируемой величины в следующий момент времени. Задача прогнозирования заключается в нахождении величины изменения ее значения, которое произойдет в следующий момент времени, на основе анализа прошлых значений, а также истории изменения других факторов, влияющих на динамику прогнозируемой величины.

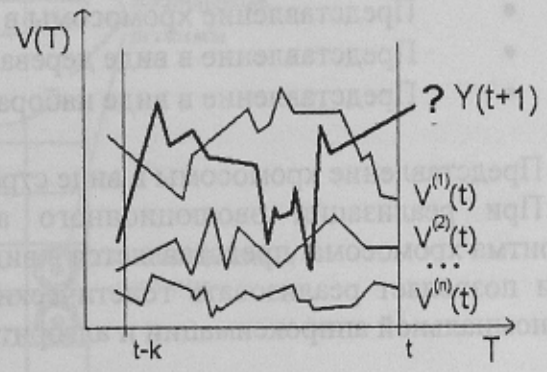

Рисунок 1 - К постановке задачи

При этом наиболее важным является правильное определение точного значения прогнозируемой величины.

Математически это может быть сформулировано следующим образом: требуется получить выходной сигнал $Y(t+1)$ на основе входного сигнала X $Y=g(X)$, который обеспечивал бы заданную точность прогноза.

При этом входной сигнал содержит несколько факторов, представленных совокупностью своих значений, начиная с момента времени $t-k$ до момента времени t (рис. 1):

$X=X(Y_{t-k}; v1_{t-k} \dots v1_t; v2_{t-k} \dots v2_t; \dots ; vn_{t-k} \dots vn_t)$,

где

Vi $i=1...p$ - факторы, значимо влияющие на изменение прогнозируемой величины

t – текущий момент времени

к – глубина ретроспективной выборки. В общем случае к различно для разных факторов.

Формальное описание генетического алгоритма.

Формально генетический алгоритм можно описать следующим образом:

$$ГА = (P^0, \lambda, I, s, \rho, f, t), \text{ где}$$

$P^0 = (a_1^0, ..., a_n^0)$, исходная популяция, где a_i^0 - это решение задачи, представленной в виде хромосомы.

λ - размер популяции (целое число).

l - длина каждой хромосомы.

s - оператор отбора.

ρ - оператор рекомбинации (кроссинговер, мутация и т.п.).

f - функция оптимальности

t - критерий остановки

Таким образом, для того, чтобы определить генетический алгоритм решения данной задачи, необходимо задать параметры: размер популяции, критерий остановки, оператор отбора, а также определить хромосому и генетические операторы на ней. Кроме этого, необходимо определить для хромосомы функцию, позволяющую оценивать ее близость к оптимальному решению (фитнесс-функцию).

Подходы при прогнозировании в зависимости от способа кодировки хромосом.

При применении в прогнозировании существуют различные группы эволюционных алгоритмов (по представлению хромосомы):

- Представление хромосомы в виде строки коэффициентов
- Представление в виде дерева
- Представление в виде набора продукций

Представление хромосомы в виде строки коэффициентов.

При реализации эволюционного алгоритма в виде стандартного генетического алгоритма хромосома представляется в виде вектора коэффициентов. Данное представление особи позволяет реализовать генетический подход к прогнозированию с использованием полиномиальной аппроксимации и алгоритма ZET.

Полиномиальная аппроксимация.

В этом случае ГА может использоваться для нахождения коэффициентов при независимых переменных полинома или показателе степеней переменных. Возможно также

применение различных нелинейных регрессионных моделей (логистические, экспоненциальные и т.п.).

В качестве примера можно привести, например, линейный полином 1.51+2.06*X+13.7*X^2. Очевидно, этот полином может быть представлен вектором коэффициентов К0, К1, К2.

1.51	2.06	13.70

В этом случае применяются стандартные генетические операторы.

Кроссинговер. Для двух родителей

$$k_1^1, k_2^1, \ldots, k_{l,}^1 k_{l+1}^1, \ldots, k_n^1 \text{ и } k_1^2, k_2^2, \ldots, k_{l,}^2 k_{l+1}^2, \ldots, k_n^2.$$

выбирается случайная точка кроссинговера, допустим, с индексом 1 и производится обмен подстроками. В результате кроссинговера имеем два потомка:

$$k_1^1, k_2^1, \ldots, k_{l,}^1 k_{l+1}^2, \ldots, k_n^2 \text{ и } k_1^2, k_2^2, \ldots, k_{l,}^2 k_{l+1}^1, \ldots, k_n^1.$$

Мутация. Случайным образом выбирается один ген хромосомы, значение которого также случайно изменяется в заданном диапазоне.

Алгоритм ZET. Разработан эволюционный алгоритм LGAP, который основан на модификации алгоритма ZET для заполнения пробелов [4].

В этом случае пространством поиска является двухвходовая таблица «объект-время». Строки таблицы представляют значения характеристик в один из моментов времени, а столбцы – факторы, влияющие на прогнозируемую величину. Моменты времени (t = 1,2,...,T) упорядочены в таблице по «возрасту»: самые свежие данные имеют индекс t=1, данные за предшествующий день t = 2 и т. д. до дня с индексом t = T.

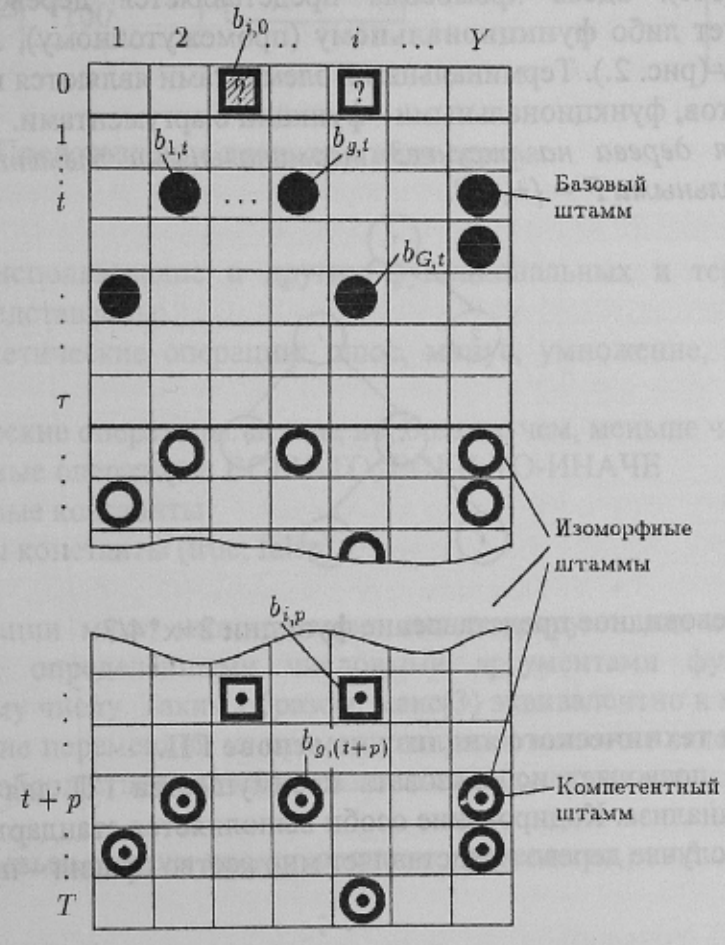

Рисунок 2 - Представление штаммов

Для прогнозирования используется некоторое подмножество элементов таблицы – базовый штамм, который и является особью – потенциальным решением. Каждый из таких штаммов может смещаться по оси времени, образуя изоморфные штаммы. Из множества базовых штаммов необходимо выбрать такой, который дает наименьшую ошибку прогнозирования – компетентный штамм (рис. 2.).

Кодирование штаммов может выполняться следующими способами:

- **Двоичное кодирование.** При значимой глубине выборки τ и числе значащих факторов - Y каждая особь может быть представлена двоичной матрицей τ×Y, которую можно для удобства развернуть в строку, содержащую τ*Y элементов. В этом случае элемент строки равен 1, если соответствующий элемент таблицы включен в базовый штамм, и 0 – если не включен.
- **Целочисленное кодирование.** При размере штамма k особь может быть задана строкой из k ячеек, в каждой из которых записан индекс элемента, учавствующего в базовом штамме.

Поскольку каждое из этих представлений является строкой, применимы стандартные операторы кроссинговера и мутации, описанные выше.

Для построения прогноза необходимо вычислить значение соответствующего элемента нулевой строки (t=0). В случае линейной зависимости можно воспользоваться прогнозированием на основе линейной регрессии.

Представление в хромосомы виде дерева.

При представлении хромосомы в виде дерева используется аппарат генетического программирования (ГП). Здесь хромосома представляется деревом, каждый элемент которого соответствует либо **функциональному** (промежуточному), либо **терминальному** (конечному) элементу (рис. 2.). Терминальными элементами являются константы, действия и функции без аргументов, функциональными - функции с аргументами.

Например, для дерева на рисунке3 терминальными элементами являются T = {2,x,4,7}, а функциональными F = {+,,/}.*

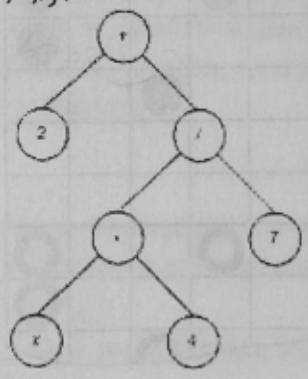

Рисунок 3 - Древовидное представление функции 2+x*4/7

Расчет правил технического анализа на основе ГП.

Данный метод позволяет использовать преимущества ГП для прогнозирования на основе технического анализа. Кодирование особи выполняется стандартными средствами ГП в виде дерева. В этом случае дерево представляет множество правил – продукций вида [1]:

ЕСЛИ отношение цены X к доходу – 10% или ниже, чем среднее
И цена X больше чем минимум цены за последние 63 дня,
ТОГДА X повышается.

Здесь функциональными элементами являются {ЕСЛИ-ТО-ИНАЧЕ, И, НЕ, <, >}, а терминальными - прежде всего прошлые значения прогнозируемой переменной (индексы лагов временного ряда).

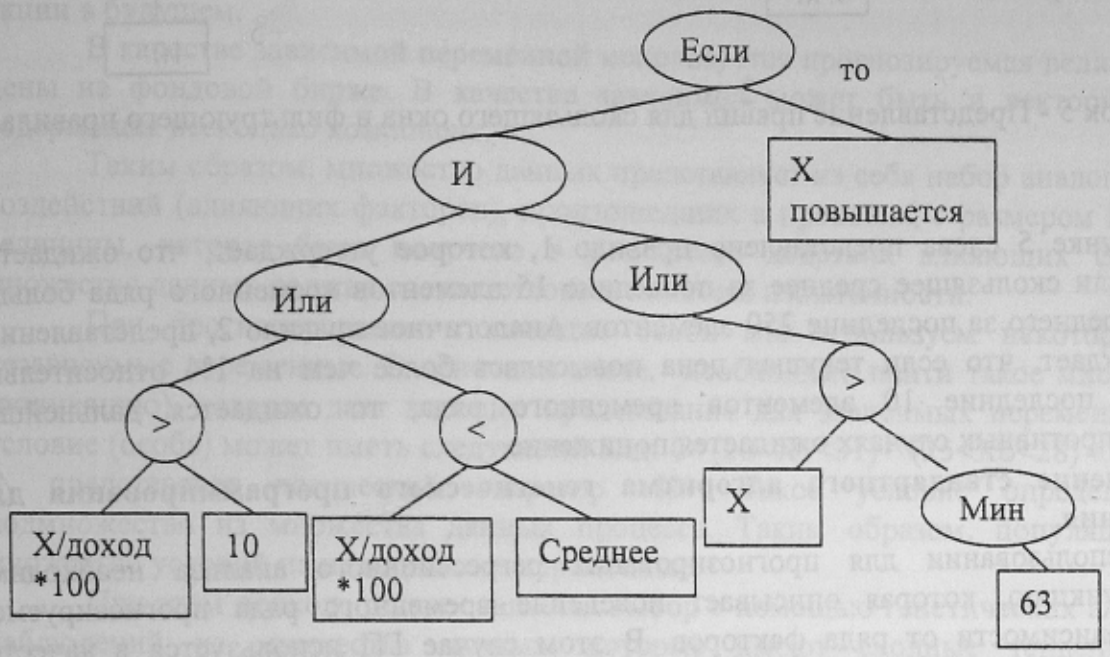

Рисунок 4 - Представление правила в виде дерева

Возможно использование и других функциональных и терминальных элементов. Например, в [5] представлены:

1. Арифметические операции: плюс, минус, умножение, деление, среднее, макс, мин, лаг, норма .
2. Логические операторы: и, или, не, больше чем, меньше чем
3. Условные операторы: ЕСЛИ-ТО, ЕСЛИ-ТО-ИНАЧЕ
4. Числовые константы
5. Булевы константы (true, false)

Здесь операции макс, мин, среднее, лаг оперируют над скользящими окнами во временных рядах, определенными числовыми аргументами функций, округленных к ближайшему целому числу. Таким образом, макс(3) эквивалентно к макс(p_{t-1}, p_{t-2}, p_{t-3}), лаг (3) возвращает значение переменной из временного ряда, взятого с задержкой 3 (p_{t-3}). Операция норма возвращает абсолютное значение разницы между двумя числами.

Пример деревьев с таким представлением приведен на рис.5.

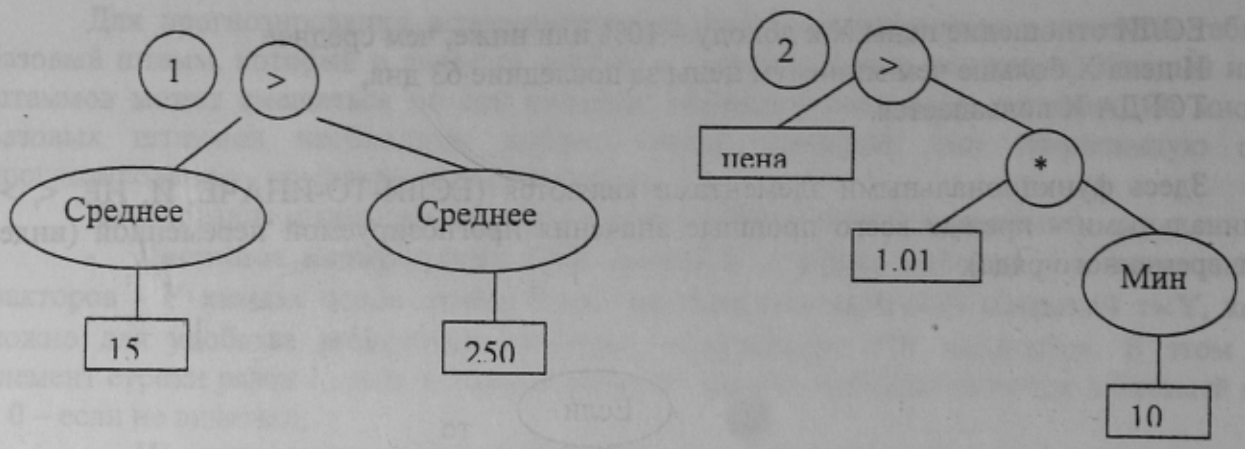

Рисунок 5 - Представление правил для скользящего окна и фильтрующего правила

На рисунке 5 слева представлено правило 1, которое утверждает, что ожидается повышение, если скользящее среднее за последние 15 элементов временного ряда больше скользящего среднего за последние 250 элементов. Аналогичное правило 2, представленное справа, утверждает, что если текущая цена повысилась более чем на 1% относительно минимума за последние 10 элементов временного ряда, то ожидается дальнейшее повышение. В противных случаях ожидается понижение.

Применение стандартного алгоритма генетического программирования для прогнозирования.

При использовании для прогнозирования регрессионного анализа необходимо определить функцию, которая описывает поведение временного ряда прогнозируемой величины в зависимости от ряда факторов. В этом случае ГП используется в качестве символьной регрессии. Здесь множество узлов и терминалов могут быть определены таким образом [6]:

Таблица 1. Множество узлов и терминалов для алгоритма.

Символ	Узлов	Описание
+	2	Сложение
-	2	Вычитание
*	2	Умножение
/	2	Защищенное деление (x/0=1)
Sin	1	Функция синуса
Cos	1	Функция косинуса
Exp	1	Экспонента
Rlog	1	Защищенный логарифм ((r log 0)=0, в противном случае это log(|x|))
XN	0	Входной параметр системы (x1 возвращает 1-й лаг (x_{-1}))
<число>	0	Случайная константа в пределах [-1,1)

Для всех описанных выше древовидных представлений (символьная регрессия и технический анализ) применяются стандартные операторы кроссинговера и мутации ГП.

Представление хромосомы в виде набора продукций. В работе [**Ошибка! Источник ссылки не найден.**] исследованы возможности применения ГА к некоторым проблемам анализа данных и прогнозирования временных рядов. Общая проблема может

быть сформулирована в следующем виде: серия наблюдений некоторого процесса имеет вид {(x1,y1), … ,(xn,yn)}, где xi= (xi1,…, xin) – независимые и yi – соответственно зависимые переменные. Например, в прогнозировании погоды независимые переменные представляют характеристики погоды в настоящем или прошлом: средняя влажность, среднее барометрическое давление, наибольшая и наименьшая температура за сутки, наличие дождя; зависимая переменная соответствует некоторым признакам будущей погоды (например, наличие дождя завтра). В задаче прогнозирования цен на фондовой бирже независимые переменные x=(x(t1), x(t2), …, x(tn)) могут представлять стоимость какой-либо акции в указанные моменты времени, а зависимая переменная y= x(tn+k) определяет стоимость акции в будущем.

В качестве зависимой переменной используется прогнозируемая величина, например, цены на фондовой бирже. В качестве зависимой может быть и векторная переменная, содержащая несколько компонент.

Таким образом, множество данных представляет из себя набор аналогичных входных воздействий (влияющих факторов), произошедших в прошлом, с размером прогнозируемой величины, которая была получена в результате действия влияющих факторов. Такое множество данных должно соответствовать гипотезе избыточности.

При прогнозировании в качестве особи мы используем некоторое условие на независимые переменные. В конечном счете, необходимо найти такое множество условий (популяцию), которое дает хорошее предсказание для зависимых переменных. Например, условие (особь) может иметь следующий вид: C=(20<X1<31) ^ (25<X6<28) ^ (19<X9<27), где ^ представляет логический оператор «И». Такое условие определяет некоторое подмножество из множества данных процесса. Таким образом, популяцию составляют множество условий на независимые переменные.

При этом подходе целью является выбор с помощью генетических алгоритмов таких наблюдений из множества данных, которые имеют сходные тенденции изменений независимых переменных, и близкие значения зависимых переменных. Эти зависимые переменные собственно и определяют прогнозируемые значения. Это условие также должно однозначно удовлетворять и такому наблюдению, для которого и выполняется прогноз. Таким образом, генетические алгоритмы выбирают из тех фактов, которые уже состоялись в прошлом, такие, которые имеют достаточно много общего с фактом, который имеет место в настоящем. Соответственно, работа особей алгоритма соответствует гипотезе локальной компактности. Можно предположить, что если из похожих фактов можно сделать близкие выводы, подобный же вывод можно сделать и из текущего наблюдения на основании гипотезы линейых зависимостей, допустим, простой операцией усреднения зависимой(ых) переменной(ых).

Каждая особь популяции представляется линейной структурой (унарным деревом). Каждый узел данного дерева имеет атрибуты: «имя переменной», «левая граница», «правая граница», «потомок». Таким образом, каждый узел определяет диапазон для какой-либо одной переменной. К нему может быть присоединен (или нет) узел – потомок. Таким образом, особь представляется унарным деревом, которое имеет максимум N узлов (которые соответствует N независимым переменным), и для каждой из них определен диапазон изменения (верхняя и нижняя граница).

Для данного представления наиболее часто применяются следующие операторы рекомбинации:

Кроссинговер. Более применимой к данному методу представления особи является версия оператора кроссинговера, основанная на равновероятном случайном выборе узлов от каждого родителей. Каждый ген (ограничение на значения переменной) потомка наследуется у одного из родителей с вероятностью $P_c \approx 0{,}5$.

Мутация. Для данного представления особей возможны следующие операторы мутации:

1. Добавление ограничения (добавляется в дерево новое ограничение).
2. Удаление ограничения (удаляется один случайный узел из дерева).
3. Расширение или сужение диапазона.
4. Сдвиг диапазона вверх или вниз:
5. Полная перестройка дерева. Особь полностью удаляется, вместо нее заново случайным образом генерируется новая особь.

Типовые фитнесс-функции, используемые при прогнозировании.

Прежде всего, необходимо сказать, что вид фитнесс-функция в значительной степени зависит от конкретной задачи и способа кодирования хромосомы. Хотя, можно сказать, что для задачи прогнозирования, где исходными данными являются временные ряды прогнозируемой величины и влияющих факторов, число различных видов фитнесс-функций весьма ограничено. В любом случае значение фитнесс-функции должна быть тем больше, чем больше отличается прогнозируемое значение от действительного. В простейшем случае фитнесс-функция может быть пропорциональна разности между этими величинами.

Поскольку обучение производится на временном ряде, для оценки качества прогноза может браться среднее значение погрешностей для каждого из элементов временного ряда

$$E = \frac{1}{n}\sum_{i=1}^{n}(x_i - \tilde{x}_i)$$, где

x_i - действительное значение прогнозной величины,

$\tilde{x}_i$ - спрогнозированная величина,

i – индекс номера элемента во временном ряде,

n – количество элементов во временном ряде.

В этом случае возможна ситуация, когда разность между значениями фитнесс - функций наилучшей и наихудшей величин слишком мала, для нормального функционирования ГА. Тогда можно в качестве фитнесс-функции можно взять среднеквадратичную ошибку

$$E = \frac{1}{n}\sum_{i=1}^{n}(x_i - \tilde{x}_i)^2 .$$

Эти типы фитнесс-функций обладают значительным недостатком – зависимостью величины погрешности от величины данных, что особенно важно при большой разнице в размерности прогнозной величины.

Для исправления этого недостатка может быть взято абсолютное отклонение прогноза $\tilde{x}_i$ от истинного значения x_i, деленное на истинное значение:

$$E = \frac{1}{n}\sum_{i=1}^{n}|\tilde{x}_i - x_i| / x_i ,$$

Такая относительная величина мало чувствительна к ошибкам прогноза больших значений и чрезмерно чувствительна к ошибкам прогноза величин, близких к нулю. Кроме того, разность между минимальным и максимальным значениями может быть различной у разных наблюдаемых характеристик и одинаковая относительная ошибка будет приемлемой для принятия решений в одних случаях и не приемлемой в других.

О точности прогноза возможно судить по величине ошибки, нормированной по разнице между максимальным и минимальным значением ряда

$$E = \frac{1}{n}\sum_{i=1}^{n} |\tilde{x}_i - x_i| / (x_{i\max} - x_{i\min})$$

Такая мера обладает одинаковой чувствительностью к ошибкам прогноза для разных значений прогнозируемой характеристики. Ее чувствительность к ошибкам тем выше, чем в меньших пределах колеблется прогнозируемая характеристика, что представляется вполне логичным.

Иногда важно знать не абсолютную величину $b_{i,0}$ характеристики в будущем, а лишь то, будет ли она больше или меньше значения $b_{i,t}$ в данный момент времени. И таких случаях применима мера точности прогноза, учитывающая лишь совпадения знаков:

$$d^* = \begin{cases} 0, & \text{если } (b_{i,0} > b_{i,t}) \text{ и } (b'_{i,0} > b_{i,t}) \\ & \text{или } (b_{i,0} < b_{i,t}) \text{ и } (b'_{i,0} < b_{i,t}) \\ 0.5, & \text{если } (b_{i,0} = b_{i,t}) \text{ и } (b'_{i,0} \neq b_{i,t}) \\ & \text{;в других случаях} \end{cases}$$

Возможно также определить нормированную дисперсию ошибки в качестве фитнесс-функции:

$$E = \frac{1}{\sigma^2} * \frac{1}{n} * \sum_{i=1}^{n} (x_i - \tilde{x}_i)^2 .$$

Нормализация дисперсии удаляет зависимость динамического интервала данных от величины множества данных.

Кроме этого можно использовать составляющие, которые зависят от размера хромосомы, например, для регулировки размера особи в генетическом программировании.

Для продукционного подхода подобные виды фитнесс-функции малоприменимы, вследствие того, что особь представляет собой условие на множестве наблюдений некоторого процесса, который необходимо прогнозировать. Соответственно, фитнесс-функция при этом подходе должна состоять из нескольких составляющих. Во-первых, близость между собой действительного значения и прогноза играет важнейшую роль. Во-вторых, качество прогноза зависит от количества аналогичных ситуаций, которые смогли отыскать в множестве наблюдений. Чем больше выборка, отобранная особью из множества наблюдений, тем лучше. В-третьих, качество прогноза зависит от того, насколько близки между собой наблюдаемые данные в выборке, отобранной особью. Чем более близки между собой данные, тем более качественна выборка. Наиболее компактная выборка, которую можно сделать, состоит из одного наблюдения. И в четвертых, возможны разнообразные дополнительные масштабные, штрафные составляющие для фитнесс-функции . Таким образом, третье условие зачастую конфликтует со вторым, и каждое из них может мало согласоваться с первым. Согласованность второго и третьего составляющих с первым возможна только при условии представительности выборки и повторяемости ситуаций для наблюдаемого процесса. Таким образом, эффективность такого подхода весьма мала для неизученных областей пространства поиска и достигает максимума при стандартных ситуациях.

Например, в [2] используется следующая фитнесс-функция:

$$f(C) = -\log\frac{\partial}{\partial 0} - \frac{\alpha}{Nc} + \Delta$$

где σ - стандартное отклонение (дисперсия) для зависимых переменных из множества данных, удовлетворяющих заданному условию C;

σ0 – стандартное отклонение (дисперсия) для зависимых переменных всего множества обучающих данных;

Nc – количество наблюдений, которые удовлетворяют данному ограничению C;

α и Δ - константы.

Приведенная фитнесс-функция имеет три составляющие, где

- первая составляющая оценивает разброс данных, отобранных данным условием. Чем ближе между собой данные в пространстве независимых переменных, тем меньше дисперсия для условия C, и тем большее значение имеет первая составляющая.

- вторая составляющая оценивает размер выборки, которую представляет данное условие. Чем большее число наблюдений удовлетворяет данному условию, тем меньше значение этой составляющей и тем больше значение целевой функции. Таким образом, вторую составляющую можно назвать «штрафом для условий с бедной статистикой». Для того, чтобы вторая составляющая была соизмерима с первой, введен масштабный коэффициент α.

- третья составляющая -Δ. введена для того, чтобы существовала возможность регулировки величины фитнесс-функции относительно нуля.

Выводы.

В работе выполнен анализ возможных подходов к решению задач прогнозирования с использованием генетических алгоритмов. Показано, что, поскольку применение эволюционного подхода к задаче прогнозирования основывается на математическом аппарате, применяемом в математической статистике, эффективность конкретной реализации генетического алгоритма зависит не только от его вида , но и от выбранного математического аппарата, наследуя от этого аппарата как многие из его преимуществ, так и многие из недостатков.

Эффективность генетического алгоритма прогнозирования также основана на правильно подобранной фитнесс-функции. Приведены различные виды фитнесс-функций, применяющиеся при прогнозировании, показаны их недостатки и преимущества .

Литература

1. Jin Li, Edward P.K. Tsang. Improving Technical Analysis Prediction: An Application of Genetic Programming. – Department of Computer Science, University of Essex, Wivenhoe Park, Colchester, United Kingdom

2. Christopher Neely, Paul Weller, Rob Dittmar. "Is Technical Analysis in the Foreign Exchange Market Profitable? A Genetic Programming Approach." – The Journal of Financial and Quantitative Analysis, Vol 32, №4 (Дек. 1997), стр 405 – 426.

3. Harri Jaske. Prediction of sunspots by GP. The 2nd Finnish Workshop on Genetic Algorithms. Universiry of Vaasa, Finland. ftp://ftp.uwasa.fi/cs/2NWGA/Jaske.ps.z

4. Н.Г. Загоруйко "Прикладные методы анализа данных" Новосибирск, Изд-во Института Математики, 1999.

M.Mitchel. An Introduction to Genetic Algorithms. MIT Press, 1998