УДК 517.534

Б.В. Шамша, Т.Б. Шатовская, А.Ю. Гуд

Локальные модели прогнозирования нелинейных временных рядов в условиях гетероскедастичности

Харьковский национальный университет радиоэлектроники

Предложено разбить исходную выборку на классы с помощью кластерного анализа. Текущее измерение отнесено к тому или иному классу на основе дискриминантного анализа. Модель прогнозирования построена в каждом классе с учетом риска, полученного по оценке волатильности стохастических временных рядов с коррекцией ошибки, условно полученной по текущему и прошлому информационному множеству.
Ключевые слова: гетероскедастичность, кластеризация, дискриминация, волатильность, риск, авторегрессия, GARCH модель, временной ряд.

Многие предметные области экономики, техники, социологии, медицины описываются нелинейными моделями, которые характеризуются резкими скачками, гистерезисом, детерминированным хаосом, наличием гетероскедастичности ошибок.

В качестве примера в работе приведен нелинейный динамический ряд, характеризующий курс валют EUR/USD за период с 22.08.2003 по 12.09.2003 (с шагом 15 минут), показанный на рис. 1.

В ряде просматривается стохастическая волатильность, изменчивость статистических характеристик, в частности дисперсии, которые меняются в зависимости от времени, неконтролируемых возмущений, «тяжелых хвостов» в распределениях, глубины погружения, кластеризации экстремумов, фрактальности и других факторов. В некоторые периоды наблюдается изменчивость в широком и в узком смысле стационарности характеристик. Как показано в работах [1] и [2], нелинейные системы характеризуются общими тенденциями в функционировании и описании. Наблюдается связь между теорией хаоса и нелинейными временными рядами, которая проявляется на основе нелинейной динамики. Эта взаимосвязь позволяет подойти с новых позиций к анализу и идентификации моделей прогнозирования на основе локализации временного ряда с помощью кластерного и дискриминантного анализов.

Исходный временной ряд разобьём на классы и построим модели прогнозирования в каждом из них. Текущую информацию отнесём к тому или иному классу с помощью дискриминантного анализа.

В работе предложено разбить временной ряд на классы методами кластерного анализа. Модель прогнозирования построена в каждом классе с учётом риска, полученного из оценки волатильности стохастических временных рядов с коррекцией ошибки, условно полученной по текущему и прошлому информационному множеству. Также предложено локальную модель прогноза подвергнуть комплексной диагностической проверке на основе анализа автокорреляционных функций, кумулятивной периодограммы остаточных ошибок и доверительных интервалов прогнозных значений. Заметим, что разработать обобщенную диагностическую проверку для всех случаев не представляется возможным, так как исходные данные имеют специфические особенности для данного момента времени, учесть которые не всегда удается. В работе основное внимание уделяется временным рядам, у которых динамические параметры, характеризующие состояние объекта управления, являются нестационарными и неоднородными. Однако

временной ряд можно привести к однородному с помощью моделирования тренда и вычитания его из исходного ряда.

Для проведения кластеризации исходных данных и в будущем дискриминации текущих значений для отнесения их к тому или иному классу, разобьем исходный ряд на несколько типовых состояний, которые будут характеризоваться качественными или количественными показателями по следующему правилу.

Представим одномерный временной ряд в виде матрицы состояний в многомерном пространстве следующим образом. Пусть временной ряд состоит из N точек. В интервале от 0 до N выберем «окно», которое содержит определенное число данных, допустим, m, причем $m<int(N/2)$, где int – целая часть от деления. Первая строка матрицы X будет состоять из данных от 1 до m. Вторая строка матрицы будет начинаться со второго значения временного ряда до $m+1$. Продолжать такую процедуру необходимо, пока в последней строчке не окажутся значения от m до N. Матрица X будет содержать m строк и N-$m+1$ столбцов. Количество данных временного ряда составит $m*(N-m+1)$.

Для проведения кластеризации необходимо решить следующие проблемы:

1) определить первоначальное количество окон, величину m в каждом окне;

2) выбрать количество классов и наиболее информативные признаки в каждом из них;

3) выбрать и оценить меры близости объектов внутри классов и между классами;

4) определить критические значения признаков для перехода из одного класса в другой;

5) последовательной сдвижкой окон на одно измерение скорректировать количество классов с учетом получения однородной выборки;

6) в скорректированных классах построить модели прогнозирования с учетом наличия гетероскедастичности ошибок и сдвижки окон;

7) оценить адекватность моделей;

8) при необходимости изменить величину окон и повторить процедуры от 1 до 7;

9) построить комплекс дискриминантных функций и отнести текущее измерение к тому или иному классу.

Показать решение всех проблем в данной статье не представляется возможным. Основное внимание будет обращено на решение проблем 1, 2, 7.

Проблема выбора параметров состоит в изначальном выборе длины ряда N и длины окна m.

Формализованных рекомендаций решения данной проблемы не существует. При геометрической интерпретации параметр m является размерностью пространства, в котором исследуется траектория многомерной ломаной линии, в которую переводится исходный временной ряд. В общем случае выбор m зависит от постановки задачи. В нашем случае задачу кластеризации исходного временного ряда необходимо проводить с учетом скрытых периодичностей с неизвестными периодами. Здесь используют следующий подход. Сначала вычисляют собственные числа при максимально возможном m и определяют l. Затем проводят повторные расчеты с m несколько большим, чем l. Большие возможности решения данной задачи возлагаются на метод имитационного моделирования.

Определим количество классов равное количеству окон.

Рассмотрим вкратце принципы выбора метрики для оценки сходства.

Оценка близости кластеров существенно зависит от выбора способа нормировки данных (приведение к единому масштабу) и определения меры близости между объектами. Заметим, что измерение близости становится оправданным при одинаковом масштабе исходных данных. Существует множество способов нормировки [1], которые зависят от использования тех или иных видов статистических характеристик (размах, среднее, дисперсия, среднеквадратическое отклонение, эталонные значения исходных данных).

Многие нормировки искажают физический смысл данных.

Исследуем технологию выбора номинальных шкал или мер сходства, к которым отнесем меры близости Рао, Хемминга, Роджерса-Танимото, Жаккарда [4, 5]. Выбор конкретного измерителя должен осуществляться, прежде всего, из содержательных соображений: если предполагается равная значимость совпадения единичных и нулевых свойств, то следует применять расстояние Хемминга; если важно только наличие свойства, а не его отсутствие, — использовать коэффициенты Рао или Роджерса-Танимото.

Проведем анализ расстояния Махолонобиса, которое является универсальной метрикой в виде

$$d_M(y_i, y_j) = [(Y_i - Y_j)R^{-1}(Y_i - Y_j)^T]^{1/2} \quad (1)$$

где Yi, Yj – вектор-столбцы значений всех признаков на *i*-м и *j*-м объектах;

R^{-1} – обратная ковариационная матрица исходных данных.

Степень корреляционной связи между признаками определяет степень вырожденности ковариационной матрицы. Если корреляции близки к единичным, и дисперсии почти равны друг другу, определитель ковариационной матрицы приближается к нулю. Обратная матрица R^{-1} приобретает крайне неустойчивый вид, что приводит к большим колебаниям значений $dM(yi,yj)$. Наличие аномальных наблюдений приводит к искажению матрицы расстояний, так как матрица R^{-1} зависит от расстояния между другими признаками.

Итак, выбор меры близости кластеров существенно зависит от структуры временного ряда.

Множество исходных данных было разбито на пять групп. График исходных данных с типовыми классами изображён на рис. 1.

В качестве признаков классификации определим следующие критерии:

1) среднее значение цепного темпа роста

$$Y = (\Delta Y_1 + \Delta Y_2 + ... + \Delta Y_n)/n, \quad (2)$$

где ΔY_1 – цепной темп роста;

n – количество данных в подмножестве;

2) среднее значение размаха

$$R = (Y_{max} - Y_{min})/n, \quad (3)$$

где Y_{max} – максимальное значение в подмножестве данных;

Y_{min} – минимальное значение в подмножестве данных;

3) D – дисперсия.

Проведем исследования влияния различных факторов кластеризации на эффективность моделей прогнозирования.

Исходные данные разбили на группы и для каждой группы рассчитали значение критериев классификации; каждая из групп была отнесена в соответствии

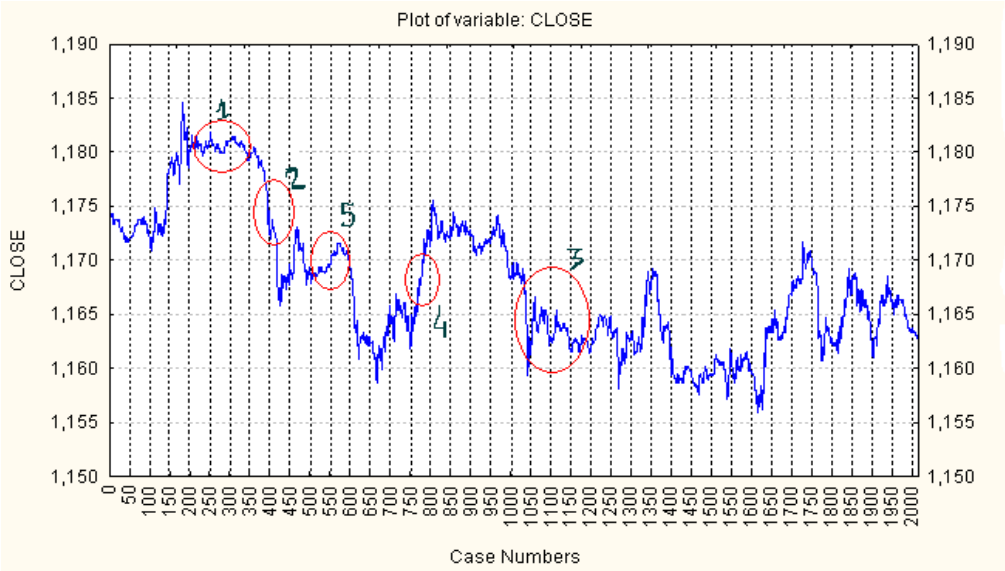

Рис. 1. Пример подмножеств данных, входящих в каждую из групп

со значениями критериев к одному из пяти классов. В каждом классе построили модели прогнозирования.

Графический и аналитический анализы показывают, что локальные временные ряды представляют собой неоднородную нестационарную последовательность, для которой характерен блуждающий характер структуры (тренды и циклы могут в будущем меняться или исчезать).

Интерес представляет оценка меры изменчивости условной дисперсии. Степень изменчивости переменной называют волатильностью (изменчивость, непостоянство). Для улучшения предсказательных свойств модели прогнозирования используем оценку волатильности. Ошибку прогноза в предыдущей точке t используем для корректировки прогноза в точке $t+1$ с учетом предсказанного значения ошибки в точке $t+1$. Предсказанные значения ошибки предлагаем оценивать с помощью метода авторегрессии в условиях гетероскедастичности (ARCH). Суть применения модели ARCH состоит в том, что если абсолютная величина y_t оказывается большой, то это приводит к повышению условной дисперсии в последующие периоды. В свою очередь, при высокой условной дисперсии, более вероятно появление больших (по абсолютной величине) значений y_t. Если значения y_t в течение нескольких периодов близки к 0, то это приводит к понижению условной дисперсии в последующие периоды. В свою очередь, при низкой условной дисперсии более вероятно появление малых значений y_t. Таким образом, ARCH-процесс характеризуется инерционностью условной дисперсии.

На первоначальном этапе синтеза модели прогноза предлагается оценить величину корреляции между ошибками, сделанными в один и тот же момент времени с различными упреждениями.

Предположим, что имеют место прогнозы с момента t для различных упреждений. Тогда ошибки таких прогнозов представим в виде:

$$e_t(\ell)=y_{t+\ell}-\hat{y}_t(\ell)=a_{t+\ell}+\Psi a_{t+\ell-1}+\ldots+\Psi_{\ell-1}a_{t+1}, \quad (4)$$

$$e_t(\ell+j)=y_{t+\ell+j}-\hat{y}_t(\ell+j)=a_{t+\ell+j}+\Psi a_{t+\ell+j-1}+\ldots+\Psi_j a_{t+\ell}+\Psi_{j+1}a_{t+\ell-1}+\ldots+\Psi_{\ell+j-1}a_{t+1}. \quad (5)$$

Генерация числовых последовательностей временных рядов без сезонных составляющих показала, что ошибки прогноза $e_t(\ell)$ и $e_t(\ell+j)$, сделанные для разных упреждений с одного и того же момента времени t, коррелированны. Функция прогноза в этом случае лежит целиком выше, либо целиком ниже фактических значений ряда.

Ошибки прогноза $e_t(\ell)$ и $e_t(t-j)$, сделанные при том же упреждении с различных моментов времени t и $t-j$, также коррелированны. Оптимальные ошибки прогноза с упреждением на шаг в перед $(\ell=1)$ некоррелированные, а ошибки прогноза с большим упреждением будут коррелированны.

Подведем итог. Итак, если ошибки прогноза коррелированны, то есть являются динамическими, тогда возможно построить модель предсказания по известным ошибкам прогноза $a_t, a_{t-1}, a_{t-2},\ldots$.

Оценим степень статистической связи между членами ряда. Гипотеза о постоянстве статической связи в данном случае несостоятельна, так как временной ряд курса валют является нестационарным. Степень статистической связи между измерениями ряда может усиливаться или уменьшаться. Имеет место динамическая стохастическая связь измерениями ряда, которую оценим с помощью рекуррентного коэффициента корреляции

$$Z_t(d) = S_t/d_t, \quad (6)$$

где $S_t = (1-\alpha)\, S_{t-1}+\alpha(\nabla y_{1t}\nabla y_{2t}),\ t = 1, 2, \ldots, T;$

$d_t = (1-\alpha)\, d_{t-1}+\alpha|\nabla y_{1t}\nabla y_{2t}|,\ 0 < \alpha > 1;$

$\nabla y_{it} = y_{it} - +y_{i(t-1)},\ i = 1,2.$

В выражении величины S_t и d_t являются экспоненциально-взвешенными скользящими средними произведений приростов и абсолютных произведений приростов двух рядов.

Параметр α - постоянная сглаживания.

Начальные значения S_0 и d_0 в работе [4] предлагается определять как простые арифметические средние произведений и абсолютных значений произведений приростов.

Пусть модель прогноза строится методом авторегрессии и проинтегрированного скользящего среднего (АРПСС). Реальное значение величины y_t в точке $t+1$, y_{t+1} известно, тогда прогноз на момент $t+1$ может быть скорректирован с учетом новой ошибки $a_{t+1}=y_{t+1}-\hat{y}_t(1)$ и предсказанные значения с точки $t+1$ вычисляются следующим образом:

$$\hat{y}_{t+1}(\ell)=p_1[y_{t+1+\ell-1}]+\ldots+p_{p+q}[y_{t+1+\ell-p-q}]+a_{t+1+\ell}-q_1[a_{t+1+\ell-1}]-\ldots-q_q[a_{t+1+\ell-q}] \quad (7)$$

Учитывая веса Ψ_j и используя прогнозы $\hat{y}_t(1),\hat{y}_t(2),\ldots,\hat{y}_t(L)$ с момента t, получим прогнозы с момента $t+1$ из формулы

$$\hat{y}_{t+1}(\ell)=y_t(\ell+1)+\Psi_\ell a_{t+1}=y_t(\ell+1)+\Psi_\ell[y_{t+1}-\hat{y}_t(1)]. \quad (8)$$

Итак, ошибка прогноза в предыдущей точке *t* используется для корректировки прогноза в точке *t+1*. Такой подход улучшает прогноз, но имеет смысл использовать для прогноза в точке ℓ еще и предсказанные значения ошибки. Для прогнозирования ошибки прогноза используем модель с авторегрессионной условной гетероскедастичностью (ARCH).

Процесс ARCH *q*-того порядка задается следующими выражениями:

$$\sigma_t^2=\omega+\gamma_1 y_{t-1}^2+\gamma_2 y_{t-2}^2+...+\gamma_q y_{t-q}^2, \quad (9)$$

при условии, что $y_t, y_{t-1}, y_{t-2},...,\sim N(0,\sigma_t^2)$.

σ_t^2 - условная по предыстории дисперсия y_t.

Если обозначить $y_t^2-E(y_t^2/y_{t-1},y_{t-2},...,y_{t-k})=\eta_t$, получим следующую запись процесса ARCH:

$$y_t^2=\omega+\gamma_1 y_{t-1}^2+\gamma_2 y_{t-2}^2+...+\gamma_q y_{t-q}^2+\eta_t \quad (10)$$

Обобщенную модель GARCH (p,q) получаем из формулы

$$\sigma_t^2=\omega+\sum_{j=1}^{p}\delta_j\sigma_{t-j}^2+\sum_{j=1}^{q}\gamma_j y_{t-j}^2. \quad (11)$$

Пусть $l=1$, $p=1$, $q=1$. Ошибка предсказания: $e_{t+l}=y_{t+l}-y_{t+l}^p=d_l$.

Условная дисперсия ошибки предсказания:

$$d_{d_l}^2=E(d_l^2/y,y_{t-1},...)=E(e_{t+l}^2/y_t,y_{t-1},...). \quad (12)$$

Условная дисперсия ошибки предсказания зависит от величины упреждения l и от предыстории $y_t, y_{t-1}, ..., y_{t-k}$.

Общее выражение для GARCH(1,1) имеет вид:

$$\sigma_{d_l}^2=\omega\frac{1-(\delta_1+\gamma_1)^{l-1}}{1-\delta_1-\gamma_1}+(\delta_1+\gamma_1)^{l-1}\sigma_{t+1}^2, \quad (13)$$

за исключением случая $\delta_1+\gamma_1=1$.

При $\delta_1+\gamma_1\prec 1$ условная дисперсия ошибки прогноза сходится к безусловной дисперсии GARCH(1,1):

$$\lim_{k\longrightarrow\infty}\sigma_{d_l}^2=\frac{\omega}{1-\delta_1-\gamma_1}. \quad (14)$$

Исследуем влияние остатков на эффективность моделей прогнозирования. Для этого рассмотрим обобщенную модель авторегрессии и проинтегрированного скользящего среднего с известными параметрами ($\hat{P}$, $\hat{Q}$) вида

$$\hat{P}(B)\nabla^d y_t=\hat{P}(B)e_t, \quad (15)$$

откуда остаточная ошибка

$$\hat{e}_t=\hat{P}(B)\nabla^d y_t\hat{Q}^{-1}(B). \quad (16)$$

С увеличением длины ряда $\hat{e}_t$ по своим характеристикам приближается к белому шуму. Выборочная автокорреляционная функция $\hat{a}_t$ покажет возможные действия для устранения отклонений от белого шума. Если модель верна, выборочные значения автокорреляционной функции $R(e_t)$ являются некоррелированными, распределёнными нормально с дисперсией n^{-1}. Статистика отклонений $R(e_t)$ от нуля позволит проверить гипотезу об адекватности модели.

Для модели АРПСС (p,d,q) используем первые k автокорреляции. Если модель адекватна, величина

$$Q_p = n\sum_{k=1}^{k} r_k^2(\hat{e}) \qquad (17)$$

имеет распределение, приближенное к χ^2 (k-p-q). Наблюдаемое значение $Q_н$ сравниваем с табличным значением χ^2 распределения Q_t. Если модель неадекватна, то имеет место неравенство $Q_p > Q_t$, для соответствующих степеней свободы δ 90…95 % квантилей.

Для исходного ряда получены следующие модели:

1. Структура (011): $\nabla y_t = (1-0{,}05B)e_t$ с прогнозом $y_{t+l}=y_{t+l\text{-}1}+e_{t+l}$-0,05 $e_{t+l\text{-}1}$.
2. Структура (022): $\nabla^2 y_t = (1-1{,}0439B+0{,}0522B^2)e_t$ с прогнозом $y_{t+l}=2y_{t+l\text{-}1}-y_{t+l\text{-}2}+e_{t+l}-1{,}0439e_{t+l\text{-}1}+0{,}0522e_{t+l\text{-}2}$.
3. Структура (111): $(1+0{,}257B)\nabla y_t = (1+0{,}31B)e_t$ с прогнозом $y_{t+l}=y_{t+l\text{-}1}+0{,}31y_{t+l\text{-}2}+e_{t+l}+0{,}257e_{t+l\text{-}1}$.
4. Структура (211): $(1-0{,}64B-0{,}08B^2)\nabla y_t = (1-0{,}69B)e_t$.
5. Структура (221): $(1+0{,}057B-0{,}03B^2)\nabla^2 y_t = (1-0{,}99B)e_t$.
6. Структура (222): $(1-0{,}65B-0{,}08B^2)\nabla^2 y_t = (1-1{,}7B+0{,}7B^2)e_t$.
7. Структура (012): $\nabla y_t = (1-0{,}05B+0{,}04B^2)e_t$ с прогнозом $y_{t+l}=y_{t+l\text{-}1}+e_t-0{,}055e_{t+l\text{-}1}+0{,}04e_{t+l\text{-}2}$.
8. Структура (021): $\nabla^2 y_t = (1-0{,}9926B)e_t$ с прогнозом $y_{t+l}=2y_{t+l\text{-}1}-y_{t+l\text{-}2}+e_{t+l}-0{,}9926e_{t+l\text{-}1}$.

Для каждой модели определены значения Q_p и $Q_{таб}$. Из восьми структур моделей прогнозирования наиболее приемлемой оказалась модель со структурой (2 2 2). Неадекватность моделей связана с тем, что данные по курсу валют имеют неоднородную структуру с элементами интервенции. Параметры модели с одинаковой структурой изменяются в зависимости от времени. Покажем это, разбив исходный ряд на три части, каждая из которых имеет характерные особенности.

Стандартная ошибка для рядов 1,2 равна $\sqrt{(0.022)^2+(0.123)^2}$ = 0,124.

Использование изложенного выше подхода является перспективной процедурой. Построение модели прогнозирования сопряжено с определенными трудно-

стями, которые обусловлены изменчивым характером временных рядов и решением проблем 1-8.

В работе показано, что метод построения модели ошибки прогноза с авторегрессионной условной гетероскедастичностью по сравнению с моделями АРПСС (ARIMA) позволяет найти более эффективные оценки. Эффективность достигается за счет уменьшения ошибки прогноза. В прогнозе на следующий *(t+1)*-й период имеет место ошибка, которая свойственна ARCH-процессу и ошибка, определяющаяся как разница между оценками параметров и истинным значением параметров. Точность прогноза зависит от текущей информации и от момента прогноза.

Список литературы

1. Хакен К. Синергетика / К. Хакен. – М.: Мир, 1990. – 46 с.

2. Малинецкий Г.Г. Современные проблемы нелинейной динамики / Г.Г. Малинецкий, А.Б. Попов. – М.: УРСС, 2000. – 372 с.

3. Андерсен Т. Статистический анализ временных рядов / Т. Андерсен. – М.: Мир, 1976. – 755 с.

4. Бокс Дж. Анализ временных рядов. Прогноз и управление / Дж. Бокс, Г. Дженкинс. – М.: Мир, 1974. – Вып. 1. – 212 с.

5. Колодізєв О.М. Дилінгові операції: навч. посібник / О.М. Колодізєв, Б.В. Шамша. – Х.: ВД «ІНЖЕК», 2009. – 456 с.

Рецензент: д-р техн. наук, проф. Петров Э.Г.,
Национальный университет радиоэлектроники, Харьков.

Поступила в редакцию 21.02.10

Локальні моделі прогнозування нелінійних часових рядів в умовах гетероскедастичності

Запропоновано розбити вихідну вибірку на класи за допомогою кластерного аналізу. Поточний вимір віднесено до того чи іншого класу на основі дискримінантного аналізу. Модель прогнозування побудовано у кожному класі з урахуванням ризику, отриманого з оцінювання волатильності стохастичних часових рядів із корекцією помилки, умовно отриманої з поточної і минулої інформаційної множини.

Ключові слова: гетероскедастичність, кластеризація, дискримінація, волатильність, ризик, авторегресія, GARCH модель, часовий ряд.

Local Models for Forecasting Nonlinear Time Series under Heteroscedasticity

The paper proposes to split the original sample into classes by cluster analysis. The current measurement will refer to a particular class by discriminant analysis. Forecast model will be built in each class, given the risk, obtained by evaluating the volatility of stochastic time series with the correction of errors, conventionally obtained from current and past information set.

Keywords: heteroscedasticity, clustering, discrimination, volatility, risk, autoregressive, GARCH model, time series.