УДК 004.032.26(045)

МЕТОД РЕШЕНИЯ ЗАДАЧИ ПРОГНОЗИРОВАНИЯ НА ОСНОВЕ КОМПЛЕКСИРОВАНИЯ ОЦЕНОК

В.М. Синеглазов, Е.И. Чумаченко, В.С. Горбатюк

Национальный технический университет Украины «КПИ»
lobach21@mail.ru

У статті розглядаються основні методи прогнозування. Також пропонується новий алгоритм, заснований на комплексуванні оцінок моделей, отриманих за допомогою різних методів (таких як штучні нейронні мережі, метод групового урахування аргументів і їх комбінації). Алгоритм було протестований на реальних даних і показав кращі результати, ніж кожен метод окремо.
Ключові слова: прогнозування часових рядів, МГУА, штучні нейронні мережі, комплексування результатів.

In this article the main forecasting methods are considered. There has been offered a new algorithm based on a complexing of estimates given by models obtained with the help of different methods (artificial neural networks, group method of data handling and their combination). Algorithm was tested on the real data and showed better results, as compared to separate methods on their own.
Key words: time series forecasting, GMDH, artificial neural networks, results complexing.

В статье рассматриваются основные методы прогнозирования. Также предлагается новый алгоритм, основанный на комплексировании оценок моделей, полученных с помощью различных методов (таких как искусственные нейронные сети, метод группового учета аргументов и их комбинации). Алгоритм был протестирован на реальных данных и показал лучшие результаты, чем каждый метод в отдельности.
Ключевые слова: прогнозирование временных рядов, МГУА, искусственные нейронные сети, комплексирование результатов.

Вступление

Задача прогнозирования всегда была и будет одной из наиболее интересных и изучаемых задач, так как ее решение дает возможность судить о будущем имея лишь текущие данные. С другой стороны, окончательное ее решение вероятнее всего не будет найдено никогда - поскольку количество факторов, влияющих на поведение прогнозируемого процесса обычно стремится к бесконечности. В данной статье предлагается метод нахождения приближенного решения этой задачи, основанный на на переборе различных моделей, выборе лучших (по внешнему критерию) моделей, и, наконец, комплексировании оценок, полученных с помощью выбранных моделей.

1. Постановка задачи

Пусть заданы n дискретных отсчетов $\{y_1,\ldots,y_n\}$ в последовательные моменты времени $\{t_1,\ldots,t_n\}$. Тогда рассматриваемая в данной статье задача прогнозирования (рис. 1) состоит в предсказании значения y_{n+k} в некоторый

будущий момент времени t_{n+k}, где k – длительность (упреждение) прогноза, в виде некоторой функциональной зависимости:

$$y_{n+k} = F(y_1, \ldots, y_n),$$

где F – функциональный преобразователь.

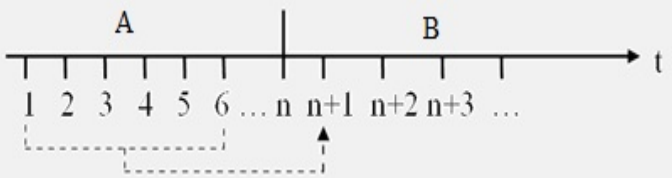

Рис. 1. Графическая иллюстрация постановки задачи прогнозирования: A – известные значения; B – прогнозируемый период

Предполагается, что реальная математическая модель прогнозируемого процесса имеет вид:

$$y_t = h(y_{t-1}, y_{t-2}, \ldots, y_{t-p}, \varepsilon_{t-1}, \ldots, \varepsilon_{t-q}) + \varepsilon_t$$

где h - некая неизвестная гладкая функция, ε_t – случайная составляющая. Также предполагается, что математическое ожидание $E(\varepsilon_t \mid y_{t-1}, y_{t-2}, \ldots, y_{t-p}) = 0$ и ε_t имеет некоторую конечную дисперсию σ^2. Очень важным является знание таких характеристик ε_t, как закон распределения, математическое ожидание, дисперсия.

2. Анализ существующих методов

Искусственные нейронные сети. Основной элемент нейронной сети – формальный (искусственный) нейрон. Он представляет собой математическую модель биологической нервной клетки.

ИНС представляют собой систему соединённых и взаимодействующих между собой искусственных нейронов.

ИНС не программируются в привычном смысле этого слова, они обучаются. В процессе обучения нейронная сеть способна выявлять сложные зависимости между входными и выходными данными, а также выполнять обобщение.

Способности нейронной сети к прогнозированию напрямую следуют из ее способности к обобщению и выделению скрытых зависимостей между входными и выходными данными. После обучения сеть способна предсказать будущее значение некоторой последовательности на основе нескольких предыдущих значений и/или каких-то существующих в настоящий момент факторов.

Главным преимуществом ИНС перед другими методами прогнозирования является то, что сети с одинаковым успехом могут прогнозировать процессы, регулярная составляющая которых имеет любой закон изменения значений, в

то время как большинство остальных методов лучше всего подходит для процессов, регулярная составляющая которых принадлежит к определенному классу (очевидно, что метод полиномиального сглаживания лучше всего подходит для процессов с полиномиальной регулярной составляющей, метод сглаживания рядами Фурье – для процессов с периодической регулярной составляющей и т.д.). Еще одним важным преимуществом нейронных сетей является возможность обучения.

Метод группового учета аргументов. Метод группового учета аргументов (МГУА) [2] - это набор алгоритмов прогнозирования (а точнее математического моделирования), который основывается на разбиении исходных данных на две выборки: обучающую и проверочную и использовании опорных функций некоторого вида, параметры которых находятся из обучающей выборки, а проверка того, насколько хорошо они моделируют заданный ряд, выполняется на проверочной выборке.

Комбинация МГУА и ИНС. Данный подход был предложен в [3]. Суть метода - использование подхода МГУА (разбиение выборки на подвыборки и постепенное усложнение модели), но вместо полиномиальных опорных функций используются ИНС с очень простой структурой. Это увеличило применимость метода, так как большинство прогнозируемых процессов нелинейные по своей природе, а также упрощает его использование и увеличивает автономность, так как не требует задания явного вида модели, а только ее входов.

Основные шаги данного метода.

1. Из данных вида $X = \{x_n, n = 1 \ldots N\}$, где N – количество отсчётов, формируются 2 исходные матрицы

$$X^{(0)} = \begin{bmatrix} x_1 & x_2 & \ldots & x_k \\ x_{k+2} & x_{k+3} & \ldots & x_{2*k+1} \\ \ldots & \ldots & \ldots & \ldots \\ x_{N-k-1} & x_{N-k} & \ldots & x_{N-1} \end{bmatrix}, y = \begin{bmatrix} x_{k+1} \\ x_{2*k+2} \\ \ldots \\ x_N \end{bmatrix},$$

где k – размер окна. С помощью этих матриц в дальнейшем будет осуществляться обучение сетей – каждый вектор-строка $\vec{x}_n = \{x_{n1}, x_{n2}, \ldots, x_{nk}\}, n = 1 \ldots m$ (m – количество строк матрицы) матрицы $X^{(0)}$ и соответствующее ему значение y_n являются обучающим примером, а вектор-столбец $\vec{x}_z = \{x_{z1}, x_{z2}, \ldots, x_{zk}\}, z = 1 \ldots k$ - отдельной переменной.

2. Определяется вид опорных функций, а именно от каких переменных они будут зависеть, например, $f_c = f(x_i, x_j)$ или $f_c = f(x_i, x_j, x_i * x_j, x_i^2, x_j^2)$, где $c = 1 \ldots C_k^2$ - количество вариантов выбора двух переменных из k возможных (в отличие от МГУА, где также определяется конкретный вид функций, например $f_ñ = a_0 + a_1 * x_i + a_2 * x_j$).

3. Составляются C_k^o (o – количество переменных, учитываемых в опорных функциях, для описанного выше примера o=2) многослойных персептронов (МП) с одним выходом, одним скрытым слоем с малым количеством нейронов (около 3), и нужным количеством входов (для опорных функций вида $f_c = f(x_i, x_j)$ сеть должна иметь 2 входа, для $f_c = f(x_i, x_j, x_i * x_j, x_i^2, x_j^2)$ - 5 входов и т.д.).
4. Каждый персептрон сопоставляется с конкретной опорной функцией, а именно, выбираются переменные, которые будут подаваться на входы сети (например, для опорных функций вида $f_c = f(x_i, x_j, x_i * x_j, x_i^2, x_j^2)$ один МП будет работать с переменными $x_1, x_2, x_1 * x_2, x_1^2, x_2^2$ а другой с переменными $x_1, x_k, x_1 * x_k, x_1^2, x_k^2$), и обучается, используя только примеры из обучающей выборки.
5. На этом шаге следует составить исходные данные для следующей итерации алгоритма. Для этого следует определить среднеквадратическую ошибку (СКО) каждого МП на проверочной выборке и отобрать k лучших сетей, после чего составить новую матрицу

$$X^{(1)} = \begin{bmatrix} h_{11} & h_{12} & \dots & h_{1k} \\ h_{21} & h_{22} & \dots & h_{2k} \\ \dots & \dots & \dots & \dots \\ h_{m1} & h_{m2} & \dots & h_{mk} \end{bmatrix},$$

где h_{ij} - значение выхода j-ой из k лучших сетей при подаче на её входы i-го примера, $i = 1 \dots m, j = 1 \dots k$ (либо исходная переменная).
6. Выполняется следующая итерация, но в качестве исходных примеров берется уже матрица $X^{(1)}$. Итерации выполняются, пока значение СКО сетей на проверочной выборке уменьшается, либо пока не будет достигнута требуемая СКО.

*Метод взвешенного скользящего среднего [1].*Метод основывается на простой модели, которая предполагает, что текущее значение $y_t, (t = 1 \dots n)$, ряда $\{y_1, \dots, y_n\}$ является взвешенной суммой некоторого количества предыдущих значений и некоторой случайной составляющей. Тогда каждому значению присваивается вес тем больший, чем более «свежее» значение

$$y_t = C + \varepsilon_t,$$

$$C = \frac{\sum_{i=1}^{p} \alpha_i * y_{t-i}}{\sum_{i=1}^{p} \alpha_i},$$

где α_i – соответствующие веса значений.

3. Метод решения задачи прогнозирования на основе комплексирования оценок

Для начала сформируем некоторый искусственный прогнозируемый процесс, имеющий математическую модель вида

$$f(t) = 0.1t^3 - 10\mathrm{t}^2 + 6 + 300\sin(\mathrm{t}), \mathrm{t} = 1\ldots100,$$

и добавим к математической модели нормальный шум $\varepsilon(\mathrm{t})$ с характеристиками $\mathrm{E}(\varepsilon) = 1000, \sigma(\varepsilon) = 500$:

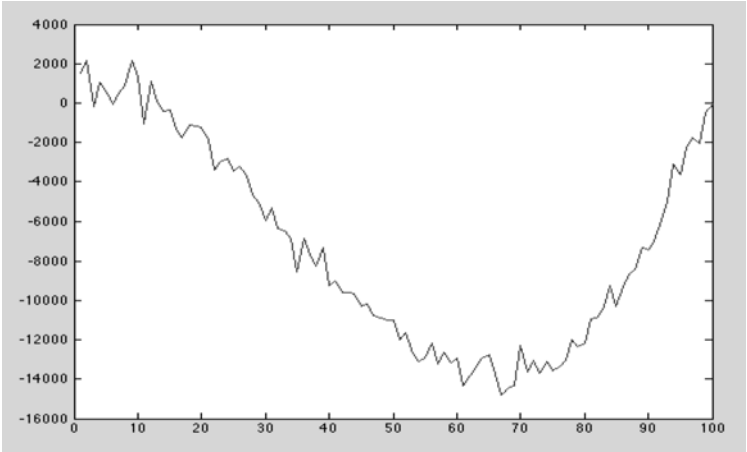

Рис. 2. График искусственно сгенерированного процесса

Попробуем идентифицировать истинную модель процесса с помощью нейронной сети. Для этого используем многослойный персептрон с одним скрытым слоем с 10 нейронами в нем, и с 5 нейронами во входном слое (активационная функция нейронов скрытого слоя - сигмоида, выходного слоя - линейная; метод обучения - Левенберга-Марквардта). После обучения ПСКО прогноза сети на всей выборке ($MSPE = \dfrac{\sum_i (\hat{\mathrm{y}}_i - \mathrm{y}_i)^2}{\sum_i (\mathrm{y}_i)^2}$) $MSPEnet = 0.0079$:

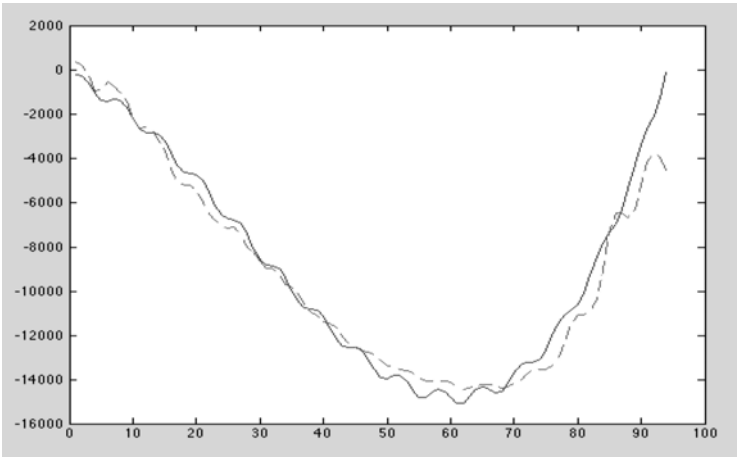

Рис. 3. Прогноз, полученный с помощью ИНС. ― ― ― - прогноз, ――― - истинная модель процесса

Как видим, точность полученной модели недостаточна. Попробуем предварительно обработать данные для улучшения точности получаемой модели. Для этого используем фильтр нижних частот. После применения фильтра и обучения новой сети, ПСКО новой сети стала $MSPEfiltered = 0.0038$ - уменьшилась в 2 раза!

Попробуем использовать метод группового учета аргументов вместо ИНС. Для построения модели будем использовать квадратичные опорные функции от 2-х аргументов. В результате имеем ПСКО полученной модели $MSPEgdmh = 0.0026$.

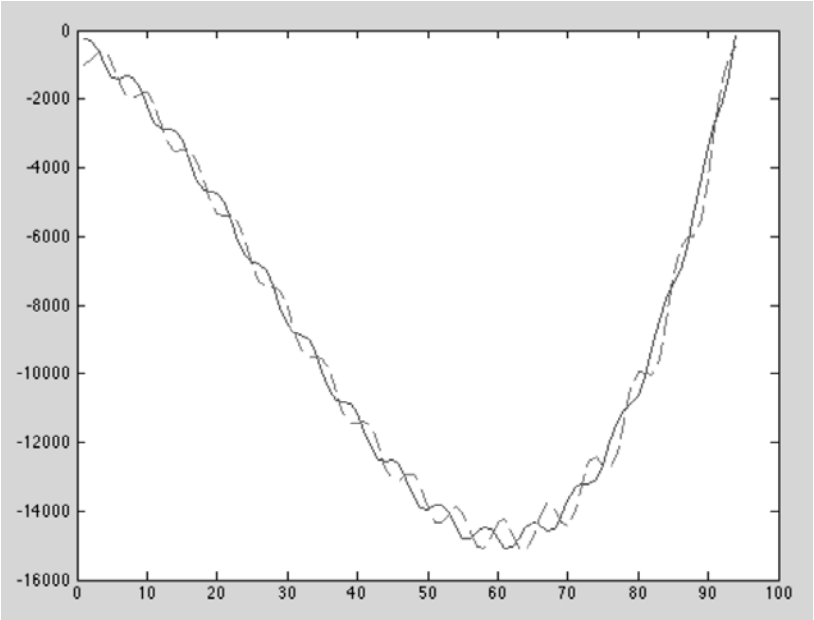

Рис. 4. Прогноз, полученный с помощью МГУА. ― ― ― - прогноз, ——— - истинная модель процесса

И, наконец, используем предложенный метод. Основная идея метода - комплексирование оценок, полученных с помощью различных моделей. Под комплексированием понимается взвешенная сума оценок, полученных с помощью сгенерированного набора моделей. Весовые коэффициенты определяются с помощью внешнего критерия оптимальности моделей - дисперсии на экзаменационной выборке. Множество моделей получается перебором вариантов разбиения исходной выборки на подвыборки и перебором выбранных методов прогнозирования. Таким образом, имея k_1 вариантов разбиения на подвыборки и k_2 методов прогнозирования, получаем k_1*k_2 различных моделей. Для получения окончательного прогноза, имея вектор входных данных $\vec{x}$ необходимо выполнить следующую последовательность действий.

1. Подать этот вектор на вход каждой модели, таким образом получив вектор оценок $\vec{\hat{y}} = \left[\hat{y}_1, \ldots, \hat{y}_{k_1*k_2}\right]^T$.

2. Получить окончательный прогноз $\hat{y}$ как взвешенную сумму элементов вектора оценок $\vec{\hat{y}}$.

Эти шаги требуют определения весовых коэффициентов $\alpha_i, i = 1 \ldots k_1 * k_2$.

Предложенный метод состоит из следующих этапов.

1. Предварительная обработка данных

Если известно, что прогнозируемый процесс не содержит информативных выбросов, то существует несколько основных алгоритмов избавления от случайных выбросов:

- Простейший алгоритм, основанный на характеристиках случайной величины, согласно которому выбросами считаются все значения, отклонённые от среднего на величину, большую, чем 2…3 среднеквадратических отклонений σ^2.
- Алгоритм Tukey 53H [7], который заключается в построении сглаженной последовательности, используя медианный фильтр и фильтр скользящего среднего, после чего выбросами будут считаться все исходные значения, отклонение которых от значений сглаженной последовательности больше чем некоторый наперед заданный порог k.
- Описанный выше фильтр нижних частот, в различных реализациях (например, фильтр Бесселя).

2. Разбиение исходной выборки на подвыборки

- *По порядку.* В обучающую выборку отбираются первые $C_1 * N$ точек, в проверочную - последующие $C_2 * N$ точек и в экзаменационную - оставшиеся точки, то-есть $(1 - C_1 - C_2) * N$ (где N - общее количество точек, $C_1 + C_2 < 1; C_1, C_2 > 0$ - коэффициенты, обычно выбирают $C_1 = 0.6, C_2 = 0.2$).
- *Случайным образом.* Аналогично предыдущему методу, только точки отбираются не по порядку, а случайным образом, но пропорции между подвыборками сохраняются.
- *Каждый i-й.* В проверочную выборку отбирается каждая i-я точка, из оставшихся точек в экзаменационную отбирается каждая j-я, все оставшиеся точки отбираются в обучающую выборку.
- *По дисперсии.* Все точки ранжируются по дисперсии (под точкой понимается один пример) и потом отбираются в выборки аналогично первому методу.

3. Получение моделей с помощью перебора методов

Для каждого из вышеперечисленных методов разбиения на выборки строится модель прогнозируемого процесса, используя методы прогнозирования: ИНС, МГУА, комбинация МГУА и ИНС. Таким образом, получаем 4*3=12 различных моделей.

4. Комплексирование полученных результатов

Для получения реального прогноза используется комплексирование прогнозов всех полученных на предыдущем этапе моделей. То есть, имея

оценки прогнозируемого значения $\vec{\hat{y}} = [\hat{y}_1, \ldots, \hat{y}_{12}]^T$, прогноз самого значения будет равен $\hat{y} = \sum_{i=1}^{12} \alpha_i * \hat{y}_i$. Коэффициенты α_i определяются, используя внешний критерий - дисперсию прогноза i-ой модели на экзаменационной выборке: $\alpha_i = \frac{1}{\sigma_i}$, где σ_i - СКО прогнозов i-ой модели на экзаменационной выборке от реальных значений прогнозируемого процесса. После вычисления всех α_i их следует нормализовать по формуле $\alpha_i^n = \frac{\alpha_i}{\sum_{i=1}^{12} \alpha_i}$.

Используя предложенный метод достигается ПСКО равная $MSPEbest = 0.0020$:

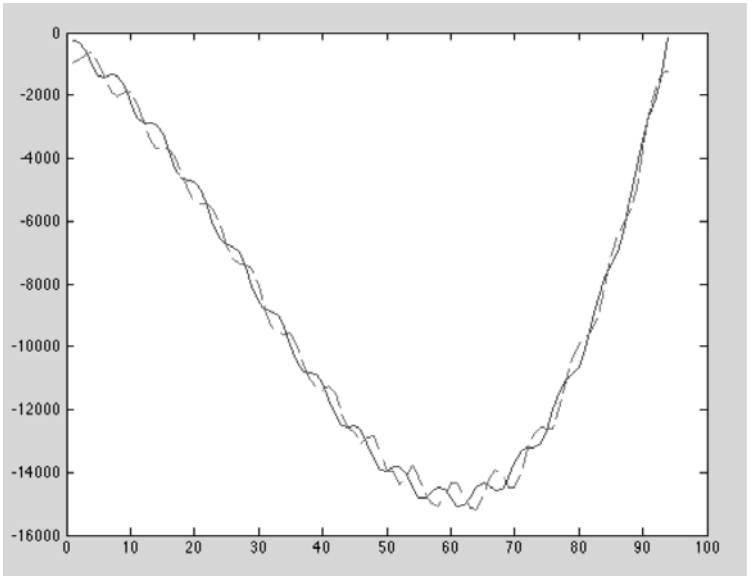

Рис. 5. Прогноз, полученный с помощью предложенного метода. ——— - прогноз, ——— - истинная модель процесса

4. Практическое применение

Для проверки предложенного метода были использованы публично доступные данные по годовым ценам на медь в период с 1800 по 1907 г. Также, для оценки работы метода был построен и обучен многослойный перцептрон с 2 скрытыми слоями (5 и 8 нейронов в 1 и 2 скрытом слое соответственно) и 5 входами.

Сравнение результатов, полученных при использовании предложенного метода и ИНС показаны на Рис. 6.

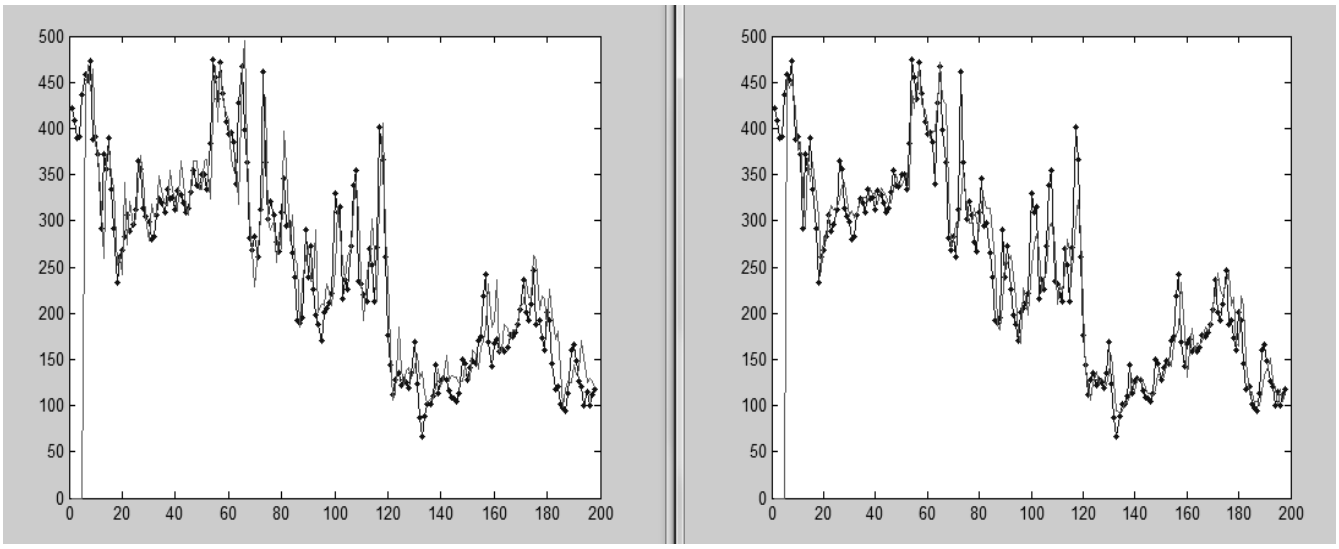

Рис. 6. Сравнение прогнозов ИНС (слева) и предложенного метода (справа). —◆— - истинные значения процесса, —— - прогноз

5. Заключение

В данной статье предложен метод, использующий комплексирование оценок моделей, построенных другими методами. Среди рассмотренных наиболее перспективными методами являются ИНС (изза их универсальности и возможности обучаться), МГУА (этот метод позволяет получить точный прогноз даже при малой размерности выборки исходных данных и ее сильной зашумленности) и их комбинация. Поэтому именно эти методы использовались для генерации моделей.

Предложенный алгоритм показал лучшие результаты по сравнению с методами МГУА, ИНС и их комбинация. Это говорит о возможности его дальнейшего использования для решения задач прогнозирования. Для дальнейшего развития алгоритма следует рассмотреть использование других внешних критериев для определения весовых коэффициентов.

Основными преимуществами алгоритма являются:

- более точная модель прогнозируемого процесса (с точки зрения минимума критерия СКО на всей выборке);

- как результат использования МГУА и ИНС в качестве метода построения промежуточных моделей предложенный метод справляется с выборками малых размеров;

- вследствие применения предварительной обработки данных предложенный метод является менее чувствительным к выбросам в данных;

Недостатки:

- так как используются сразу несколько методов прогнозирования, общее время построения модели с помощью предложенного метода может быть сравнительно большим. По этой же причине для программной реализации метода требуются сравнительно большие вычислительные мощности.

Литература

1. Алесинская Т.В. Методы скользяшего среднего и экспоненциального сглаживания / Т.В. Алесинская //Экономико-математические методы и модели : уч. пособие по решению задач по курсу. – Таганрог : Изд-во ТРТУ, 2002. – 153 с.

2. Івахненко А.Г. Метод групового урахування аргументів – конкурент методу стохастичної апроксимації / А.Г. Ивахненко // Автоматика. – 1968. – № 3. – С. 58-72.

3. Мак-Каллок У.С. Логическое исчисление идей, относящихся к нервной активности / У.С. Мак- Каллок, В.Питтс // Автоматы ; [под ред. К.Э. Шеннона и Дж. Маккарти]. – М. : Изд-во иностр. лит., 1956. – С. 363-384.

4. Amir F.A. A comparison between neural-network forecasting techniques – case study: river flow forecasting /F.A. Amir, I.S. Samir // IEEE Transactions on neural networks. – 1999, march. – Vol. 10, № 2. – С. 402-409.

5. Mohsen H. Artificial neural network approach for short term load forecasting for Illam region / H. Mohsen, S. Yazdan // World Academy of Science, Engineering and Technology 28 2007. – С. 280-284.

6. Jerome T.C. neural networks and robust time series prediction / T.C. Jerome, R.M. Douglas, L.E. Atlas // IEEE transactions on neural networks. –1994, march. – Vol. 5, № 2. – С. 240-254.

7. Klevecka Irina. Pre-Processing of Input Data of Neural Networks: The Case of Forecasting Telecommunication Network Traffic / Irina Klevecka, Janis // Telektronikk 3/4.2008. – С. 168-178.

8. http://robjhyndman.com/TSDL/micro-economic/.