УДК 004.6 **doi: 10.20998/2218-1849.2020.01.08**

Черенков Игорь Александрович, канд. тех. наук; Тел: +38 (050) 4 02 62 12; E-mail: sm261245@gmail.com
Национальный технический университет «Харьковский политехнический институт», ул. Кирпичёва, 2, г. Харьков, Украина, 61002

ПРОГНОЗИРОВАНИЕ НА ОСНОВЕ НОВОСТНОГО ПОТОКА ПОСРЕДСТВОМ АССОЦИАТИВНЫХ ПРАВИЛ

В динамически меняющейся рыночной среде актуальным является решение проблемы прогнозирования данных, представленных в виде временных рядов, в частности, прогнозирования цен. Среди существующих методов решения задачи ценового прогнозирования наибольшее распространение получили методы математической статистики, в частности экспоненциального сглаживания и авторегрессионные модели. Для этих методов характерно, что прогнозирование осуществляется на основе значений цены продукта, при этом факторы, влияющие на формирование цены, включается в прогноз опосредственно через исторические ценовые значения, что негативно сказывается на качестве прогноза, поскольку разные наборы внешних и внутренних факторов могут приводить к одинаковому значению цены.Точность прогнозов может быть повышена как за счёт оптимизации алгоритма прогноза на основе ассоциативных правил, так и за счёт оптимизации методов идентификации событий в новостном потоке. Экспериментально подтверждено превосходство методов ценового прогнозирования на основе новостного потока посредством ассоциативных правил над регрессионными методами. Обосновано, что его применение целесообразно в тех случаях, когда необходима максимальная точность прогнозов, т.к. суммарные затраты на прогноз, включая формирование множества ассоциативных правил, значительно больше, чем для регрессионных методов. Рассмотрены методы краткосрочного ценового прогнозирования на примере рынка полимеров, которые могут быть применены к рынку электроэнергетики.

Ключевые слова: *ценовое прогнозирования; ARIMA; экспоненциальное сглаживание; ассоциативные правила.*

Черенков Ігор Олександрович, канд. тех. наук; Тел: +38 (050) 4026212; E-mail: sm261245@gmail.com
Національний технічний університет «Харківський політехнічний інститут», вул. Кирпичова, 2, г. Харьків, Україна, 61002

ПРОГНОЗУВАННЯ НА ОСНОВІ НОВОСТНОГО ПОТОКУ ПОСЕРЕДНИЦТВОМ АСОЦІАТИВНИХ ПРАВИЛ

У динамічно мінливому ринковому середовищі актуальним є вирішення проблеми прогнозування даних, представлених у вигляді часових рядів, зокрема, прогнозування цін. Серед існуючих методів вирішення задачі цінового прогнозування найбільшого поширення набули методи математичної статистики, зокрема експоненціального згладжування і авторегресійні моделі. Для цих методів характерно, що прогнозування здійснюється на основі значень ціни продукту, при цьому чинники, що впливають на формування ціни, включається в прогноз опосередковано через історичні цінові значення, що негативно позначається на якості прогнозу, оскільки різні набори зовнішніх і внутрішніх факторів можуть призводити до однакового значенням цени. Точность прогнозів може бути підвищена як за рахунок оптимізації алгоритму прогнозу на основі асоціативних правил, так і за рахунок оптимізації методів ідентифікації подій в овостном потоці. Експериментально підтверджено перевагу методів цінового прогнозування на основі новинного потоку за допомогою асоціативних правил над регресійний методами. Обґрунтовано, що його застосування доцільно в тих випадках, коли необхідна максимальна точність прогнозів, тому що сумарні витрати на прогноз, включаючи формування безлічі асоціативних правил, значно більше, ніж для регресійних методов. Рассмотрени методи короткострокового цінового прогнозування на прикладі ринку полімерів, які можуть бути застосовані до ринку електроенергетики.

Ключові слова: *цінове прогнозування; ARIMA; експоненціальне згладжування; асоціативні правила.*

Cherenkov Igor Alexandrovich, cand. tech. Science; Tel: +38 (050) 4 02 62 12 ; E-mail: sm261245@gmail.com
National Technical University "Kharkiv Polytechnic Institute", str. Kyrpychova, 2, Kharkiv, Ukraine, 61002

FORECASTING BASED ON THE NEWSLETTER THROUGH ASSOCIATIVE RULES

In a dynamically changing market environment, it is relevant to solve the problem of forecasting data presented in the form of time series, in particular, price forecasting. Among the existing methods for solving the problem of price forecasting, the most widespread methods are mathematical statistics, in particular exponential smoothing and autoregressive models. It is typical for these methods that forecasting is carried out on the basis of product price values, while factors affecting the formation of prices are included in the forecast directly through historical price values, which negatively affects the quality of the forecast, since different sets of external and internal factors can lead to the same value of the price. The accuracy of forecasts can be improved both by optimizing the forecast algorithm based on associative rules, and by optimizing the methods for identifying events in ovostnom stream. The superiority of the methods of price forecasting based on news flow through associative rules over regression methods has been experimentally confirmed. It is proved that its use is advisable in those cases when the maximum accuracy of forecasts is necessary, because the total forecast costs, including the formation of many associative rules, are much larger than for the regression methods. Short-term price forecasting methods are considered on the example of the polymer market, which can be applied to the electricity market.

Key words: *price forecasting; ARIMA; exponential smoothing; associative rules.*

Введение

В динамически меняющейся рыночной среде актуальным является решение проблемы прогнозирования данных, представленных в виде временных рядов, в частности, прогнозирования цен. Задача анализа и прогнозирования временных рядов заключается в определении регулярной составляющей, включающей тренд и сезонную компоненту. Тренд является некоторой линейной или нелинейной зависимостью, отображающей главную закономерность в данных. Сезонная компонента отображает влияние циклических процессов на данные. Наибольший интерес для краткосрочного прогнозирования представляет периодическая, сезонная компонента. Эта компонента выражается зависимостью порядка k между каждым i-м и каждым (i-k)-м элементом. Не существует универсального способа определения регулярной составляющей временного ряда.

Постановка задачи. Сравним существующие способы краткосрочного прогнозирования, базирующиеся на математическом аппарате временных рядов, с подходами прогнозирования на основе новостного потока посредством поиска ассоциативных правил на примере рынка полимеров.

Основная часть. Среди существующих методов решения задачи ценового прогнозирования наибольшее распространение получили методы математической статистики, в частности экспоненциального сглаживания и авторегрессионные модели. Для этих методов характерно, что прогнозирование осуществляется на основе значений цены продукта, при этом факторы, влияющие на формирование цены, включается в прогноз опосредственно через исторические ценовые значения, что негативно сказывается на качестве прогноза, поскольку разные наборы внешних и внутренних факторов могут приводить к одинаковому значению цены.

Опишем группу методов экспоненциального сглаживания. В основе этих методов лежит построение экспоненциально взвешенных, усреднённых значений по всему временному ряду.

Так для исходного ряда $X = \{x_1,...,x_T\}$ метод экспоненциального сглаживания подразумевает построение ряда по следующей реккурентной формуле [1]:

$$F_t = \begin{cases} x_t, t = 1, \\ \alpha x_{t-1} + F_{t-1}(1-\alpha), t > 1, \end{cases} \quad (1)$$

где:

- F_t – сглаженный ряд;
- α – коэффициент сглаживания $\alpha \in (0,1)$, который задаёт уровень подавления колебаний и шума исходного ряда.

Метод экспоненциального сглаживания довольно часто используется для задач краткосрочного прогнозирования временных рядов, однако, у данного метода есть существенный недостаток. Область его применения сводится исключительно к краткосрочному прогнозированию, т.к. в модели не учитываются сезонные колебания и тренд.

Для учёта этих двух составляющих используют, среди прочих методов, модель Хольта-Уинтерса [2]. Для временного ряда $X = \{x_1,...,x_T\}$ будущее значение определяются по формуле:

$$\begin{aligned} F_{t+m} &= (s_t + mb_t)c_{t-L+((m-1) \pmod L)}, \\ s_t &= \alpha \frac{x_t}{c_{t-L}} + (1-\alpha)(s_{t-1} + b_{t-1}), \\ b_t &= \beta(s_t - s_{t-1}) + (1-\beta)b_{t-1}, \\ c_t &= \gamma \frac{x_t}{s_t} + (1-\gamma)c_{t-L}, \end{aligned} \tag{2}$$

где:

- s_t – сглаженное значение прогноза;
- b_t – составляющая тренда;
- c_t – сезонная компонента;
- α – параметр сглаживания модели $\alpha \in (0,1)$;
- β – параметр сглаживания тренда $\beta \in (0,1)$;
- γ – параметр сезонного сглаживания $\gamma \in (0,1)$.

Правильный подбор параметров (α, β, γ) определяет качество функционирования модели и её прогнозов.Другая часто используемая группа методов включает подходы, основанные на авторегрессионных моделях, для которых построение будущих значений ряда осуществляется по формуле:

$$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t, \tag{3}$$

где: ϕ_i – параметры модели; c – константа; ε_t – белый шум.

Задача исследователя заключается в расчете параметров.

Для моделирования случайных ошибок ряда используют модель скользящего среднего:

$$X_t = \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t, \tag{4}$$

где: θ_i – параметры модели; $\varepsilon_{t-q},...,\varepsilon_t$ – ошибки.

Производными моделями от вышеописанных являются авторегрессионные модели скользящего среднего ARMA (p, q) и ARIMA (p, d, q), содержащих p авторегрессионных составляющих и q скользящих средних:

$$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t \cdot \quad (5)$$

Модель ARIMA (p, d, q) предназначена для моделирования нестационарных процессов, разности временного ряда порядка d подчиняются модели ARMA (p, q) [2]:

$$\Delta^d X_t = c + \sum_{i=1}^{p} a_i \Delta^d X_{t-i} + \sum_{j=1}^{q} b_j \varepsilon_{t-j} + \varepsilon_t \quad (6)$$

На данном этапе развития информационных технологий, эффективнее будет построение ценовых прогнозов, основанных на новостном потоке, в результате чего сохраняется причинно-следственная связь между событием и его влиянием на цену. Подобное решение задачи прогнозирования на основе новостного потока может быть реализовано с помощью подходов на основе ассоциативных правил [3].

К недостаткам подходов на основе новостных потоков посредством ассоциативных правил следует отнести невозможность явного учёта тренда в прогнозах. Также подход прогнозирования с помощью ассоциативных правил предполагает предварительную добычу множества правил, что связано с дополнительными затратами. На данный момент существует довольно много алгоритмов поиска ассоциативных правил, главное отличие которых в быстродействии и эффективности используемой памяти.

Опишем подход прогнозирования на основе новостных потоков посредством ассоциативных правил. Для множества добытых правил $\{r_i\}, r_i \in R, i \in I$, где r_i – есть последовательность (набор) событий, предшествующая изменению цены, задача краткосрочного ценового прогнозирования на основе множества ассоциативных правил формулируется как задача правильной идентификации сложившейся рыночной ситуации, в виде соответствующей ей правила r_i. Каждому добытому правилу r_i в процессе добычи данных ставятся в соответствие два атрибута: s_i – поддержка, характеризующая абсолютную частоту появления правила в исходной выборке; c_i - достоверность, в данном случае вероятность возникновения ценового изменения при появлении набора событий из r_i. Проблематика данного подхода заключается в следующем: при идентификации сложившейся ситуации и выборе правила может возникнуть неопределённость, т.к. добываемые ассоциативные правила r_i обладают разной достоверностью и поддержкой, что делает определение сложившейся ситуации нетривиальной задачей. Правило может иметь как очень высокую поддержку, т.е. быть очевидным правилом, так и напротив иметь очень низкую, являясь неочевидным правилом. Как следствие качество прогнозов напрямую зависит от используемого алгоритма идентификации сложившейся ситуации. Предлагается использовать следующий алгоритм:

Для заданного уровня достоверности C и поддержки S отобрать все правила, для которых $c_i \geq C$, $s_i \geq S$ соответственно. Для выбранного времени (дня) $t \in T$ найти наиболее подходящее правило на основе следующей последовательности.

1. Задать $n = 1$;
2. Отобрать множество правил $\{r_i\}_t^n$, удовлетворяющих текущей ситуации на рынке в момент времени $t \in T$;
3. Если найдено только одно правило, то сформировать прогноз;

4. Если найдено больше одного правила, сравнить прогнозные значения, если прогноз направлен одну сторону (вверх/вниз), то сформировать суммарный прогноз;
5. Если найдено больше одного правила и прогнозные значения противоречивы, то сформировать из множества $\{r_i\}_t$ множество $\{r_i\}_t^c$ путём исключения менее достоверных правил в соответствие с условием:

$$\tilde{c}_i^{\max} - c_i \le \Delta C, \tag{7}$$

где: $\tilde{c}_i^{\max}$ – правило с наибольшим параметром достоверности в текущем множестве правил; $\Delta C = \alpha \tilde{c}_i^{\max}$ – допустимая погрешность; α – коэффициент погрешности.

Аналогично сформировать множество правил $\{r_i\}_t^s$ путём исключения менее неочевидных правил в соответствие с условием:

$$\tilde{s}_i^{\max} - s_i \le \Delta S, \tag{8}$$

где: $\tilde{s}_i^{\max}$ – правило с наибольшим параметром поддержки в текущем множестве правил; $\Delta S = \beta \tilde{s}_i^{\max}$ – допустимая погрешность; β – коэффициент погрешности.

6. Задать n=n+1, сформировать новое множество $\{r_i\}_t^n = \{r_i\}_t^s \bigcup \{r_i\}_t^c$.
7. Если $\{r_i\}_t^n \ne \{r_i\}_t^{n-1}$ перейти к п.3. Иначе в соответствии с принципом бритвы Окамма выбрать наиболее очевидное правило с $\tilde{c}_i^{\max}$ из $\{r_i\}_t^s$ и сформировать прогноз.

Следует отметить, что значения C и S непосредственно влияют на качество прогнозов, и определение их оптимальных значений является отдельной исследовательской задачей.

Экспериментальная часть. Вышеописанные методы краткосрочного прогнозирования были применены к временным рядам цен рынка полимеров Российской Федерации. В качестве входной информации использовались выборка ценовых значений по ПВХ за 2010 – 2011 гг. и соответствующий ей новостной поток. Размер выборки ценовых значений составил 800 записей, новостной выборки – 2700, при этом в качестве обучающей выборки были взяты первые 600 значений цены и соответствующие им 2100 новостных событий. Из оставшихся значений были сформированы две контрольные выборки. В рамках краткосрочного прогнозирования горизонт прогноза цены составлял +1 день.

Качество работы методов прогнозирования оценивалось на основе моделей, построенных с минимизированным значением функции правдоподобия. Множество ассоциативных правил было получено с помощью алгоритма SPADE [4]. Для метода прогнозирования на основе ассоциативных правил были использованы следующие значения C = 80 % и S = 7.

Для оценки качества прогнозов использовался критерий MAPE, отражающий усреднённую абсолютную величину ошибок в процентах, в соответствии с формулой:

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|p_i - \tilde{p}_i|}{p_i} \cdot 100\%, \tag{9}$$

где: p_i – оригинальное значение ряда; $\tilde{p}_i$ – прогнозируемое значение; N – размер ряда.

Полученные значения ошибок прогнозов MAPE позволяют оценить качество работы методов прогнозирования. Экспериментальные значения MAPE приведены в табл.1.

Таблица 1

Экспериментальные значения MAPE

	Выборки	
Метод	Контрольная 1	Контрольная 2
Экспоненциальное сглаживание	2,2%	21,5%
ARIMA	19,4%	19,1%
Ассоциативные правила	12,9%	13,2%

Выводы

1. Таким образом, прогнозы, получаемые на основе ассоциативных правил на 6% точнее, чем прогнозы на основе методов регрессионного анализа. Большая точность достигается благодаря тому, что прогнозы на основе новостных потоков посредством ассоциативных правил позволяют непосредственно включать события, влияющие на формирование цены, в прогнозное значение, в то время как регрессионные методы, включают эти события опосредственно.
2. Точность прогнозов может быть повышена как за счёт оптимизации алгоритма прогноза на основе ассоциативных правил, так и за счёт оптимизации методов идентификации событий в новостном потоке.
3. В целом, метод прогнозирования на основе новостного потока посредством ассоциативных правил является перспективным и требует дальнейших исследований. Его применение целесообразно в тех случаях, когда необходима максимальная точность прогнозов, т.к. суммарные затраты на прогноз, включая формирование множества ассоциативных правил, значительно больше, чем для регрессионных методов.

Список использованной литературы:

1. Афанасьев В. Н. Анализ временных рядов и прогнозирование / В. Н. Афанасьев, М. М. Юзбашев // М. – Инфра-М, 2010. – 320 с.
2. Керимов А. К. Анализ и прогнозирование временных рядов / А. К. Керимов // Издательство Российского Университета дружбы народов: М. – 2005. – 140 с.
3. Черенков И. А. Автоматический поиск данных из новостей на примере рынка полимеров / И. А. Черенков // Системы обработки информации: Харьков. – 2011. – № 8. – С. 156 – 159.
4. Zaki M. Spade: an Efficient Algorithm for Mining Frequent Sequences / M. Zaki // Machine Learning.: Kluwer Academic Publishers. – 2001. – Vol. 42. – P. 31 – 60.

Referenses:

1. Afanasj'ev V. N. Analyz vremennыkh rjadov y proghnozyrovanye / V. N. Afanasj'ev, M. M. Juzbashev // M. – Ynfra-M, 2010. – 320 s.
2. Kerymov A. K. Analyz y proghnozyrovanye vremennыkh rjadov / A. K. Kerymov // Yzdateljstvo Rossyjskogho Unyversyteta druzhbы narodov: M. – 2005. – 140 s.
3. Cherenkov Y. A. Avtomatycheskyj poysk dannыkh yz novostej na prymere rынka polymerov / Y. A. Cherenkov // Systemы obrabotky ynformacyy: Kharjkov. – 2011. – # 8. – S. 156 – 159.
4. Zaki M. Spade: an Efficient Algorithm for Mining Frequent Sequences / M. Zaki // Machine Learning.: Kluwer Academic Publishers. – 2001. – Vol. 42. – P. 31 – 60.

Прийнята до друку 18.12. 2019