7. Krzanowski W. A criterion for determining the number of groupsin a dataset using sum of squares clustering [Text] / W. Krzanowski, Y. Lai // Biometrics. – 1985. – № 44. – pp. 23–34.
8. Sugar C. Finding the number of clusters in a data set: An information theoretic approach [Text] / C. Sugar, G. James // J. of the American Statistical Association. – 2003. – № 98. – pp. 750–763.
9. Calinski, R. B. Dendrite method for cluster analysis [Text] / R. B. Calinski, J. A. Harabasz // Communications in Statistics. – 1974. – № 93. – pp. 1–27.
10. Семенкин, Е. С. Методы оптимизации в управлении сложными системами [Текст]: учебное пособие / Е.С. Семенкин, О.Э. Семенкина, В.А. Терсков; Россия. Министерство внутренних дел. – Красноярск: Сибирский юридический институт, 2000. – 254 с. – ISBN 5–93182–008–6.
11. Статюха Г.О. Вступ до планування оптимального експерименту [Текст]: навч. посіб. / Г.О. Статюха, Д.М. Складанний, О.С. Бондаренко. – К.: НТУУ «КПІ», 2011. – 124 с. – 300 пр. ISBN 978-966-622-408-1.

У роботі пропонується метод прогнозування знаків приростів часових рядів, який базується на застосуванні в комплексі комбінованих моделей селективного типу, складовими яких є індикатори плинних середніх, та попередньої кластеризації часових рядів за методом К-найближчих сусідів

Ключові слова: часовий ряд, прогнозування, знак приросту, кластеризація, метод найближчих сусідів, комбінована модель прогнозування, плинна середня

В работе предлагается метод прогнозирования знаков приростов временных рядов, который базируется на применении в комплексе комбинированных моделей селективного типа, составляющими которых являются индикаторы скользящих средних, и предварительной кластеризации временных рядов по методу К-ближайших соседей

Ключевые слова: временной ряд, прогнозирование, знак прироста, кластеризация, метод ближайших соседей, комбинированная модель прогнозирования, скользящая средняя

УДК 004:519.2

МЕТОД ПРОГНОЗУВАННЯ ЗНАКІВ ПРИРОСТІВ ЧАСОВИХ РЯДІВ

О. Ю. Берзлев
Аспірант
Кафедра кібернетики і прикладної математики
Ужгородський національний університет
вул. Університетська 14, м. Ужгород, Україна, 88000
E-mail: berzlev@gmail.com

1. Вступ

Відомо, що більшість часових рядів, для яких виникає задача прогнозування, зокрема рядів економічної природи, як правило, характеризуються нестаціонарністю і нестійкістю відносно їх середнього рівня. Переважна більшість класичних статистичних моделей та відповідних методів (експоненціальні, лінійні регресійні, авторегресійні типу ARIMA [1-4]) не призначені для прогнозування нестаціонарних часових рядів, а ті, які для цього призначені (ARIMAX, нелінійні регресійні тощо) характеризуються складністю оцінювання численних параметрів та ідентифікації функціональних залежностей. З огляду на це, окрім задачі прогнозування майбутніх значень рядів, застосовуються інші специфічні задачі, серед яких: ідентифікація моментів локальних екстремумів [5], прогнозування знаків приростів рядів. Остання розглядається в даній роботі.

На фінансовому і валютному ринках часто виникає задача передбачення короткочасної динаміки часового ряду без розрахунку безпосередньо прогнозних значень. Тобто управління процесом прогнозування в даному випадку передбачає вибір або побудову такої моделі, яка б розраховувала прогноз знаку приросту значення часового ряду на одну точку вперед з необхідною максимальною точністю. Моделі такого типу зазвичай застосовуються для визначення напрямку руху ціни валютних пар і можуть використовуватися для визначення екстремальних точок або точок розвороту ринку, тобто таких точок, які вказують на подальший напрямок руху ціни. На практиці для підвищення точності прогнозування знаків приростів застосовують специфічні моделі та методи.

Питання розробки моделей та методів прогнозування знаків приростів висвітлені в роботі [6], зокрема в ній пропонується модель прогнозування знаків приростів рядів з нестабільним характером коливань. Але на сьогоднішній день не розроблено універсальної методики вирішення цієї задачі, яка б повністю задовольняла цілі прогнозиста, аналітика або інвестора в частині забезпечення необхідної точності прогнозів незалежно від структури часових рядів.

Для вирішення даної задачі автором пропонується метод, який базується на використанні в комплексі

комбінованих моделей прогнозування та попередньої кластеризації часових рядів. Також автором запропоновані деякі критерії оцінки якості прогнозування для даної задачі.

Актуальність цих досліджень має як практичне, так і теоретичне значення для розробки математичного інструментарію, що може бути використаний в подальших дослідженнях.

2. Ціль і задачі дослідження

Ціллю дослідження є розробка методу прогнозування знаків приростів часових рядів, який базується на використанні в комплексі попередньої кластеризації рядів за методом K-найближчих сусідів та комбінованих моделей, складовими яких є математичні інструменти технічного аналізу для прогнозування динаміки, а саме індикатори плинних середніх.

3. Постановка задачі прогнозування знаків приростів

Скінчену послідовність вимірювань, які фіксуються в дискретні моменти часу $t_i \in S$, $i=\overline{1,n}$, S – деяка дискретна множина, будемо називати дискретним часовим рядом $\{z_i\}_{i=1}^{n}=\{z_1,z_2,\ldots,z_n\}=\{z(t_1),z(t_2),\ldots,z(t_n)\}$, t_1 – початковий момент часу.

Розглянемо формальну постановку задачі прогнозування знаків приростів часового ряду. На основі ряду $\{z_i\}_{i=1}^{n}$ побудуємо ряд, який складається з перших різниць $\{\Delta z_i\}_{i=1}^{n}$, де $\Delta z_i = z_i - z_{i-1}$, $i=\overline{2,n}$. Позначимо через $\{\chi_i\}_{i=2}^{n}$ знаковий ряд, де $\chi_i = \mathrm{sgn}(\Delta z_i)$. Прогноз знаку приросту, який розраховується в точці n на τ точок вперед позначимо через $\hat{\chi}_\tau(n)$. Функціональну залежність, на основі якої прогнозується короткострокова динаміка часового ряду, в даній роботі знак приросту часового ряду на одну точку вперед ($\tau=1$), назвемо моделлю прогнозування знаків приростів і позначимо через F. Формально її можна записати так: $\hat{\chi}_{n+1}=\hat{\chi}_1(n)=F(z_{n-m+1},z_{n-m+2},\ldots,z_n)$.

Якщо існує достатньо історичних даних спостережень часового ряду, то перед реалізацією прогнозу доцільно оцінити якість моделей прогнозування на даному часовому ряді. Оцінки можуть бути використані для уточнення моделей або для побудови довірчих інтервалів прогнозів. Для побудови критеріїв оцінки якості прогнозування, розрахуємо прогнози знаків приростів для ряду $Z'=\{z_i\}_{i=n-m+1}^{n}$ довжини m на основі ряду $\{z_i\}_{i=n-m-q+1}^{n-m}$ довжини q:

$$\hat{\chi}_1(n-m)=F(z_{n-m-q+1},z_{n-m-q+2},\ldots,z_{n-m}),$$

$$\hat{\chi}_1(n-m+1)=F(z_{n-m-q+2},z_{n-m-q+3},\ldots,z_{n-m+1}),$$

$$\vdots$$

$$\hat{\chi}_1(n-1)=F(z_{n-q},z_{n-q+1},\ldots,z_{n-1}),\ m+q\le n,$$

де $\hat{\chi}_1(i)$ – прогноз знаку приросту, який реалізується в момент t_i (в точці i) на 1 точку вперед, $i=\overline{n-m,n-1}$. Послідовність таких прогнозів ретроспективного ряду позначимо через $\hat{X}=\{\hat{\chi}_1(i)\}_{i=n-m}^{n-1}$.

Для побудови критеріїв оцінки якості прогнозування необхідно визначити, які з отриманих прогнозних значень послідовності $\hat{X}$, слід врахувати в цільовій функції критерію. Підпослідовність послідовності $\hat{X}$, яка використовується для оцінки точності прогнозування назвемо оцінювальною і позначимо через $\hat{X}^*=\{\hat{\chi}_1(k_j)\}_{j=1}^{v}=\{\hat{\chi}_{k_j}\}_{j=1}^{v}$, $k_j\in[n-m,n-1]$, $v\le m$, $m<n$, v – кількість елементів послідовності $\hat{X}^*$. Послідовність $\hat{X}^*$ може бути побудована з найбільш значимих прогнозів за допомогою експертного оцінювання.

Можна визначити такі критерії оцінки якості прогнозу для задачі прогнозування знаків приростів:

$$-\ I^1=I^1(Z',\hat{X}^*)=\frac{1}{v}\sum_{j=1}^{v}\omega_{k_j}\hat{\chi}_{k_j}\mathrm{sign}(\Delta z_{k_j}),$$

де $\Delta z_{k_j}=z_{k_j}-z_{k_j-1}$ – прогнозні прирости.

Тут і далі $k_j\in[n-m,n-1]$, $v\le m$, $m<n$, $\sum_{j=1}^{v}\omega_{k_j}=1$;

$$-\ I^2=I^2(Z',\hat{X}^*)=\frac{1}{v}\sum_{j=1}^{v}\omega_{k_j}\mu_{k_j},$$

$$\mu_{k_j}=\begin{cases}1, (\hat{\chi}_{k_j}\Delta z_{k_j}>0)\vee(\hat{\chi}_{k_j}=0\wedge\Delta z_{k_j}=0)\\ 0, \hat{\chi}_{k_j}\Delta z_{k_j}<0\\ h, (\hat{\chi}_{k_j}=0\wedge\Delta z_{k_j}\ne 0)\vee(\hat{\chi}_{k_j}\ne 0\wedge\Delta z_{k_j}=0),\end{cases}$$

де $h\in[0,1)$ – поріг, який дозволяє врахувати прогноз нульового приросту. На практиці рекомендується обирати $h\in[0.2,0.5]$;

- оскільки на практиці випадок, коли одночасно $\hat{\chi}_{k_j}=0$ і $\Delta z_{k_j}=0$ виникає рідко, то I^2 можна замінити оцінкою, яка використовує функцію Хевісайда:

$$I^3=I^3(Z',\hat{X}^*)=\frac{1}{v}\sum_{j=1}^{v}\omega_{k_j}H_h(\hat{\chi}_{k_j}\Delta z_{k_j}), \quad (1)$$

$$H_h(x)=\begin{cases}0, x<0\\ h, x=0\\ 1, x>0\end{cases},\ x\in\mathbb{R}. \quad (2)$$

Нехай потрібно оцінити якість L моделей прогнозування типу F, які були протестовані на ретроспективному ряді. Позначимо через $\hat{X}_p$ – оцінювальний ряд, який був отриманий за p-ю моделлю при прогнозуванні ретроспективного ряду $\{z_i\}_{i=n-m+1}^{n}$ на 1 крок вперед. Тоді оптимальною буде вважатися та модель, якій відповідає максимальне значення наведених критеріїв $I^j\left(Z',\hat{X}_p^*\right)\to\max$ або $1-I^j\left(Z',\hat{X}_p^*\right)\to\min$, $j=\overline{1,3}$, $p=\overline{1,L}$.

4. Основна частина дослідження

Постановка задачі. Нехай задана множина моделей прогнозування $f_1,f_2,\ldots,f_L$, на основі яких в точці n ряду $\{z_i\}_{i=1}^{n}$ можуть бути розраховані оцінки знаків приростів $\chi_1^p(n)$, $p=\overline{1,L}$. На основі множини даних моделей та ретроспективного ряду $\{z_i\}_{i=n-m+1}^{n}$ в точці n розраху-

вати найбільш точну прогнозну оцінку знаку приросту на одну точку вперед.

Розв'язання задачі. Запропонований метод прогнозування базується на послідовному виконанні взаємопов'язаних алгоритмів: алгоритм кластеризації часового ряду [7-9] та алгоритм реалізації комбінованої моделі прогнозування [6,10].

Кластеризація часового ряду на основі методу K - найближчих сусідів. Кластером довжини m часового ряду $\{z_i\}_{i=1}^{n}$, який представляється скінченою послідовністю дійсних чисел, будемо називати підпослідовність $\{z_{k_j}\}_{j=1}^{m}$ з m елементів, $m<n$, $k_{j+1}=k_j+1$ для $j=\overline{1,m-1}$ (порядок слідування елементів у підпослідовностях такий же, як і у часовому ряді). Кластери можуть представлятися безпосередньо як підпослідовності елементів вхідного часового ряду або шляхом введення відстаней між елементами в середині кластерів. В даному методі будемо розглядати представлення кластерів на основі знакових послідовностей. Якщо $\{\chi_i\}_{i=2}^{n}$ – знакова послідовність часового ряду $\{z_i\}_{i=1}^{n}$, тоді знакові кластери будуть мати вигляд: $\chi_{(m)}^{s}=\{\chi_{k_1^s},\chi_{k_2^s},\ldots,\chi_{k_m^s}\}=\{\chi_{k_j^s}\}_{j=1}^{m}$, $k_{j+1}^s=k_j^s+1$, $\chi_{k_1^s}=\chi_s$, $s=\overline{2,n-m}$, де $\chi_{(m)}^{s}$ – кластер, що складається з m елементів, s – індекс початкового елемента ряду. Кластер $\chi_{(m)}^{n-m}$ будемо називати опорним, всі інші кластери $\chi_{(m)}^{s}$, $s=\overline{2,n-m-1}$ будемо називати неопорними. Очевидно, що число знакових неопорних кластерів з m елементів, що побудовані на основі ряду з n елементів рівне $n-m-1$.

Слід зазначити, що термін кластер (pattern), яким послуговуємося в даній роботі, використовується в [7]. У роботі [8] використовується термін vector, також у різних авторів зустрічаються терміни set, pieces тощо. Для визначення опорного кластеру використовуються також терміни останній придатний вектор (last available vector), історія ряду тощо.

Оскільки кожен кластер може бути представлений точкою в m - вимірному просторі, можемо розрахувати міри близькості або метричні відстані між опорним та всіма неорними кластерами. В якості мір близькості можуть бути використані: відстані Евкліда, Мінковського, Махаланобіса, або у випадку представлення кластерів на основі знакових послідовностей: міри подібності Хеммінга, Роджерса-Танімото тощо. В результаті застосування алгоритму кластеризації за методом K - найближчих сусідів, отримаємо K неопорних кластерів, подібних до опорного (відстань яких до опорного кластеру мінімальна). Позначимо їх через $\chi_{(m)}^{y-m+1}\in\aleph$, $y\in[m+1,n-1]$, $\aleph$ – множина K кластерів, подібних до опорного кластеру $\chi_{(m)}^{n-m}$.

Реалізація комбінованої моделі прогнозування на основі плинних середніх. Враховуючи специфіку поставленої задачі, будемо формувати множину моделей прогнозування на основі моделей, які базуються на математичних інструментах технічного аналізу, а саме трендових індикаторах плинних середніх, які призначені для прогнозування динаміки часового ряду. Для побудови множини моделей можуть бути обрані такі індикатори:

- $\chi_1(n)=\operatorname{sign}\left(\frac{1}{p}\sum_{i=1}^{p}z_{n-i+1}\right)$ – прогноз знаку приросту на основі простої плинної середньої з періодом $p>0$;

- $\chi_1(n)=\operatorname{sign}\left(\frac{1}{\nu_p}\sum_{i=1}^{p}(p-i+1)z_{n-i+1}\right)$, $\nu_p=\sum_{i=1}^{p}i$ – зважена плинна середня;

- $\chi_1(n)=\operatorname{sign}\left(\sqrt[p]{\prod_{i=1}^{p}z_{n-i+1}}\right)$ – геометрична плинна середня;

- $\chi_1(n)=\operatorname{sign}\left(\prod_{i=1}^{p}(z_{n-i+1})^{p-i+1}\right)^{\frac{1}{\nu_p}}$, $\nu_p=\sum_{i=1}^{p}i$ – зважена геом. плинна середня.

Ціллю управління процесом прогнозування в комбінованих моделях є врахування в оцінці прогнозу особливих «корисних» характеристик кожної прогнозної моделі.

Виділення цих характеристик здійснюється в першу чергу завдяки механізмам селекції та гібридизації моделей.

В даному методі обмежимося селективним принципом побудови прогнозу.

Нехай на основі деякої міри близькості в точці z_n було визначено K кластерів $\chi_{(m)}^{y-m+1}\in\aleph$, $y\in[m+1,n-1]$, $\operatorname{card}(\aleph)=K$, подібних опорному кластеру $\chi_{(m)}^{n-m}$. Останнім елементом кожного з кластерів $\chi_{(m)}^{y-m+1}$ будуть елементи z_y. В якості критерію селекції (відбору найбільш точних моделей) в точках z_y скористаємося оцінкою (1), з функцією Хевісайда з параметром $h=0$ і оцінювальними послідовностями, які представляються кластерами $\chi_{(m)}^{y-m+1}\in\aleph$. Для спрощення в точках z_y будемо відбирати з множини моделей єдину модель, для якої критерій I^3 максимальний. Позначимо прогнози знаків приростів відібраних моделей на одну точку вперед, які розраховуються в точках z_y через $\chi_1^d(y)$, $d=\overline{1,K}$. Підрахуємо в кожному з кластерів $\chi_{(m)}^{y-m+1}\in\aleph$ кількість додатних приростів, позначимо її через γ^+ і кількість від'ємних приростів – γ^-.

Тоді прогноз знаку приросту на одну точку вперед, який розраховується в точці z_n можна визначити за формулою:

$$\chi_1(n)=\operatorname{sign}\left(\phi_h-\frac{K+1}{2}\right), \tag{3}$$

$$\phi_h=\sum_{d=1}^{K}H_h(\chi_1^d(y))+H_h(\gamma^+-\gamma^-), \tag{4}$$

де H_h – функція Хевісайда з параметром h (2).

Отже, запропонований метод складається з таких кроків:

1. Передпрогнозний аналіз часового ряду, побудова опорного кластеру і множини неопорних кластерів. Знаходження K неопорних кластерів, подібних до опорного на основі певної міри близькості за методом K - найближчих сусідів.

2. Реалізація комбінованої моделі прогнозування на основі плинних середніх, механізм селекції в якій виконується на основі деякого критерію селекції, наприклад, за правилом (2).

3. Реалізація прогнозу за формулами (3),(4).

5. Висновки і чисельні результати

Описана методика була протестована і реалізована в програмному середовищі. Для тестування було обрано ряди валютних пар: EUR-USD, EUR-JPY, EUR-GBP за останні 5 років (щоденні дані). Кожен ряд по більш ніж 3000 вимірювань. Були реалізовані окремо плинні середні (прості, зважені, геометричні, зважені геометричні) та комбінована модель на основі даних плинних середніх з періодом $p=5$. Кластеризація часових рядів відбувалась за методом K - найближчих сусідів, в якості міри близькості було обрано відстань Евкліда ($m=2$, $K=5$). Були розраховані середні похибки прогнозування (на основі критерію I^3 (1), (2) у відсотках) знаків приростів на одну точку вперед на ділянках ряду в 500 точок згідно з запропонованим методом $h=0$ (3), (4). Для вказаних часових рядів описана методика дозволила в середньому підвищити точність прогнозування знаків приростів відносно наївного алгоритму на 7%, відносно алгоритму Лукашина на 3%, відносно середнього показника по плинним середнім на 1%. Наприклад, для раду EUR-USD точність запропонованого методу складає 53,07%, в той час як точність наївного алгоритму, в якому застосовується гіпотеза про те, що знак приросту в попередній точці збережеться і в наступній, складає 45,65%.

Наукова новизна. Для підвищення точності прогнозування знаків приростів часових рядів запропоновано метод, який базується на комбінованій моделі, в основі якої лежать плинні середні, з попередньою кластеризацією часового ряду за методом K - найближчих сусідів.

Практична цінність роботи в тому, що запропонована методика може використовуватися в якості складової інформаційних прогнозних систем, зокрема таких, які використовуються на валютному ринку для підвищення точності прогнозування знаків приростів часових рядів на одну точку вперед. Проведений порівняльний аналіз результатів прогнозування запропонованого методу з наївним підходом, алгоритмом Лукашина, звичайною комбінованої моделлю на основі плинних середніх (без попередньої кластеризації) дозволяє зробити висновок, що запропонований підхід дає можливість підвищити точність прогнозування знаків приростів.

Література

1. Vercellis C. Business intelligence: data mining and optimization for decision making / C. Vercellis. – John Wiley & Sons, Ltd., Publication, 2009. – 417 p.
2. Box G.E.P. Time series analysis: forecasting and control / G.E.P. Box, G.M. Jenkins. – San Francisco: Holden-Day, 1976. – 575 p.
3. Brown Robert G. Statistical forecasting for inventory control [Текст] / R.G. Brown. – US: McGraw-Hill Inc., 1959. – 223 p.
4. Holt Charles C. Forecasting trends and seasonal by exponentially weighted averages [Текст] / C. Holt // International Journal of Forecasting. – 1957. – Vol.20, no.1. – P. 5-10.
5. Берзлев, А.Ю. Оценка эффективности прогнозирования и принятия решений на финансовом рынке [Текст] / А.Ю. Берзлев // «Problems of Computer Intellectualization», V.M. Glushkov Institute of Cybernetics of NAS of Ukraine. – Kyiv-Sofia: ITHEA, 2012. – С. 249-257.
6. Лукашин, Ю.П. Адаптивные методы краткосрочного прогнозирования временных рядов [Текст] / Ю.П. Лукашин. – М.: Финансы и статистика, 2003.– 416 с.
7. Singh S. Pattern Modeling in Time-Series Forecasting [Текст] / S. Singh // Cybernetics and Systems. An International Journal. – 2000. – Vol. 31, no. 1. – P. 49-65.
8. Fernández-Rodríguez F. Nearest-Neighbour Predictions in Foreign Exchange Markets [Текст] / F. Fernández-Rodríguez, S. Sosvilla-Rivero, J. Andrada-Félix // Fundacion de Estudios de Economia Aplicada. – 2002. – no.5. – 36 p.
9. Keogh, E. An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback [Текст] / E. Keogh, M. Pazzani // 4th Int'l Conference on Knowledge Discovery and Data Mining. 1998 Aug 27-31. – New York. – P. 239-241.
10. Берзлев, О.Ю. Адаптивні комбіновані моделі прогнозування біржових показників [Текст] / О.Ю. Берзлев, М.М. Маляр, В.В. Ніколенко // Вісник Черкаського держ. технолог. ун-ту. Серія: технічні науки. – 2011. – № 1. – С. 50-54.