Комп’ютерне моделювання на основі стохастичних моделей генерації штучних даних

Файнзільберг Леонід
кафедра біомедичної кібернетики
Національний технічний університет «КПІ імені Ігоря Сікорського»
Київ, Україна
fainzilberg@gmail.com

Computer modeling using stochastic models of artificial data generation

Fainzilberg Leonid
Department of Biomedical Cybernetics
Igor Sikorsky Kyiv Polytechnic Institute
Kyiv,Ukraine
fainzilberg@gmail.com

***Анотація* — розвивається спосіб експериментальної оцінки ефективності комп'ютерних алгоритмів шляхом використання стохастичних моделей генерації штучних даних, що мають статистичні характеристики адекватні реальним спостереженням. Показано, що такий підхід надає змогу отримати інформацію, яка недоступна під час оброблення реальних даних, що спостерігаються в умовах спотворень.**

***Abstract* — the method of experimental evaluation of the effectiveness of computer algorithms is being developed by using stochastic models of artificial data generation, which have statistical characteristics adequate to real observations. It is shown that this approach makes it possible to obtain information that is not available during the processing of real data observed under distortion conditions.**

Ключові слова — стохастична модель; генерація штучних даних; алгоритм навчання; діагностична ознака

Keywords — stochastic model; generation of artificial data; distortion; learning algorithm; diagnostic feature

I. ВСТУП

Нагадаємо, що загальний термін «технологія» поєднує два поняття – «техно» (грец. «téchnë» – мистецтво) та «логія» (грец. «logos» – наука). Тобто технологія це мистецтво перетворити деяку сировину в продукт, а наука технології полягає у виборі найбільш ефективних методів такого перетворення. З цього випливає, що наука створення інформаційних технологій (ІТ) полягає в побудові найбільш ефективних комп’ютерних алгоритмів, що реалізують окремі етапи ІТ.

Часто неможливо формально побудувати ефективний комп’ютерний алгоритм, ґрунтуючись на мінімумі (максимумі) деякого критерію $\Im$. В таких ситуаціях не залишається іншого, як будувати алгоритм неформально та оцінювати його ефективність експериментально.

Для проведення таких досліджень не завжди доцільно, а іноді і неможливо спиратися на реальні спостереження. Тоді використовують штучні дані, які генеровані за допомогою математичних моделей.

Саме такий підхід надав нам змогу свого часу отримати новий результат під час дослідження модифікованого методу оптимальної зупинки, який, на відміну від традиційного, передбачає послідовний пошук найкращої альтернативи з заданою поступкою [1].

Продемонструємо результативність використання генеративних моделей на прикладах розв’язування двох інших задач.

II. ПОРІВНЯЛЬНИЙ АНАЛІЗ ШВИДКОДІЇ АЛГОРИТМІВ РОЗЕНБЛАТА ТА КОЗИНЦЯ

Нехай у N-вимірному просторі ознак $x_1,...x_N$ задана скінчена вибірка спостережень

$$X = \{(x_1^{(N)}, y_1),...,(x_K^{(N)}, y_K)\}$$

з відомою приналежністю до класів V_1 і V_2, що виражена в формі

$$y_j = \begin{cases} +1, & \text{якщо } x_j^{(N)} \in V_1, \\ -1, & \text{якщо } x_j^{(N)} \in V_2, \end{cases} \qquad j = 1,...,K,.$$

де $x_j^{(N)} \triangleq (x_1,...,x_N)$, $j = 1,...,K$, а K – кількість елементів.

Передбачається, що спостереження класів V_1 і V_2, можуть бути розділені лінійною дискримінантною функцією $D(x)=\langle w,x\rangle \triangleq \sum_{i=0}^{N} w_i x_i$, в якій запис $\langle w,x\rangle$ означає скалярний добуток $(N+1)$-вимірних векторів параметрів $w=(w_0,w_1,...,w_N)$ та розширених векторів $x=(1,x_1,..x_N)$.

Задача полягає в визначені вектора параметрів $w_0,w_1,...,w_N$ за якими спостереження класів V_1 і V_2 можна розділити.

Відомо два алгоритми розв'язування такої задачі: алгоритм Розенблата та алгоритм Козинця [2].

Алгоритм Розенблата передбачає такі кроки:

Крок 1. Довільним чином назначають початкове значення вектора $w^{(0)}$.

Крок 2. Послідовно вибирають спостереження $x_j^{(t)}=(x^{(N)},1)$, $j=1,...,K$ з навчальної вибірки і визначають значення дискримінантної функції $D\left(w^{(t-1)},x_j^{(t)},\right)$ для поточного значення вектора $w^{(t-1)}$, $t=1,2,...$.

Крок 3. Якщо виконується умова $D\left(w^{(t-1)},x_j^{(t)}\right)y_j>0$, тобто знаки дискримінантної функції $D\left(w^{(t-1)},x_j^{(t)},\right)$ і індикаторної змінної y_j співпадають, то корекція $w^{(t-1)}$ не здійснюється. Якщо ж виконується умова

$$D\left(w^{(t-1)},x_j^{(t)}\right)y_j<0, \qquad (1)$$

що свідчить про помилку під час класифікації $x_j^{(t)}$, то вектор параметрів модифікують таким чином:

$$w^{(t)}=\begin{cases} w^{(t-1)}+\gamma x_j^{(t)}, & \text{якщо } D\left(w^{(t-1)},x_j^{(t)}\right)<0 \text{ та } y_j=+1, \\ w^{(t-1)}-\gamma x_j^{(t)}, & \text{якщо } D\left(w^{(t-1)},x_j^{(t)}\right)>0 \text{ та } y_j=-1, \end{cases}$$

де $0<\gamma<1$ – константа, що визначає темп корекції.

Кроки 2-3 повторюють доки всі точки вибірки будуть правильно класифіковані.

Відмінність алгоритму навчання Козинця полягає в тому, що на кожному кроці $t=1,2,...$ проводиться пошук такого спостереження $x_j^{(t)}=(1,x^{(N)})$, $j=1,...,K$ вибірки, яке за поточним значенням $w^{(t-1)}$ неправильно класифікується. Якщо таких спостережень немає $\forall j=1,...,K$, то алгоритм завершує свою роботу.

Якщо ж знайдено спостереження $x_j^{(t)}$, для якого виконується умова (1), то проводиться корекція $w^{(t-1)}$ наступним чином:

$$w^{(t)}=(1-\gamma^{(t)})\cdot w^{(t-1)}+\gamma^{(t)}\cdot x_j^{(t)},$$

де

$$\gamma=\arg\min_{\gamma}\left|(1-\gamma^{(t)})\cdot w^{(t-1)}+\gamma^{(t)}\cdot x_j^{(t)}\right|.$$

Відомі оцінки зверху числа ітерацій t^0 обох алгоритмів:

$$t^0\le\begin{cases}\dfrac{Q^2}{\varepsilon^2} & \text{для алгоритма Розенблата,}\\ \dfrac{Q^2}{\varepsilon^2}\ln\dfrac{Q^2}{\varepsilon^2} & \text{для алгоритма Козинця,}\end{cases} \qquad (2)$$

де $Q=\max\limits_{i\in[1,K]}\left|x_i^{(N)}\right|$, $\varepsilon=\min\limits_{x^{(N)}\in \mathrm{Co}(X)}\left|x^{(N)}\right|>0$, а $\mathrm{Co}(X)$ – опукла оболонка множини X.

На перший погляд з порівняння оцінок (2) може здаватися, що алгоритм Розенблата завжди збігається з меншим числом ітерацій. Однак, такий висновок не правомірний, оскільки згідно з [3] наведені оцінки (2) досить грубі. Тому питання про те, який з алгоритмів має переваги за швидкістю досі залишався відкритим.

Розглянемо результати статистичний експерименту, який надає відповідь на це питання і передбачав багаторазову генерацію масивів двовимірних векторів:

$Q_m^{(j)}=\{q_{m,t}=(x_{m,k},y_{m,k}),\ k=1,...,K\}$, $j=1,2$, $m=1,...,M_0$,

які заздалегідь можуть бути розділені лінійною дискримінантною функцією. Для цього на кожному m-му випробуванні випадковим чином обираються параметри «прихованої» прямої, відносно якої генеруються точки $q_{m,k}\in Q_m^{(1)}$ і $q_{m,k}\in Q_m^{(2)}$, $k=1,...,K$.

Генеровані дані паралельно оброблювались алгоритмами Розенблата і Козинця. Для кожного $m=1,2,...,M_0$ визначають кількості ітерацій $U_{m,1}(K)$ і $U_{m,2}(K)$, витрачених першим і другим алгоритмами під час оброблення фіксованої кількості спостережень $K=10,...,200$. Моменти t_0 зупинку алгоритмів визначає умова

$$\left\langle w^{t_0},x_j\right\rangle\cdot y_j>0, \qquad \forall j=1,...,K.$$

За результатами M_0 експериментів обраховується процентне співвідношення кількості ітерацій $U_1(K)$ і $U_2(K)$ витрачених кожним алгоритмом під час навчання, що надає змогу визначити лідера за швидкодією.

Експерименти показали, що для $K<40$ приблизно в 20 % випадків алгоритми потребували однакову кількість ітерацій. Зі збільшенням K алгоритм навчання Козинця виявлявся абсолютним лідером: $U_1(K)>U_2(K)$ (рис. 1).

Експерименти також показали, що швидкість збіжності алгоритму Козинця менш чутлива до розташування точок у просторі ознак.

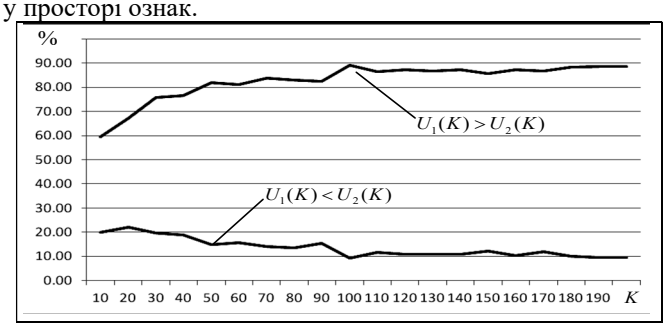

Рис 1. Залежності процента перемог алгоритмів від кількості спостережень

III. ОЦІНЮВАННЯ ТОЧНОСТІ ВИЗНАЧЕННЯ ДІАГНОСТИЧНИХ ОЗНАК ЗА МЕТОДОМ ФАЗАГРАФІЇ

Фазаграфія – інноваційний метод в кардіології, який передбачає перехід від скалярного сигналу $z(t)$ до траєкторії на фазовій площині з координатами $z(t), \dot{z}(t)$, де $\dot{z}(t)$ – швидкість зміни сигналу [4]. Клінічні випробування на вибірках реальних спостережень підтвердили діагностичну цінність оригінальної ознаки – показника β_T симетрії зубця T ЕКГ у фазових координатах.

Однак такими експериментами не обмежилося дослідження цінності β_T. Виявилося, що навіть в однієї здорової людини показник β_T може зазнавати чималих змін протягом короткого інтервалу часу. Тому потрібно було побудувати доказовий експеримент, який би підтвердив, що динаміка β_T свідчать про його високу чутливість, а не обумовлена випадковими збуреннями, що супроводжують реєстрацію ЕКГ у реальних умовах.

Такий експеримент з елементами дедуктивного підходу (від загальної моделі до конкретних висновків) проводився за допомогою стохастичної моделі породження штучної ЕКГ реалістичної форми [5]. Відповідно до моделі послідовність циклів ЕКГ генерують шляхом випадкових спотворень параметрів A_i, μ_i, $b_i^{(1)}$, $b_i^{(2)}$ еталонного циклу (корисного сигналу)

$$z_0(t) = \sum_i A_i \exp[-\frac{(t-\mu_i)^2}{2[b_i(t)]^2}], \qquad (2)$$

який має вигляд суми несиметричних функцій Гауса.

Несиметричність окремих фрагментів еталону (2), зокрема, симетрії зубця T досягається за рахунок виконання умови

$$b_i(t) = \begin{cases} b_i^{(1)}, \text{ якщо } t < \mu_i, \\ b_i^{(2)}, \text{ якщо } t > \mu_i \end{cases} \qquad b_i^{(1)} \neq b_i^{(2)}$$

Доведена теорема, згідно з якою функція $z_0(t)$ та її перша похідна $\dot{z}_0(t)$ неперервні у всіх точках області визначення, в тому числі, в точках $t = \mu_i$, в яких функції $b_i(t)$ розривні внаслідок виконання $b_i^{(1)} \neq b_i^{(2)}$. Це дало змогу організувати доказовий експеримент, який полягав у багаторазовій генерації та оброблені методом фазаграфії штучних ЕКГ за еталонами (2) з відомими значеннями $\beta_T = b_T^{(2)} / b_T^{(1)}$ (рис. 2).

Така організація експерименту надала змогу оцінювати точність визначення діагностичної ознаки β_T еталонного циклу $z_0(t)$. Зауважимо, що під час оброблення реальних ЕКГ, спотворених завадами, інформація про точне значення β_T недоступна.

Експерименти показали, що в широкому діапазоні $\beta_T \in [0{,}3\ 3{,}0]$ під час оброблення $M \geq 50$ циклів модельної ЕКГ в умовах випадкових неадитивних збурень $\varepsilon_{Tm}^{(1)} \leq 50\ \%$, $\varepsilon_{Tm}^{(2)} \leq 50\ \%$ та адитивної завади $h(t_k) \leq 50\ \%$ метод фазаграфії забезпечує високу точність оцінки β_T: стандартна помилка $\hat{\beta}_T$ становила лише 0,021, а середня відносна помилка не перевищувала 2,64 %, що достатньо для практичного використання методу.

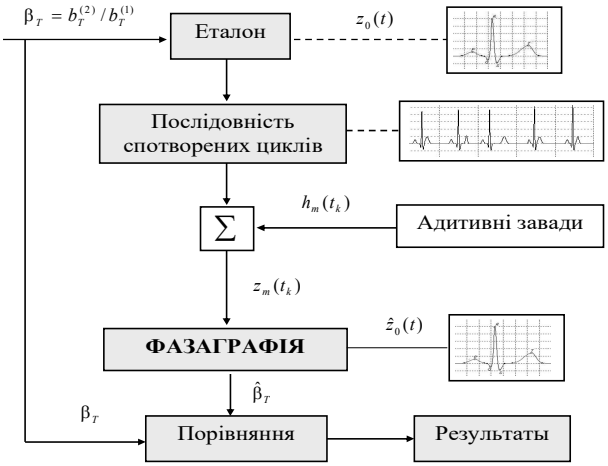

Рис 1. Схема доказового експерименту

ЛІТЕРАТУРА REFERENCES

[1] Л. Файнзільберг., Ю. Яременко Комп'ютерне моделювання модифікованого методу оптимальної зупинки. Proceedings of the international scientific conference «Information technologies and computer modeling». 2018. Івано-Франківськ: Прикарпатський національний університет імені Василя Стефаника. С. 270-273.

[2] L.S. Fainzilberg N.A. Matushevych. Comparative evaluation of convergence's speed of learning algorithms for linear classifiers by statistical experiments method. Cybernetics and computer engineering. 2018. N. 2 (192). P. 6-22. https://doi.org/10.15407/kvt192.02

[3] M. Schlesinger, V. Hlavac. Ten lectures on statistical and structural pattern recognition. 2002. Dodrecht- Boston- London: Kluwer Academic Publishers. 519 p.

[4] L.S. Fainzilberg. New Approaches to the analysis and interpretation of the shape of cyclic signals. Cybernetics and systems analysis. 2020. Vol. 56, N. 4. P. 665-674. https://doi.org/10.1007/s10559-020-00283-0

[5] L.S. Fainzilberg., T.Yu. Bekler, G.A. Glushauskene. Mathematical model for generation of artificial electrocardiogram with given amplitude-time characteristics of informative fragments. Journal of automation and information sciences. 2011. Vol. 43. Issue 9. P. 20-33. https://doi.org/10.1615/JAutomatInfScien.v43.i9.20