УДК 004.9 42

О. П. Гожий, канд. техн. наук

ПІДХІД ДО ОЦІНЮВАННЯ НЕВИЗНАЧЕНОСТЕЙ В ЗАДАЧАХ ПРОГНОЗУВАННЯ

Анотація. *В роботі наведена класифікація невизначеностей які мають місце при вирішенні задач прогнозування. Невизначеності при вирішенні задач прогнозування проявляються у вигляді імовірнісних розподілів, у вигляді суб'єктивних ймовірностей і у вигляді інтервальних невизначеностей. Розглянуто особливості статистичних, структурних невизначеностей, невизначеностей при відсутності спостережень, невизначеностей амплітудного типу. Окрім того розглянуто невизначеність параметрів прогнозної моделі. Запропоновані методи їх подолання*

Ключові слова. *Невизначеність, прогнозування, ймовірносний розподіл, інтервальна невизначеність, параметри моделі, авторегресія, авторегресія з ковзним середнім, фільтр Калмана*

O. Gozhyi, PhD.

APPROACH TO THE EVALUATION OF UNCERTAINTY IN FORECASTING PROBLEMS

Abstract. *The paper shows the classification of uncertainties in solving problems of forecasting. Uncertainties in solving prediction manifested in the form of probability distributions in the form of subjective probabilities in the form of interval uncertainty. The features of statistical and structural uncertainties, uncertainties in the absence of observations, the uncertainty of the amplitude type. And also consider the uncertainty of parameters predictive model. The proposed methods of overcoming them.*

Keywords *The uncertainty, forecasting, probability distribution, interval uncertain, parameters model, autoregressive, autoregressive moving average of, Kalman filter*

А. П .Гожий, канд. техн. наук

ПОДХОД К ОЦЕНИВАНИЮ НЕОПРЕДЕЛЕННОСТЕЙ В ЗАДАЧАХ ПРОГНОЗИРОВАНИЯ

Аннотация. *В работе приведена классификация неопределенностей которые имеют место при решении задач прогнозирования. Неопределенности при решении задач прогнозирования проявляются в виде вероятностных распределений, в виде субъективных вероятностей и в виде интервальных неопределенностей. Рассмотрены особенности статистических, структурных неопределенностей, неопределенностей при отсутствии наблюдений, неопределенностей амплитудного типа. Кроме того рассматривались неопределенности параметров прогнозной модели. Предложенны методы их преодоления.*

Ключевые слова. *Неопределенность, прогнозирования, вероятностное распределение, интервальная неопределенность, параметры модели, авторегрессия, авторегрессия со скользящим средним, фильтр Калмана*

Вступ. При досліджені складних систем завжди мають місце невизначеності різного типу. Це залежить від різних причин, головна з яких це нестача інформації. Функціонування складних систем має залежність від цілого ряду причин, які можливо систематизувати у вигляді схеми невизначеностей різного типу [2– 3]. Але детальний опис невизначеностей для вирішення кожної конкретної задачі має свої особливості. Окрім того методи подолання невизначеності працюють ефективно коли системно використовуються декілька підходів. В статті розглядаються типи невизначеностей та методи їх подолання при вирішенні задач прогнозування

Мета роботи. Дослідити методи оцінювання різних типів невизначеностей при вирішенні задач прогнозування, та запропонувати методи їх подолання.

Основна частина. При вирішенні задач прогнозування та в подальшому при прийнятті рішень необхідно враховувати, що невизначеність проявляється по різному:

• у вигляді імовірнісних розподілів (розподіл випадкової величини, точно відомо, але невідомо яке конкретно значення прийме випадкова величина);

• у вигляді суб'єктивних ймовірностей (розподіл випадкової величини невідомий, але відомі ймовірності окремих подій, задані експертним шляхом);

• у вигляді інтервальної невизначеності (розподіл випадкової величини невідомий, але відомо, що вона може приймати будь-яке значення із області визначення.

Опис та обробку невизначеностей при при вирішенні задач прогнозування можна представляти по різному, в залежності від їх типу.

Обробка статистичних невизначеностей. Статистичні невизначеності зустрічаються найчастіше. Вони пов'язані з розвитком моделей та оцінкою прогнозів та викликаються такими факторами:

•похибки вимірювань (шум), який є практично у всіх випадках збору даних незалежно від їх походження;

•стохастичні зовнішні збурення, що, як правило, негативно впливають на досліджуваний процес;

•значні випадкові викиди;

•мультіколінеарність, яка вимагає спеціальних методів обробки даних для зменшення ступеня взаємної кореляції між окремими часовими рядами.

Найчастіше засобами, які використовуються для відокремлення шуму і зовнішніх стохастичних збурень, є цифрові та оптимальні фільтри. Серед них фільтр Калмана є самим розповсюдженим інструментом обробки даних [8]. Цифрові фільтри (ЦФ) допомагають вибрати для подальшої обробки потрібну смугу частот за допомогою лінійних сруктур, які можуть бути представлені у вигляді рівнянь авторегресії (AR), або авторегресії з ковзним середнім (ARMA) такого типу:

$$y(k)=\sum_{i=1}^{p} a_i y(k-i)+\sum_{j=1}^{q} b_j x(k-j), \quad (1)$$

де $x(k)$ – вхідна послідовність вимірювань; $y(k)$ – відфільтрована вихідна послідовность; $\theta=[a_1 ... a_p\, b_1 ... b_q]^T$ – вектор параметрів, який визначає смугу частот фільтра; p і q – порядки авторегресії та ковзного середнього, відповідно. Правильно спроектований адаптивний фільтр Калмана забезпечує можливість оцінювання коваріацій стохастичних збурень і шумів вимірювань, а також оцінку короткострокових прогнозів. Оптимальне проектування фільтра вимагає модель процесу (системи) у просторі станів:

$$\mathrm{x}(k)=A(k,k-1)x(k-1)+B(k,k-d)u(k-1)+w(k) \quad (2)$$

$$z(k)=H\,x(k)+v(k), \quad (3)$$

де $\mathrm{x}(k)$ – вектор простору станів, що складається з n елементів; $u(k)$ – це m-вимірний вектор управлінь; $w(k)$ – вектор випадкових зовнішніх збурень; $A(k,k-1)$ – це матриця ($n\times n$) динаміки системи; $B(k)$ – це матриця ($m\times n$) коефіцієнтів управління; k – дискретний час, пов'язаний з безперервним часом t через період дискретизації $T_s : t=kT_s$; d – час затримки системи по входу; $z(k)$ – це вектор вимірювання з r елементами; H = [1 0] матриця коефіцієнтів виміру; $v(k)$ – це вектор шуму (похибок вимірів). Подвійний аргумент часу $(k,\,k-1)$ означає, що змінна з таким аргументом використовується у момент часу k, але її значення формується на основі попередніх вимірів до $(k,\,k-1)$.

Основною перевагою моделі (2) та (3) є те, що вона включає в себе дві складових випадкових процесів $w(k)$ і $v(k)$; отже, така модель більш адекватно відповідає реальності, ніж лінійна регресія. Основне завдання оптимального фільтра полягає у обчисленні векторних оцінок стану разом з урахуванням статистичних характеристик (коваріацій) двох випадкових процесів. Такий підхід забезпечує можливість для удосконалення оцінок вектора стану. Основне рівняння оптимального фільтра:

$$\begin{aligned}\hat{\mathbf{x}}(k)&=A(k)\hat{\mathbf{x}}(k-1)+B(k)u(k-1)+\\&+K(k)[z(k)-H(k)A(k)\hat{\mathbf{x}}(k-1)]\end{aligned}, \quad (4)$$

де $K(k)$ – оптимальний коефіцієнт фільтра у матричної формі; $H(k)$ – матриця коефіцієнтів виміру (дуже часто має діагональний вигляд).

Коефіцієнт обчислюється шляхом мінімізації функціоналу

$$J=\min_{\mathbf{K}} E\left\{[\hat{\mathbf{x}}(k)-x(k)]^T[\hat{\mathbf{x}}(k)-x(k)]\right\}, \quad (5)$$

де $x(k)$ точне значення вектора стану, яке може бути знайдено за допомогою детермінованої частини моделі (2). У лінійному дискретному випадку коефіцієнт досить легко обчислюється шляхом знаходження розв'язку рівняння Ріккаті.

Таким чином, оптимальний фільтр забезпечує можливість для зниження невизначеності у вигляді впливу двох випадкових процесів $w(k)$ і $v(k)$, а також оцінювання невимірюваних компонентів вектора стану, коли відповідні компоненти коваріаційних матриць відомі. В задачах прогнозування особливо корисні адаптивні версії фільтрів, які є найбільш придатними для практичних застосувань (при роботі в режимі реального часу і при пакетній обробці). Вони підходять для повторного оцінювання системи матриць $A(k)$ і $B(k)$, а також коваріацій двох випадкових процесів [7].

Невизначеность через відсутність спостережень. Для її подолання у часових рядах часто використовуються такі методи:

– просте усереднення, коли це можливо (коли тільки кілька значень відсутні);

– формування прогнозних оцінок за допомогою побудованої моделі з використанням наявних вимірювань;

– формування відсутньої оцінки за допомогою розподілів і їх параметрів, які визначаються на основі наявної частини вибірки даних;

– використання методів оптимізації, скажімо, відповідні форми ЕМ-алгоритмів [6];

– експоненційне згладжування.

Найпростішою моделлю, яка може бути використана для оцінювання прогнозів, є AR (1):

$$y(k)=a_0+a_1\,y(k-1)+\varepsilon(k), \quad (6)$$

де a_0, a_1 параметри моделі; випадковий процес $\varepsilon(k)$ враховує можливі невизначеності структури моделі (відсутність відповідних регресорів, зовнішні випадкові збурення, похибки обчислення параметрів), а також можливі похибки вимірів.

Якщо параметри a_0, a_1 відомі, можливо обчислити прогноз на крок вперед, як умовне математичне очікування у вигляді:

$$\begin{aligned}\hat{y}(k+1)&=E_k[y(k+1)]=E_k[y(k+1)\,|\,y(k),y(k-1),.\\..,\varepsilon(k),\varepsilon(k-1),...]&=a_0+a_1E_k[y(k)]=a_0+a_1y(k)\end{aligned}, \quad (7)$$

оскільки $y(k)$ на момент k приймає відоме значення. Ітеративно можна отримати функцію прогнозування на довільну кількість кроків S [2]:

$$\begin{aligned}\hat{y}(k+s)&=E_S[y(k+s)]=a_0\left(\sum_{i=0}^{S-1}a_1^{\,i}\right)+\\&+a_1^{\,S}y(k)=a_0\sum_{i=0}^{S-1}a_1^{\,i}+a_1^{\,S}y(k)\end{aligned} \quad (8)$$

Послідовність прогнозних оцінок $\{\hat{y}(k+i)\}\ i=1,...,\ s$, буде збіжною, якщо $|a_1|<1$:

$$\lim_{s\to\infty} E_k[y(k+s)] = \frac{a_0}{1-a_1}, \quad (9)$$

Останній вираз означає, що для стаціонарного AR або ARMA процесів оцінки умовних прогнозів асимптотично $(s\to\infty)$ збігаються до безумовного середнього (довгостроковий прогноз). Подальше зменшення невизначеності можливе завдяки застосуванню кількох методів прогнозування і комбінування окремих прогнозів з використанням відповідних вагових коефіцієнтів.

Невизначеність параметрів моделі. Зазвичай невизначеності оцінок параметрів моделі, таких як упередженість і непослідовність пов'язані з низькою інформативністю даних або даними, що не відповідають нормальному розподілу. Така ситуація може мати місце у випадку мультіколінеарності регресорів та істотного впливу нелінійності процесу, що з деяких причин не були прийняті до уваги при побудові моделі. Коли об'єм вибірки даних незадовільний для побудови моделі, він може бути розширений за рахунок застосування імітаційного моделювання або спеціальних методів, таких як метод групового врахування аргументів (МГУА). Якщо дані не відповідають нормальному розподілу, то може бути використано метод максимальної правдоподібності або відповідні процедури Монте-Карло для ланцюгів Маркова [4; 10]. Останні методи можуть бути застосовані з незначними обчислювальними витратами, коли кількість параметрів невелика.

Структурна невизначеність. При використанні в задачах прийняття рішень структура моделі повинна практично завжди оцінюватись за допомогою даних. Це означає, що елементи структури моделі майже завжди приймають лише наближені значення. Коли модель побудована для прогнозування, аналізуються кілька кандидатів і вибирають кращу з них за множиною статистик якості моделі. Як правило, використовуються такі методи боротьби із структурними невизначеностями: поступове покращення порядку моделі (AR (*P*) або ARMA (*P, Q*)), застосування адаптивного підходу до моделювання та автоматичного пошуку «кращих» структур з використанням складних статистичних критеріїв якості; адаптивна оцінка часу затримки і типу розподілу даних та його параметрів; опису виявлених нелінійностей процесів за альтернативними аналітичними формами з подальшою оцінкою адекватності моделі і якості прогнозів. Комплексний критерій для моделювання і прогнозування може виглядати, наприклад, таким чином:

$$J = \left|1-R^2\right| + \alpha \ln\left[\sum_{k=1}^{N} e^2(k)\right] + \left|2-DW\right| + \beta \ln(MAPE) + U \to \min_{\hat{\theta}_i}, \quad (10)$$

де R^2 коефіцієнт детермінації; $\sum_{k=1}^{N} e^2(k)=\sum_{k=1}^{N}[y(k)-\hat{y}(k)]^2$ – сума квадратів похибок моделювання; DW – статистика Дарбіна-Уотсона; $MAPE$ – це середня абсолютна похибка у відсотках (на один крок прогнозування); U – це коефіцієнт Тейла, який характеризує корельованість похибок моделі; α, β – підібрані відповідним чином вагові коефіцієнти; $\hat{\theta}_i$ – це вектор параметрів для *i*-ї моделі-кандидата.

Мінімізувати критерій J можна на множині цих векторів, тобто один вектор параметрів для однієї моделі-кандидата. Потім із кандидатів обирається краща модель. Критерій такого типу використовується для автоматичного вибору кращої моделі-кандидата. Можливе подальше удосконалення складних критеріїв. Але при побудові критерію важливо не добавляти велику кількість членів у праву сторону рівняння (10).

Невизначеність амплітудного типу. Використання випадкових (тобто з випадковою амплітудою або рівнем) змінних, які не піддаються вимірюванню, приводить до необхідності використання нечітких множин для опису таких ситуацій. Змінні з випадковою амплітудою можуть бути описані за деяким розподілом ймовірностей, якщо виміри можливі і їх можна проаналізувати за прийнятний проміжок часу. Однак, деякі змінні не можуть бути виміряні в принципі. У таких ситуаціях для змінної можна призначити множину можливих значень у лінгвістичній формі (наприклад: значення змінної = {*дуже низьке*, *низьке*, *середнє*, *високе*, *дуже високе*}). Існує необхідний набір математичних операцій, які повинні застосовуватися до обробки таких нечітких змінних.

Імовірнісні невизначеності. Використання випадкових величин приводить до необхідності побудови імовірнісних розподілів та їх застосування в процедурах прийняття рішень. Зазвичай значення спостережень відомі лише наближено, хоча відомі межі для цих значень. Розподіли ймовірностей дуже корисні для опису таких ситуацій. При роботі з дискретними результатами присвоюються ймовірності конкретним даним, використовуючи функцію маси. Вона показує, скільки потрібно «ваги» (або маси), щоб призначити її кожному спостереженню або виміру. Відповідь на запитання про значення того чи іншого результату буде його масою. Для більш глибокого розуміння того, що відбувається, необхідно використати аксіоми Колмогорова. Якщо дві або більше змінних аналізується одночасно, то необхідно будувати та використовувати спільні розподіли. Спільні розподіли дозволяють обчислювати умовні ймовірності. Дуже корисним для виконання імовірнісних обчислень є поняття умовної незалежності:

$$P(x,y\,|\,z)=P(x\,|\,z)\ P(y\,|\,z),$$

де x і y незалежні події. Незалежності є дуже зручними для аналізу, але події повинні бути дійсно незалежними [3]. Гранична ймовірність $P(B)$, може бути обчислена за допомогою умовних ймовірностей. Умовна ймовірність того, що відбувається в цілому, може бути отримана з умови: $P(A/B)=P(B/A)P(A)/P(B)$.

Щоб вирішити проблему опису невизначеності, можна розглянути різноманітні байєсівські моделі, які представлені у вигляді так званого байєсівського формалізму програмування. Ця множина моделей включає в себе байєсівські мережі [9], динамічні мережі Байєса (ДБН), приховані моделі Маркова (дискретні і безперервні) [4], фільтри Калмана [8], байєсівські карти та інше.

При цьому структура процесу обробки інформації включає в себе такі етапи:

1. Опис проблеми та формулювання основного питання у вигляді: $P(X_i / D, Kn)$, де X_i визначає одну змінну, тобто те, що має бути оціненено з використанням специфічних механізмів логічного висновку; D, Kn – дані і знання стосовно подій, що аналізуються.

2. Використання попередніх знань та експериментальних даних для встановлення типової структури та визначення параметрів.

3. Вибір та застосування відповідної методики логічного висновку.

4. Перевірка якості кінцевого результату.

Ця схема також придатна для використання в режимі адаптації параметрів моделі. Обчислюються нові оцінки експериментальних даних або данних від експертів, які можуть бути використані, наприклад, для оцінки попередніх розподілів або структури Баєсівської мережі (БМ) або динамічної Баєсівськоі мережі (ДБМ).

Висновки. Опис наведених типів невизначеностей і запропонованих методів їх подолання, дозволяє більш якісно враховувати невизначеності у задачах прогнозування і будувати на основі їх обробки більш якісні прогнози. Представлені підходи до опису та оцінювання невизначеностей різного типу дозволяють системно підійти до вирішення задач прогнозування і ефективно використовувати їх при створенні систем підтримки прийняття рішень.

Список використаної літератури

1. Бидюк П. И. Построение и методы обучения байесовских сетей / П. И. Бидюк, А. Н. Терентьев // Таврический вестник информатики и математики. – Симферополь : – 2004. – № 2. – С. 139 – 153.

2. Бідюк П. І. Прогнозування фінансових процесів на біржі з використанням індикаторів технічного аналізу / П. І. Бідюк, А. В. Федоров //Наукові вісті НТУУ «КПІ». – К. : – 2009. – № 2. – С. 59 – 64.

3. Бідюк П. І. Побудова системи адаптивного прогнозування фінансово-економічних та її застосування / П. І. Бідюк, А. В. Федоров / Збірник наукових праць Миколаївського державного університету ім. П. Могили. – Миколаїв : – 2009. – Вип.104. – Том 117. – С. 119 – 129. Url: *http://lib.chdu.edu.ua/pdf/naukpraci /computer /2009/ 117-104-12 p.*

4. Шведов А. С. О методах Монте-Карло с цепями Маркова / А. С. Шведов // Экономический журнал ВШЭ. – М. : – 2010. – Т. 14. – № 2. – С. 227 – 243

5. Cooper G.F., (1990), The Computational Complexity of Probabilistic Inference Using Bayesian Belief Networks, *Artificial Intelligence*, Vol. 42, No. 2-3, pp. 393 – 405

6. Kasitskij A., Bidyuk P., and Gozhyi A., (2014), Effective Expectation Maximization Algorithm Implementation Using Multicore Computer Systems, *IAPGOS,* 4(4), pp. 35 – 37 (In English). Url: *http://e-iapgos.pl/abstracted.php?level =4&id _issue=876277.*

7. Neal P., and Rao T.S., (2007), MCMC for Integer-valued ARMA, Processes, *Journal of Time Series Analysis*, 28, pp. 92 – 110.

8. Strid I., and Walentin K., (2009), Block Kalman Filtering for Large-Scale DSGE Models*, Computational Economics, Springer,* Vol. 33 (3), pp. 277 – 304

9. Zgurowskii M.Z., Bidyuk P., and Terentyev A.M., (2008), Methods of Constructing Bayesian Networks Based on Scoring Functions, *Cybernetics and System Analysis*, Vol. 44, No. 2, pp. 219 – 224.

10. Winters-Hilt S., (2006), Hidden Markov Model Nariants and their Application, *BMC Bioinformatics,* No. 7 (2), pp. 1 – 14.

Отримано 17.05.2015

References

1. Bidyuk P.I., and Terentyev A.N. Postroenie i metody obuchenia bayesovich setey [Construction and Training Methods Bayesian Networks], (2004), *Tauride Journal of Computer Science and Mathematics*, *Publ.,* Simpferopol, Ukraine, No. 2, pp. 139 – 153. (In Russian).

2. Bidyuk P.I, and Fedorov A.V. Prognozirovanie finansovich procesjv na birgi z vikoristanyam technicheskogo analizu [Predicting Financial Processes on the Stock Exchange Using Technical Analysis Indicators], (2009), *Research news National Technical University "KPI"*, *Publ.,* Kiev, Ukraine, No. 2, pp. 59 – 64 (In Ukrainian).

3. Bidyuk P.I, and Fedorov A.V. Pobudova systemi adaptivnogo prognozirovania finansovo-ekonomi-shnich pokaznikiv ta ii zastosuvannya [Building a System of Adaptive Forecasting Financial, Economic and Its Application], (2009), *Mogila State University, Publ.,* Nikolaev, Ukraine, No. 117, pp. 119 – 129 (In Ukrainian), available at: Url: *http://lib.chdu.edu.ua/pdf/naukpraci /computer /2009/ 117-104-12 p.*

4. Schvedov A.S. O metodach Monte-Karlo s cepyami Markova [About Monte Carlo Markov Chains], (2010), *Economic Journal Higher School of Economics, Publ.,* Moscow, Russian Federation, Vol. 14, No. 2, pp. 227 – 243 (In Russian), available at: Url: *http://library.hse.ru/e-resources/HSE_economic journal /content.htm ?vo-lume=14&issue=2.*

5. Cooper G.F., (1990), The Computational Complexity of Probabilistic Inference using Bayesian Belief Networks, *Artificial Intelligence*, Vol. 42, No. 2-3, pp. 393 – 405

6. Kasitskij A., Bidyuk P., and Gozhyi A., (2014), Effective Expectation Maximization Algorithm Implementation using Multicore Computer Systems, *IAPGOS,* 4(4), pp. 35 – 37 (In English), available at: Url: *http://e-iapgos.pl/abstracted.php?level =4&id _issue=876277.*

7. Neal P., and Rao T.S., (2007), MCMC for Integer-valued ARMA Processes, *Journal of Time Series Analysis,* 28, pp. 92 – 110.

8. Strid I., and Walentin K., (2009), Block Kalman Filtering for Large-Scale DSGE Models*, Computational Economics, Springer,* Vol. 33 (3), pp. 277 – 304.

9. Zgurowskii M.Z., Bidyuk P., and Terentyev A.M., (2008), Methods of Constructing Bayesian Networks Based on Scoring Functions, *Cybernetics and System Analysis,* Vol. 44, No. 2, pp. 219 – 224 (In English).

10. Winters-Hilt S., (2006), Hidden Markov Model Variants and their Application, *BMC Bioinformatics,* No. 7 (2), pp. 1 – 14.

Гожий
Олександр Петрович,
канд. техн. наук, доцент, доцент кафедри інформа-ційних технологій і прог-рамних систем Чорноморсь-кого державного ун-ту
ім. П. Могили,
вул.68 Десантників, 10, Миколаїв, Україна, 54003,
+38(067)7501247.
E-mail: alex_daos@mail.ru