Событийное моделирование и прогноз финансовых временных рядов*

А. А. Романенко
angriff07@gmail.com
Московский физико-технический институт, ФУПМ, каф. «Интеллектуальные системы»

Финансовые временные ряды обычно сильно зашумлены и зависят от других временных рядов (например, курс доллара или пошлины на таможне). Но насколько сильна эта зависимость, какие факторы учитывать при их прогнозировании, однозначно определить непросто. В работе для прогнозирования поведения целевого ряда используется разметка временных рядов. Предлагается алгоритм порождения признаков из размеченных временных рядов и генетический алгоритм отбора признаков на размеченных временных рядах.

Ключевые слова: *временные ряды, разметка временных рядов, логистическая регрессия, прогнозирование событий.*

Введение

В данной работе ставится задача прогнозирования динамики роста финансовых временных рядов. Это задача является экономически важной и сложной. Сложность обусловлена тем, что такие временные ряды обычно сильно зашумлены и зависят от других временных рядов (курс доллара, пошлины на таможне, и т.д.), но степень этой зависимости однозначно определить сложно.

Различают два основных подхода к прогнозированию цен: технический анализ [1] и фундаментальный [2]. Оба подхода сейчас активно применяются при прогнозировании цен на различные активы и имеют своих критиков и сторонников. Прогнозирование методами технического анализа основано на анализе временных рядов и индикаторов, прогнозирование методами фундаментального анализа — на анализе экономической ситуации и новостей.

В данной работе для прогнозирования динамики роста временных рядов предлагается использовать методы событийного моделирования к временным рядам с выделенными на них трендами [3]. Для выделения трендов к временным рядам применим технологию разметки [4, 5, 6]. Для прогнозирования используем нейронные сети [7, 8], а для поиска наилучшего набора признаков — генетический алгоритм.

В следующем разделе будет поставлена задача прогнозирования. Затем будет описан способ ее решения, а также вычислительный эксперимент на реальных данных и представлены его результаты.

Постановка и предлагаемое решение задачи

Постановка задачи. Пусть $\mathbf{f}_i(t),\ i = 1, \ldots, a$ — данные временные ряды, t —номера отсчётов. Предполагаем, что во временных рядах нет пропущенных значений. Задача состоит в том, чтобы по известным значениям временных рядов

$$\mathbf{f}_i(t),\ i = 1, \ldots, a,\ t = 1, \ldots, T$$

спрогнозировать для временного ряда $\mathbf{f}_1$ увеличится его значение в момент времени $T + 1$ или уменьшится.

Научный руководитель В. В. Стрижов

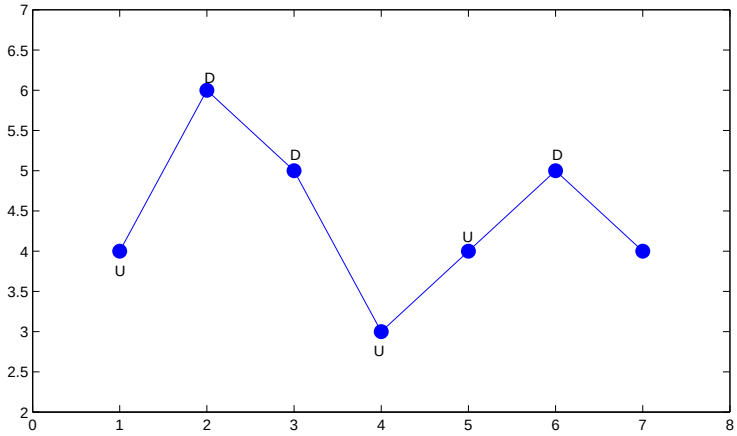


Рис. 1. Пример разметки в алфавите $\mathcal{M} = \{up, down\}$

Разметка временного ряда.

Определение 1. *Множеством меток называется конечное множество* $\mathcal{M} = \{m_1, \ldots, m_r\}$, *которое задается экспертом.*

Пример множества меток: $\mathcal{M}$= $\{up, down, plt\}$, где "*up*" — метка для обозначения точек возрастания, "*down*" — убывания, "*plt*" — метка для обозначения плато. На рис. 1 показан пример разметки временного ряда в алфавите $\mathcal{M}$= $\{up, down\}$.

Зафиксируем множество меток $\mathcal{M}$. Определим разбиение временного ряда на сегменты $\bar{\mathbf{s}} = (s_1, \ldots, s_V)$: $s_k = \{\mathbf{f}(i), \mathbf{f}(i+1), \ldots, \mathbf{f}(i+l_k)\}$, $s_{k_1} \bigcap s_{k_2} = \varnothing$ при $k_1 \neq k_2$, $\bigcup\limits_{k=1}^{V} s_k =$ $= \{\mathbf{f}(1), \ldots, \mathbf{f}(T)\}$.

Определение 2. *Разметкой временного ряда* $\mathbf{f}(t)$, $t = 1, \ldots, T$ *назовем пару* $(\bar{\mathbf{s}}, \bar{\mathbf{m}})$: $\bar{\mathbf{m}} =$ $= (m_1, \ldots, m_U)$, $m_i \in \mathcal{M}$.

Предлагается произвести разметку всех временных рядов в алфавите $\mathcal{M}$= $\{1, -1\}$. Метка "1" ставиться участку временного ряда, на котором его значения растут; метка "$-$ -1" ставится, если значения не увеличиваются. Вообще говоря, длина сегмента разметки как внутри одного ряда, так и в разных рядах может менятся. Но мы будем считать, что разметка *синхронная*, то есть длины сегментов и их начала для всех временных рядов совпадают. В таком случае можно рассматривать не изначально данные временные ряды, а последовательности из 1 и -1. Тогда задача прогнозирования сводится к прогнозированию появления в первой последовательности 1 или -1.

Порождение признаков и прогноз временного ряда. После процедуры разметки получим a последовательностей $\mathbf{f}_i$ одинаковой длины T из 1 и -1.

Для порождения признаков используем идеи из [9, 10]. Выберем натуральное число b, назовем его *глубиной логирования*. Каждому $\mathbf{f}_1(k+1)$ поставим в соответствие матрицу размера $a \times b$:

$$\begin{pmatrix} \mathbf{f}_1(k-b+1) & \ldots & \mathbf{f}_1(k-1) & \mathbf{f}_1(k) \\ \mathbf{f}_2(k-b+1) & \ldots & \mathbf{f}_2(k-1) & \mathbf{f}_2(k) \\ \vdots & \ddots & \vdots & \vdots \\ \mathbf{f}_{a-1}(k-b+1) & \ldots & \mathbf{f}_{a-1}(k-1) & \mathbf{f}_{a-1}(k) \\ \mathbf{f}_a(k-b+1) & \ldots & \mathbf{f}_a(k-1) & \mathbf{f}_a(k) \end{pmatrix}$$

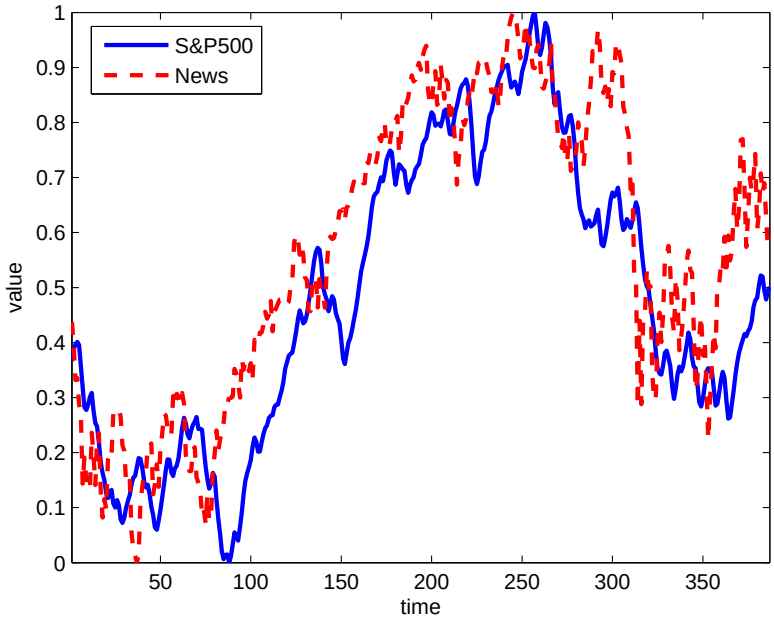


Рис. 2. Используемые временные ряды

Теперь векторизуем ее и получим вектор $\boldsymbol{x}_k$:

$$\mathbf{x}_k = \begin{pmatrix} \mathbf{f}_1(k-b+1) \\ \dots \\ \mathbf{f}_1(k) \\ \vdots \\ \mathbf{f}_a(k-b+1) \\ \dots \\ \mathbf{f}_a(k) \end{pmatrix}. \tag{1}$$

В итоге получим множество прецедентов

$$({\mathbf{x}_k}^T, y_k), \text{ где } y_k = \mathbf{f}_1(k+1), \ k = b, \dots, T-1,$$

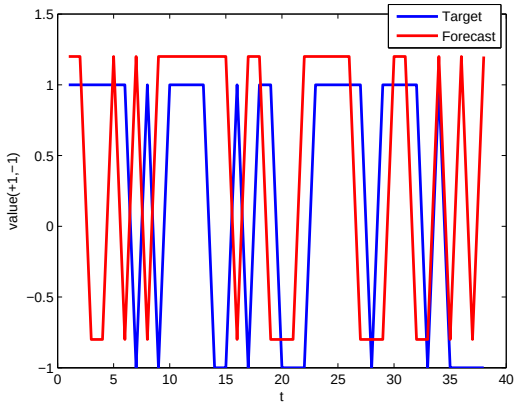
$\mathbf{x}_k$ — признаковое описание k-го объекта, $y_k \in \{+1, -1\}$ — класс, к которому он относится. Остается выбрать модель алгоритма и метод обучения, чтобы решить задачу классификации на два класса. В работе в качестве алгоритма кластеризации используются нейронные сети. Для улучшения качества классификации можно воспользоваться методами отбора признаков. Предлагается использовать генетический алгоритм отбора признаков.

Инстересно исследовать зависимость результата прогнозирования от глубины логирования b. Это исследование проведено на практике, и его результаты представлены ниже.

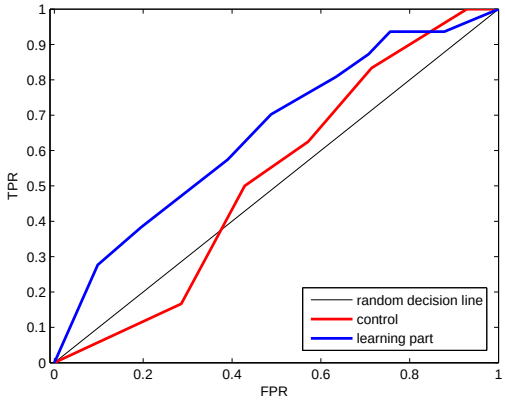
Генетический алгоритм отбора признаков. Пусть α и β — бинарные строки длины $a \times b$:

$$\alpha = (\alpha_1, \dots, \alpha_{ab}),$$
$$\beta = (\beta_1, \dots, \beta_{ab}).$$

Тогда $\alpha_i = 1$ означает, что при прогнозе учитывается i-ый признак; $\alpha_i = 0$ — не учитывается.

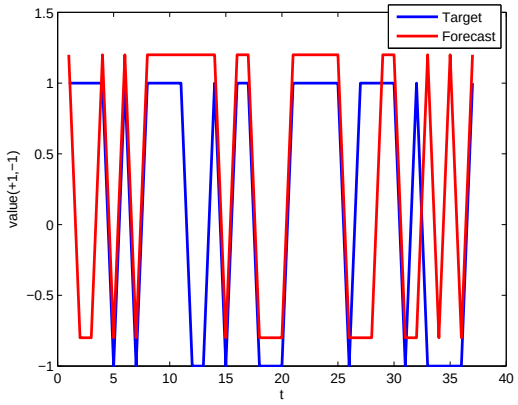


(a) Прогноз

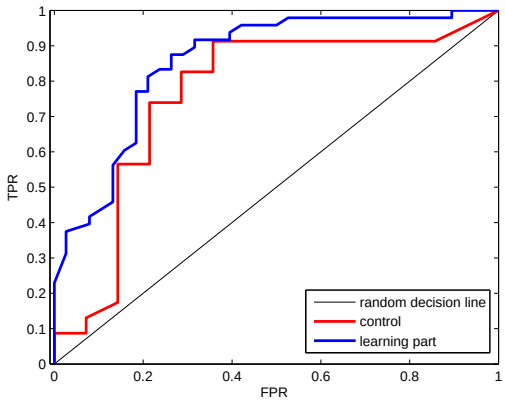


(b) ROC-кривая на обучении и контроле

Рис. 3. Результаты при $b = 3$: $Error = 44\%$, $AUC_1 = 0{,}64$, $AUC_2 = 0{,}52$

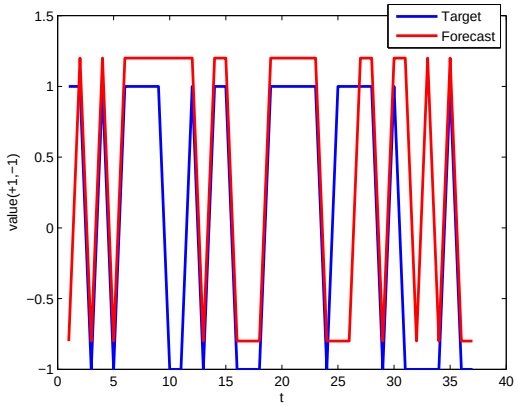


(a) Прогноз

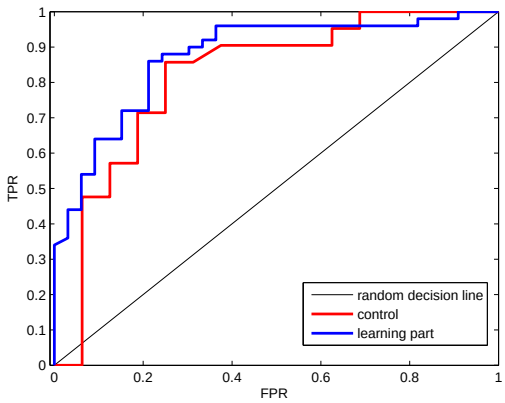


(b) ROC-кривая на обучении и контроле

Рис. 4. Результаты при $b = 6$: $Error = 24\%$, $AUC_1 = 0{,}86$, $AUC_2 = 0{,}76$



(a) Прогноз



(b) ROC-кривая на обучении и контроле

Рис. 5. Результаты при $b = 9$: $Error = 18\%$, $AUC_1 = 0{,}87$, $AUC_2 = 0{,}81$

Для поиска оптимального набора признаков будем порождать α и β с помощью следующего генетического алгоритма. Ему на вход подаются следующие параметры: I_{max} — ограничение на количество итераций, $Error_{max}$ — допустимый процент ошибок на обучающей выборке, n — число признаков, q — число мутаций при порождении новой пары признаков. Алгоритм останавливается, когда на полученном наборе признаков достигнут порог ошибки или будет сделано изначально заданное количество итераций I_{max}.

ПРОЦЕДУРА `Поиск оптимального набора признаков`

Вход: $I_{max}, Error_{max}, n, q$ —ограничение на количество итераций, требуемая точность, число признаков,число мутаций

сгенерировать случайным образом бинарные строки α и β длины n :

$\alpha = rand(1, n)$;

$\beta = rand(1, n)$;

инициализация счетчика

$i = 1$;

$BEST\_Error = 1$;

для $i = 1, \ldots, I_{max}$

обучить нейронную сеть по признакам i, где $\alpha_i = 1$:

$Error = train(X_\alpha, y)$;

если $Error < BEST\_Error$ **то**

$BEST\_SET = \alpha$;

$BEST\_Error = Error$;

обучить нейронную сеть по признакам i, где $\beta_i = 1$:

$Error = train(X_\beta, y)$;

если $Error < BEST\_Error$ **то**

$BEST\_SET = \beta$;

$BEST\_Error = Error$;

если $BEST\_Error < Error_{max}$ **то**

выход

СКРЕЩИВАНИЕ

запомнить старые α и β и выбрать случайное целое $k \in \{1, \ldots, n-1\}$:

$k = random(n)$;

$\alpha_{old} = \alpha$;

$\beta_{old} = \beta$;

$\alpha = [\alpha_{old}(1:k), \beta_{old}(k+1:end)]$;

$\beta = [\beta_{old}(1:k), \alpha_{old}(k+1:end)]$;

МУТАЦИЯ

изменить значения в α и β на q произвольных позициях:

для j=1,. . . ,q

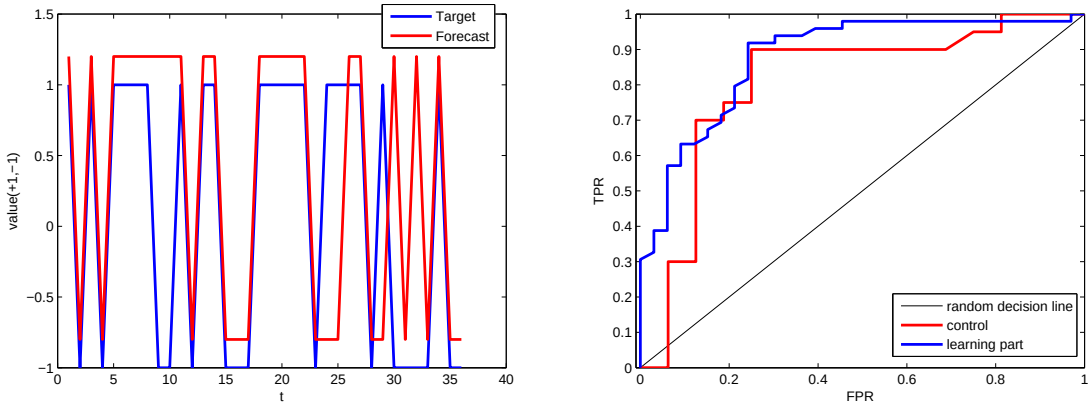
$k = random(n)$;

$\alpha_k = |\alpha - 1|$;

$k = random(n)$;

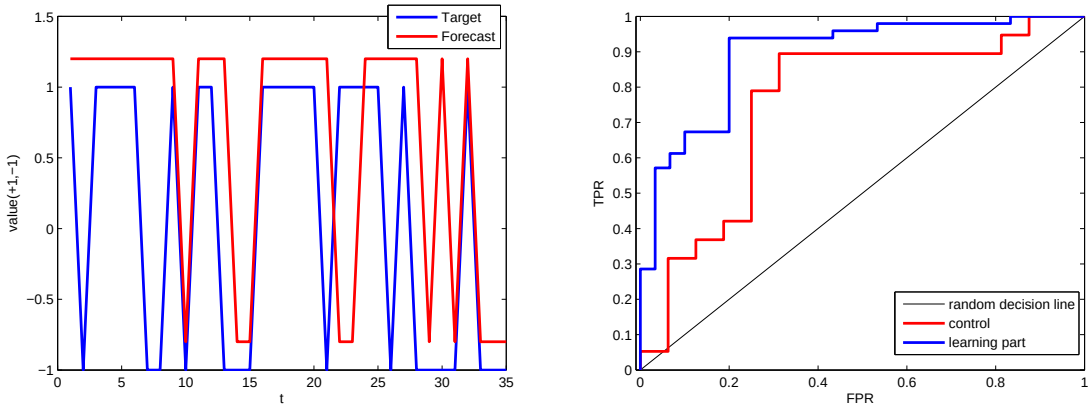
$\beta_k = |\beta - 1|$;

вернуть $BEST\_SET$



(a) Прогноз (b) ROC-кривая на обучении и контроле

Рис. 6. Результаты при $b = 11$: $Error = 19\%$, $AUC_1 = 0{,}88$, $AUC_2 = 0{,}80$



(a) Прогноз (b) ROC-кривая на обучении и контроле

Рис. 7. Результаты при $b = 15$: $Error = 28\%, AUC_1 = 0{,}89$, $AUC_2 = 0{,}75$

Вычислительный эксперимент

Описание временных рядов. Спрогнозируем динамику роста индекса S&P-500 по его истории и по временному ряду, характеризующему поток новостей по заданной тематике. Количество временных отсчетов $T = 394$. Так как в работе нас не интересуют абсолютные значения временного ряда, то можно нормировать временные ряды. Для временного ряда $\mathbf{f}(t)$ нормированный ряд будет вычисляться так:

$$\hat{\mathbf{f}}(t) = \frac{\mathbf{f}(t) - \mathbf{f}_{min}}{\mathbf{f}_{max} - \mathbf{f}_{min}}, \ t = 1, \ldots, T,$$

где $\mathbf{f}_{min}$ и $\mathbf{f}_{max}$ — минимальное и максимальное значения временного ряда $\mathbf{f}$. На рис. 2 показаны используемые нормированные временные ряды.

Результаты вычислительного эксперимента. Ниже представлены полученный в результате прогноза ряд (красным цветом) прогнозируемый ряд (синим цветом) для некоторых значений глубины логирования и ROC-кривые для полученного прогноза на обучающей выборке и на контроле. Для удобства прогноз на графике изображен немного

выше прогнозируемого ряда. В качестве обучающей выборки были выбраны 70% всех прецедентов, в качестве контрольной — оставшаяся часть. Для каждого значения глубины логирования приведены значения площади под ROC-кривыми на обучении AUC_1 и на контроле AUC_2.

Из графиков видно, что сначала при росте значения глубины логирования b качество прогноза улучшается (увеличивается площадь под ROC-кривой AUC_2 на контрольной выборке): рис. 3 – 4. При определенном значении глубины логирования ошибка AUC_2 достигает своего максимума (рис. 5). При дальнейшем росте b значение AUC_2 падает (рис. 6 – 7). Это означает, что происходит переобучение: несмотря на то, что площадь под ROC-кривой на обучающей выборке AUC_1 растет с ростом b, значение AUC_2 падает, а $Error$ увеличивается.

Таким образом, получили, что для данных временных рядов наилучшее значение глубины логирования $b = 9$, при этом ошибка на контрольной выборке равна $Error = 18\%$, площадь под ROC-кривой равна $AUC = 0{,}81$.

Код и данные для проведения эксперимента находятся по адресу [11].

Литература

[1] Achelis S. B. *Technical analysis from A to Z* / New York:McGraw Hill, 2001.

[2] Ritchie J. C. *Fundamental Analysis: A Back-To-The Basics Investment Guide to Selecting Quality Stocks* / Irwin Professional Pub, 1996.

[3] Кононенко Д. С. *Прогнозирование событий* // Машинное обучение и анализ данных, 2010, Т. 1, № 1, С. 113–115.

[4] Колесникова С. И. *Особенности применения эталонных моделей для разметки временного ряда при распознавании состояний сложного объекта.* //Управление, вычислительная техника и информатика, 2011, Т. 1, № 1, С. 31–36.

[5] Чехович Ю. В. *Элементы алгебраической теории синтеза обучаемых алгоритмов выделения трендов.* // Диссертация на соискание степени магистра, ФУПМ МФТИ(ГУ), 2003.

[6] Чехович Ю. В. *Об обучаемых алгоритмах выделения трендов.* // Искусственный интеллект, 2002.

[7] Воронцов К. В. *Лекции по линейным алгоритмам классификации*, www.machinelearning.ru/, 2011

[8] Головко В. А. *Нейронные сети: обучение, организация и применение* /ИПРЖР, 2001.

[9] Филипенков Н. В. *О задачах анализа пучков временных рядов с изменяющимися закономерностями.* // Диссертация на соискание степени магистра, ВМК МГУ, 2006.

[10] Филипенков Н. В. *Об алгоритмах прогнозирования процессов с плавно меняющимися закономерностями.* // Диссертация на соискание степени кандидата наук, ВЦ РАН им. А. А. Дородницына, 2010.

[11] Исходные временные ряды, http://bit.ly/uCf4XV, 2011.