УДК 681.513.2, 519.246.8

ПРИМЕНЕНИЕ ДОВЕРИТЕЛЬНЫХ ИНТЕРВАЛОВ В ЗАДАЧАХ ПРОГНОЗИРОВАНИЯ ФИНАНСОВЫХ РЯДОВ С ПОМОЩЬЮ МГУА

Поднебесная Г.А[1], Поднебесная А.А.[2]

[1] *МНУЦ ИТС АН и МОН Украины,*

[2] *Киевский Национальный университет им. Т.Г.Шевченко*

pidnebesna@irtc.org.ua

Представлена методика аналізу ефективності роботи прогнозуючої системи за допомогою побудови довірчих інтервалів, вводяться основні поняття та наведено приклад застосування цієї методики для конкретної прогнозуючої системи, побудованої за МГУА. За вхідні дані взято реальний часовий ряд котирування валют ринку Форекс за період з 1 січня по 30 червня 2010 року.

Ключові слова: прогнозуюча система, довірчий інтервал, помилка прогнозу, прогнозування часових рядів, МГУА.

A method for analysis of work of a forecasting system by construction of confidence intervals is presented, basic concepts are entered and a example of application of the given method for the concrete forecasting system based on GMDH is given. As input data the real time series of quotations of currencies of the Foreks market is taken for a period from January, 1 to June, 30, 2010.

Keywords: the forecasting system, confidence intervals, error of prediction, prediction of time series, GMDH

Представлена методика анализа эффективности работы прогнозирующей системы с помощью построения доверительных интервалов, вводятся основные понятия и приведен пример применения данной методики для конкретной прогнозирующей системы, основанной на МГУА. В качестве входных данных взят реальный временной ряд котировок валют рынка Форекс за период с 1 января по 30 июня 2010 года.

Ключевые слова: прогнозирующая система, доверительный интервал, ошибка прогноза, прогнозирование временных рядов, МГУА.

Введение. При применении любой прогнозирующей системы естественно возникает вопрос: насколько хорошо она работает. Для ответа необходимо ввести некоторые количественные характеристики полученного прогноза. Существует ряд показателей, которые могут использоваться для оценки эффективности прогнозирующей системы. В зависимости от поставленных задач они могут быть самыми разными. Одним из видов подобной характеристики является изучение доверительных интервалов, т.е. меры того, насколько далеко отклоняется спрогнозированное значение от значения реального.

В данной статье представлена методика анализа оценки работы прогнозирующей системы с помощью построения доверительных интервалов. Вводятся основные понятия и приведен пример применения данной методики для конкретной прогнозирующей системы, основанной на МГУА [1, 2].

В качестве данных взят реальный финансовый ряд котировок валют рынка Форекс за период с 1 января по 30 июня 2010 года.

Методика оценивания прогноза с помощью доверительных интервалов

Пусть зафиксированы некоторые обобщенные данные в виде временного ряда, статистические характеристики которого не меняются.

Вместе с данными рассмотрим некоторую обобщенную прогнозирующую систему, получающую на вход данные и выдающую на выходе набор прогнозных значений в заданных точках. Для определенности введем следующие обозначения (см. рис.1):

- M – последовательность данных (статистическая выборка), используемая для прогнозирования;
- $T_1, T_2, ..., T_i, ..., T_n$ - координаты прогнозируемых точек, $\{1 \le i \le n\}$;
- y_{F_i} – значение спрогнозированных данных, $\{1 \le i \le n\}$;
- y_{R_i} – значение реальных данных, $\{1 \le i \le n\}$;
- y_{R_0} – последняя точка входных данных (множества M), являющаяся одновременно начальным уровнем, относительно которого будет рассматриваться прогноз.

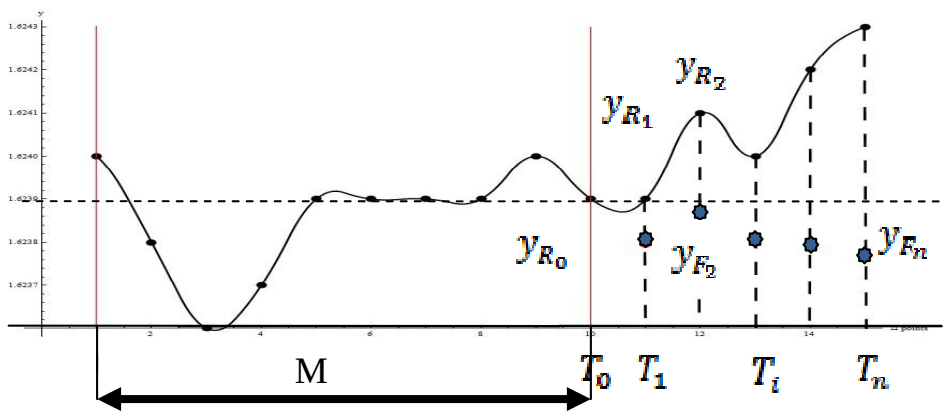

Рис. 1. Обобщенная прогнозирующая система

Отметим, что нельзя говорить в общем случае о точности работы прогнозирующей системы. Это имеет смысл, только если рассматривать прогнозирующую систему и конкретный вид данных, для которых строится прогноз, используя эту систему. Только в этом случае можно говорить о каких-либо качественных и количественных характеристиках прогноза.

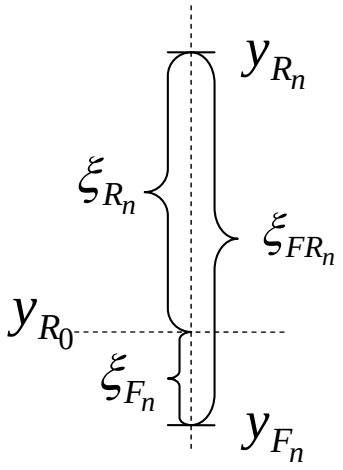

Рис. 2. Исследуемые случайные величины

Для того, чтобы сделать некоторые выводы и охарактеризовать работу прогнозирующей системы на указанных данных, введем несколько случайных величин (рис. 2) и исследуем их поведение. Обозначим:

- разность между прогнозным и реальным значением (ошибка прогноза):

$$\xi_{FR_i} = y_{F_i} - y_{R_i}; \qquad (1)$$

- отклонение прогнозного значения от начального уровня

$$\xi_{F_i} = y_{F_i} - y_{R_0}; \qquad (2)$$

- отклонение реальных данных от начального уровня:

$$\xi_{R_i} = y_{R_i} - y_{R_0},\ i \in \{0,...,n\}. \qquad (3)$$

Искомыми характеристиками прогнозирующей системы будут доверительные интервалы этих случайных величин.

Действительно, говоря о прогнозном значении, нелишним будет указать, насколько оно удалено от реального и с какой вероятностью (доверительный интервал ошибки прогноза). Еще две случайные величины (отклонения от начального уровня) введены для сравнения: насколько стоит использовать данную прогнозирующую систему на этих данных или результат будет лучше, если в качестве прогноза брать случайную точку с распределением случайной величины ξ_{R_i}.

Одним *испытанием* будем считать следующее: выбрали множество M заданной длины, построили прогноз для точки, находящейся на заданном расстоянии, посчитали для нее три значения (1) - (3).

Таким образом, проведя достаточное количество испытаний, мы будем иметь набор значений для определенных выше случайных величин (1) - (3), по которым можно будет делать выводы о доверительных интервалах.

Следует учесть, что говорить о доверительных интервалах и других вероятностных характеристиках (стандартное отклонение, среднее), можно лишь при условии, что статистические характеристики ряда данных не изменяются на исследуемом временном промежутке, т.е. что распределения исследуемых величин не зависят от того, какую часть данных взяли для анализа.

При расчетах использовались следующие статистические понятия.

Математическое ожидание случайной величины оценивается как выборочное среднее:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{4}$$

Дисперсия случайной величины оценивается как выборочная дисперсия:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{5}$$

Под «*доверительным интервалом*» понимается следующее [3].
Обозначим истинное значение измеряемой величины x, ее полученное среднее арифметическое значение - $\bar{x}$, а погрешность измерения этой величины Δx. Пусть α означает вероятность того, что результат измерений отличается от истинного значения на величину, не большую чем Δx, т.е.

$$P\{\bar{x} - \Delta x < x < \bar{x} + \Delta x\} = \alpha \tag{6}$$

Вероятность α - *доверительная вероятность* (интервал надежности).
Интервал $(\bar{x} - \Delta x, \bar{x} + \Delta x)$ – *доверительный интервал*.

Для нормально распределенной случайной величины выполняется правило «трех сигм» [3], которое дает возможность легко построить доверительный интервал с определенной доверительной вероятностью:
для трех значений доверительных вероятностей истинное значение величины отличается от своего среднего на известную величину:

- стандартному отклонению $\sigma = \sqrt{S^2}$ соответствует доверительная вероятность 0.68,
- удвоенному стандартному отклонению (2σ) – доверительная вероятность – 0.95,
- утроенному (3σ) – 0.997.

Пример применения метода на реальных данных

Данные. Рассматривался реальный временной ряд котировок валюты GRBCHF (отношение цены британского фунта стерлингов к швейцарскому франку, в поинтах) рынка Форекс за период с января 2010 года по июнь 2010 года включительно [4].

В архиве данные содержатся в следующем формате:
<TICKER><DTYYYYMMDD><TIME><OPEN><HIGH><LOW><CLOSE>,
где :

- TICKER – код валюты,
- DTYYYYMMDD – дата в формате год.месяц.день (например, 20100325),
- TIME –номер минуты с момента начала дня,
- OPEN – «цена открытия», т.е. первое значение цены за текущую минуту,
- HIGH – максимальное значение, которого достигает цена за текущую минуту,

- LOW – минимальное значение, которого достигает цена за текущую минуту,
- CLOSE – «цена закрытия», т.е. последнее полученное значение цены за текущую минуту.

Итак, имеем для каждой минуты набор из четырех значений (цен). Обрабатывая их, можно получить дополнительную информацию о поведении данных. Мы же будем рассматривать «сглаженные» данные, а именно:

$$mean\_HL = \frac{< HIGH > + < LOW >}{2}$$

Использование таких усредненных данных дает возможность несколько сгладить стохастический шум реальных данных, а значит получить, возможно, более точный прогноз.

Следует обращать внимание при анализе на то, что для корректности использования статистических методов выборка данных должна быть независимой, а это достигается, если интервалы, по которым строятся каждый раз прогнозы, не пересекаются.

Прогнозирующая система. Основой прогнозирующей системы выступает комбинаторный алгоритм МГУА [1, 2]. В качестве моделей использовались полиномы степени от 0 до 10, коэффициенты которых рассчитывались с помощью МНК на обучающей выборке. Лучшие модели выбирались согласно критерию регулярности (таб.1).

Исследовалось несколько способов разбиения выборки данных M на «обучающую» M_A и «проверочную» M_B, $M = M_A \cup M_B$. Было проведено четыре эксперимента: рассматривались разбиения с различным количеством точек в подвыборках M_A и M_B: [5, 5],[5, 20],[20, 5],[20, 20]. Прогноз строился на 20 точек. Для каждой точки прогноза рассчитывались значения (1) - (3).

Интервал временного ряда, на котором проводились расчеты: с 10 по 182760 минуты. При каждом следующем испытании начало интервала М сдвигалось на 40 точек (минут) вперед относительно предыдущего. Будем называть это расстояние «шагом». Такая длина шага позволила использовать непересекающиеся интервалы, то есть получать независимые выборки. Получено порядка 4500 измерений для каждого эксперимента, по которым строились и анализировались доверительные интервалы (6).

Таблица 1

Параметры эксперимента

Алгоритм МГУА	COMBI			
Степень полинома	10			
Количество точек обучающей выборки M_A	5		20	
Количество точек проверочной выборки M_B	5	20	5	20
Количество точек прогноза	20			
Шаг	40			

В результате проведенных экспериментов для каждой случайной величины (1) - (3) была набрана достаточная статистика для того, чтобы оценить среднее, дисперсию и доверительные интервалы, сделать выводы о распределении. Следует обратить внимание, что в данной работе оценка доверительных интервалов производилась по определению (6), правило же «трех сигм» не использовалось.

Прогнозирующая система и программа для ее анализа раелизованы с использованием средств СКМ Mathematica 7.

Описание полученных результатов

1. Проведя расчеты для различных наборов параметров, мы получили, что *распределения случайных величин (1) - (3) отличаются от нормального*.

Для демонстрации приведены распределения ошибки прогноза для точек 1, 2 (параметры эксперимента: полиномиальный комбинаторный алгоритм МГУА, максимальная степень полинома 10, длина обучающей выборки 20, длина проверочной выборки 20). Рассмотрим гистограмму распределения для ошибки прогноза для этих точек (рис.3 – рис.6).

На графике на соответствующую гистограмму наложен график плотности нормального распределения со средним 0 и стандартным отклонением, оцененным по выборке. На левом рисунке приведена «полная» гистограмма, на правом – часть ее на доверительном интервале эмпирической плотности уровня 90%.

***Точка 1*:** (рис.3, 4)

Стандартное отклонение: 0.0340472

Доверительный интервал:

$\{-0.0346342\ \sigma,\ 0.0353187\ \sigma\} = \{-0.001179198, 0.001202503\}$.

Доверительная вероятность: 0.899803.

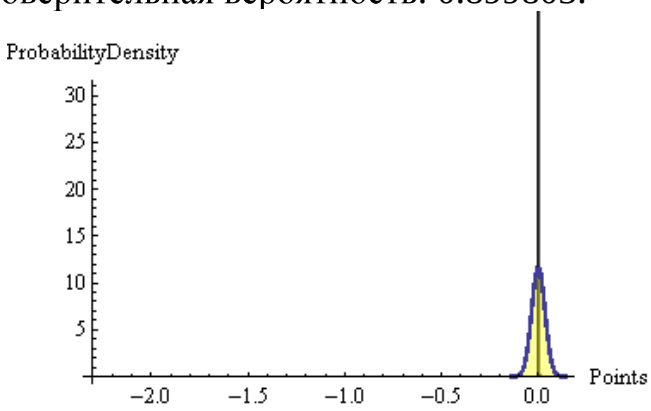

Рис. 3

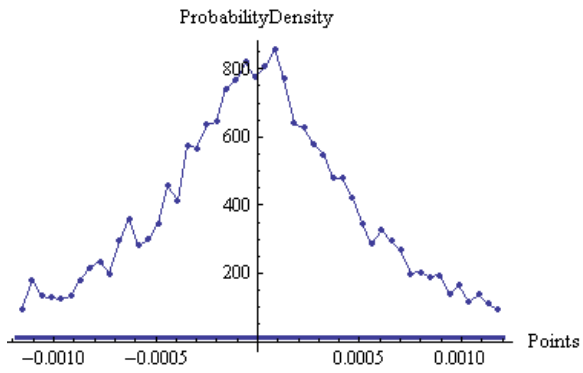

Рис. 4

***Точка 2*:** (рис.5, 6)

Стандартное отклонение: 0.0340728

Доверительный интервал:

$\{-0.0418866\ \sigma,\ 0.0401116\ \sigma\} = \{-0.001427194, 0.001366715\}$.

Доверительная вероятность: 0.899803.

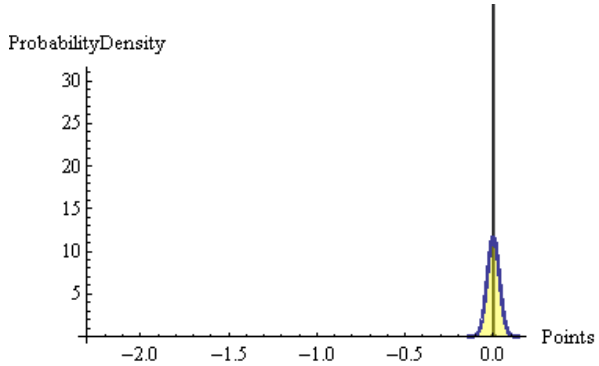

Рис. 5

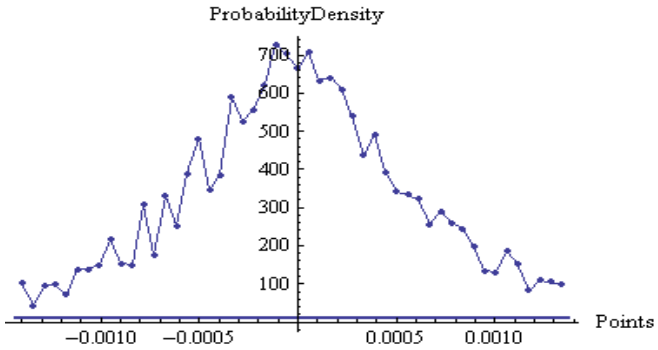

Рис. 6

Используя критерий Колмогорова [5] для всех рассмотренных точек, получили, что эмпирическая функция распределения не совпадает с нормальной.

Полученные результаты демонстрируют, что нельзя строить доверительный интервал, основываясь только лишь на среднем и стандартном отклонении, т.к. распределение получается не нормальным, и правило «трех сигм» не выполняется. Конкретнее, если пользоваться этим правилом, интервал получится в несколько раз больше доверительного интервала для случайной величины (3).

2. Далее рассмотрим, есть ли зависимость величины доверительного интервала от отношения длин обучающей и проверочной подвыборок.

На рис.7 и в таб.2 приведены график зависимости и значения величины доверительного интервала для комбинаторного алгоритма МГУА с полиномиальными базисными функциями (максимальная степень полинома 10) при разбиениях M_A, M_B: [5, 5], [5, 20], [20,5], [20,20].

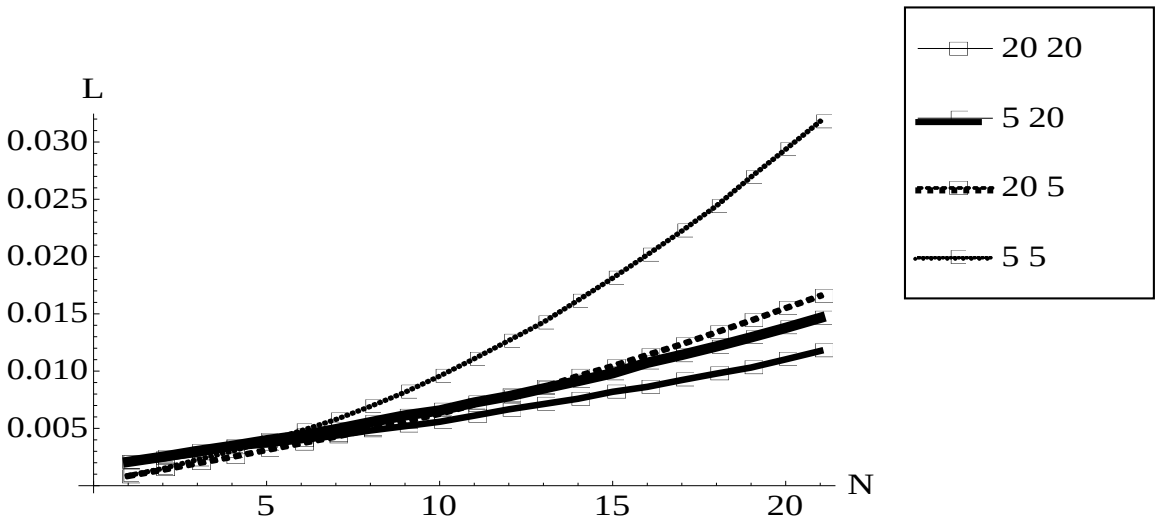

Рис.7. Зависимость величины доверительного интервала
от разбиения выборки для полиномиального базиса
(L – величина доверительного интервала, N - номер точки прогноза)

Таблица 2

Величины доверительных интервалов при различных разбиениях выборки для полиномиальных базисных функций

#	$M_A = 5,\ M_B = 5$	$M_A = 5,\ M_B = 20$	$M_A = 20,\ M_B = 5$	$M_A = 20,\ M_B = 20$
0	0,000885669	0,00208235	0,00084052	0,00201836
1	0,001494123	0,00252117	0,00137898	0,0023817
2	0,0022708	0,00302636	0,00195163	0,00279391
3	0,00301269	0,00346929	0,00250278	0,00333075
4	0,00386546	0,00396317	0,00310788	0,0036126
5	0,00480465	0,00439052	0,00369184	0,00396832
6	0,00577595	0,00493182	0,00427281	0,00436778
7	0,00691998	0,00553516	0,00493293	0,00481762
8	0,00815849	0,00613451	0,00559532	0,00519446
9	0,00956712	0,0065634	0,00626425	0,00557669
10	0,01107849	0,00725642	0,007084	0,00611221
11	0,01266981	0,00780646	0,00786414	0,00665994
12	0,01426864	0,00847944	0,00859973	0,00712402
13	0,01618347	0,00913407	0,00956592	0,00759082
14	0,01811904	0,00981696	0,01043914	0,0081979
15	0,02015171	0,01072714	0,01140973	0,00862506
16	0,0222332	0,01141878	0,01234685	0,00921863
17	0,0244251	0,01215571	0,01339252	0,00978103
18	0,0269513	0,01294645	0,01441778	0,01031416
19	0,0293566	0,01377719	0,01549584	0,01102643
20	0,0318137	0,01464555	0,01657087	0,01175834

Приведенные результаты позволяют предположить, что длина выборки $M = M_A \cup M_B$ не должна быть меньше количества точек прогноза.

Если количество точек, по которым строится прогноз, больше количества точек самого прогноза, то изменение отношения длин обучающей и проверочной подвыборок практически не влияют на длину доверительного интервала, т.е. корректность работы данной прогнозирующей системы от последовательного разбиения выборки практически не зависит.

Далее на рисунке 8 приведены графики доверительных интервалов при использовании полиномиального базиса для прогнозирования случайных величин (1) – (3).

На графике обозначены:

1 (тонкая линия) – ошибка прогноза,

2 (пунктирная линия) – отклонение реальных данных от значения в точке начала прогноза,

3 (жирная линия) – отклонение прогнозных значений от значения в точке начала прогноза.

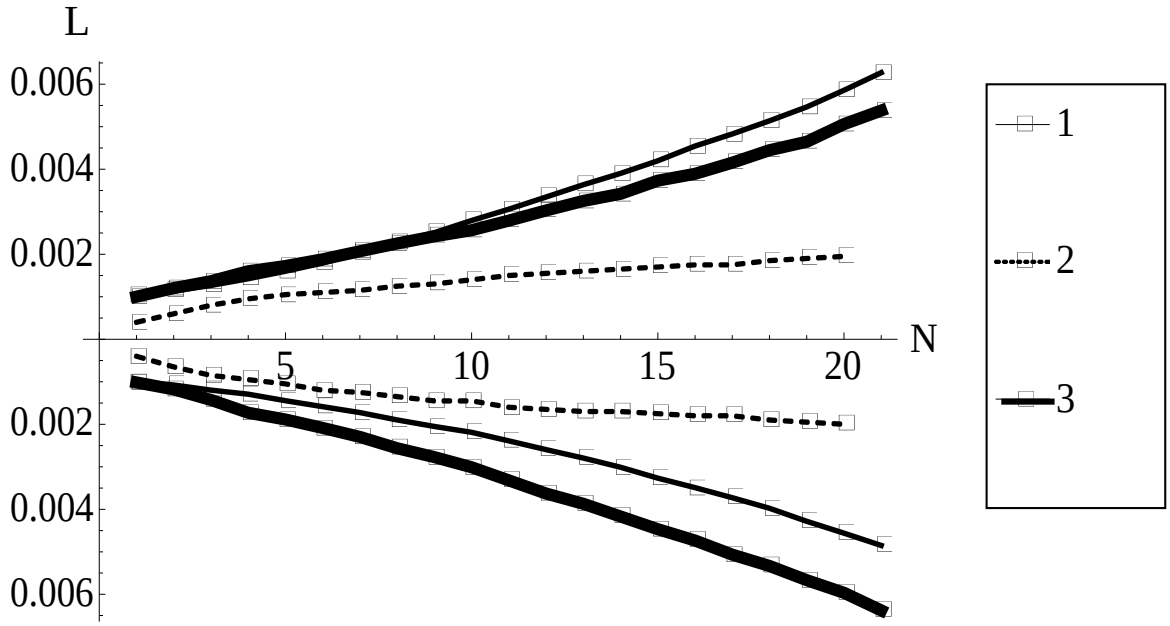

Рис. 8. Доверительные интервалы для исследуемых случайных величин (L – величина доверительного интервала, N - номер точки прогноза)

Анализируя полученные данные, можно заметить, что с увеличением интервала прогнозирования при применении полиномиального базиса доверительный интервал ошибки прогноза растет быстрее, чем соответствующий доверительный интервал для реальных данных, т.е. всегда остается больше.

Выводы и дальнейшее направление исследований

В работе описана методика анализа прогнозирующей системы с помощью построения доверительных интервалов. В качестве примера взяты: прогнозирующая система, основанная на использовании комбинаторного алгоритма МГУА с полиномиальным базисом, и реальные финансовые данные котировок валют рынка Форекс. В результате проведенных исследований можно сделать следующие выводы.

- Исследования показали, что для прогнозирования временных рядов данных рынка Форекс при использовании комбинаторного алгоритма МГУА при построении доверительного интервала нельзя пользоваться правилом «трех сигм», поскольку распределение случайных величин (1) - (3) не является нормальным.
- Величина доверительного интервала при использовании для моделирования алгоритма МГУА с полиномиальным базисом практически не зависит от соотношения длин обучающей и проверочной выборок. Количество точек для обучения не должно быть меньше количества точек прогноза.

Дальнейшим направлением исследований является применение доверительных интервалов для анализа и сравнения прогнозирующих систем, основанных на МГУА, а именно:

- изучение зависимости длины доверительного интервала от параметров прогнозирующей системы, например, от выбора базисных функций при моделировании, в частности, использование гармонического базиса;
- сравнение результатов при использовании полиномиального и гармонического базиса в комбинаторном алгоритме ;
- разработка ограничений для полиномиального алгоритма МГУА, позволяющих уменьшить количество «выбросов» прогнозируемых значений, и тем самым уменьшить стандартное отклонение;
- прогнозирование «направления движения» котировок в финансовых рядах в сторону увеличения или уменьшения.

Литература

1. Ивахненко А.Г., Степашко В.С. Помехоустойчивость моделирования. – К.: Наукова думка, 1984. - 295 с.
2. Ивахненко А.Г., Юрачковский Ю.П. Моделирование сложных систем по экспериментальным данным. - М: Радио и связь, 1987. - 120 с.
3. Зайдель А.Н. Элементарные оценки ошибок измерений. – Л.: «Наука», 1967. - 88 с.
4. http://www.forexite.com/free_forex_quotes/forex_history_arhiv.html
5. Карташов М.В. Імовірність, процеси, статистика. – К., ВПЦ «Київський університет», 2008. – 504 с.