Класифікації Електрокардіограм на Основі Метрики Левенштейна

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

Electrocardiograms Classification Based on the Levenstein Metric

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

***Анотація* — Розвивається лінгвістичний підхід до класифікації ЕКГ, заснований на переході від спостереження до кодового слова, що характеризує динаміку зміни сукупності діагностичних показників. Запропоновано оцінювати близькість ЕКГ редакторською відстанню за Левенштейном. Наведено переваги запропонованого підходу під час діагностики ішемії міокарда.**

***Abstract* — A linguistic approach to ECG classification based on the transition from observation to a code word that characterizes the dynamics of changes in the set of diagnostic indicators is being developed. It is proposed to evaluate the proximity of the ECG by distance according to Levenshtein. The advantages of the proposed approach during the diagnosis of myocardial ischemia are presented**

Ключові слова — лінгвістичний підхід; діагностичні показники; редакторська відстань.

Keywords — linguistic approach; diagnostic indicators; editorial distance.

I. ВСТУП

Дослідження у біології та медицині часто пов'язані з аналізом циклічних сигналів [1]. Типовий приклад таких сигналів - електрокардіограма (ЕКГ), яка відображає циклічний характер роботи серця [2].

Один з перспективних методом оброблення ЕКГ заснований на перетворенні вихідного сигналу на послідовність символів, що характеризують динаміку зміни форми циклів [3]. Мета доповіді – подальший розвиток лінгвістичного методу аналізу та інтерпретації циклічних сигналів на прикладі ЕКГ.

II. ПРОЦЕДУРА КОДУВАННЯ ЕКГ

Припустимо, що кожен n-й цикл ЕКГ ($n=2,...,N$) описаний сукупністю діагностичних показників $x_1,...,x_M$. Оцінюватимемо зміну m-го показника ($m=1,...,M$) за допомогою тризначної індикаторної функції

$$V_n^{(m)}=\begin{cases}+1, \text{ якщо } x_n^{(m)}-x_{n-1}^{(m)}>\varepsilon,\\ 0, \quad \text{якщо } \left|x_n^{(m)}-x_{n-1}^{(m)}\right|\le\varepsilon, \quad n=2,...,N,\\ -1, \text{ якщо } x_n^{(m)}-x_{n-1}^{(m)}<-\varepsilon,\end{cases} \qquad (1)$$

де ε – поріг нечутливості до змін m-го показника від циклу до циклу.

Можливі комбінації значень $V_n^{(1)}, V_n^{(2)},..., V_n^{(M)}$ визначають 3^M символів $\alpha_n \in \mathrm{A}$ для кожного циклу, а ланцюжок символів $S_k=\alpha_1\alpha_2\cdots\alpha_{N-1}$ формує $N-1$ розрядне слово, яке кодує сигнал $z_k(t)$, що обробляється. Наприклад, алфавіт A містить 27 символів, якщо динаміку ЕКГ характеризувати змінами трьох показників – тривалістю RR інтервалів, симетрією зубців T та амплітудами зубців R [4], які застосовують в кардіологічній практиці.

Перехід від спостереження $z_k(t)$ до кодового слова S_k надає змогу використовувати методи математичної лінгвістики для вирішення задачі аналізу та інтерпретації сигналу $z_k(t)$. Для цього будемо оцінювати близькість двох ЕКГ $z_\mu(t)$ та $z_\nu(t)$ відстанню Левенштейна $L(S_\mu,S_\nu)$ [5] між парами відповідних кодових слів S_μ і S_ν, яка визначає мінімальну кількість операцій редагування

(вставки, видалення та заміни символу) для переходу від S_μ до S_ν.

Для визначення оптимального значення порогу ε, що фігурує в (1), пропонується такий алгоритм.

Нехай маємо навчальну вибірку, що містить Q_1 спостережень класу Ψ_1 та спостережень Q_2 класу Ψ_2. Будемо кодувати спостереження класу Ψ_1 відповідно до (1) за різними фіксованими значеннями ε з певним кроком $\Delta\varepsilon$ в інтервалі $0 \le \varepsilon \le \varepsilon_{\max}$. В результаті побудуємо $\varepsilon_{\max} / \Delta\varepsilon$ матриць внутрішньокласових відстаней Левенштейна за різними дискретними значеннями $\varepsilon = 0,...\varepsilon_{\max}$:

$$\Lambda_\varepsilon^{(1)} = \begin{pmatrix} L_\varepsilon(S_1^{(1)}, S_1^{(1)}) & L_\varepsilon(S_1^{(1)}, S_2^{(1)}) & ... & L_\varepsilon(S_1^{(1)}, S_{Q_1}^{(1)}) \\ L_\varepsilon(S_2^{(1)}, S_1^{(1)}) & L_\varepsilon(S_2^{(1)}, S_2^{(1)}) & ... & L_\varepsilon(S_2^{(1)}, S_{Q_1}^{(1)}) \\ & & ... & \\ L_\varepsilon(S_{Q_1}^{(1)}, S_1^{(1)}) & L_\varepsilon(S_{Q_1}^{(1)}, S_2^{(1)}) & ... & L_\varepsilon(S_{Q_1}^{(1)}, S_{Q_1}^{(1)}) \end{pmatrix}. \quad (2)$$

За елементами кожної з матриць (2) обчислимо середню внутрішньо класову відстань $\overline{L}_\varepsilon^{(1)}$, що залежить від ε:

$$\overline{L}_\varepsilon^{(1)} = \frac{2}{Q_1 (Q_1 - 1))} \sum_{\nu=1}^{Q_1} \sum_{\mu=1}^{Q_1} L_\varepsilon(S_\mu^{(1)}, S_\nu^{(1)}). \quad (3)$$

Аналогічним чином за елементами матриць $\Lambda_\varepsilon^{(2)}$ обчислимо середню внутрішньокласову відстань $\overline{L}_\varepsilon^{(2)}$ для другого класу, що залежить від ε:

$$\overline{L}_\varepsilon^{(2)} = \frac{2}{Q_2 (Q_2 - 1))} \sum_{\nu=1}^{Q_2} \sum_{\mu=1}^{Q_2} L_\varepsilon(S_\mu^{(2)}, S_\nu^{(2)}). \quad (4)$$

Далі побудуємо $Q_1 \times Q_2$ матриць міжкласових відстаней $L_\varepsilon(S_\mu^{(1)}, S_\nu^{(2)})$ між усіма парами кодових слів для класів Ψ_1 і Ψ_2 за фіксованих значень ε в інтервалі $0 \le \varepsilon \le \varepsilon_{\max}$:

$$\Lambda_\varepsilon^{(1,2)} = \begin{pmatrix} L_\varepsilon(S_1^{(1)}, S_1^{(2)}) & L_\varepsilon(S_1^{(1)}, S_2^{(2)}) & ... & L_\varepsilon(S_1^{(1)}, S_{Q_2}^{(2)}) \\ L_\varepsilon(S_2^{(1)}, S_1^{(2)}) & L_\varepsilon(S_2^{(1)}, S_2^{(2)}) & ... & L_\varepsilon(S_2^{(1)}, S_{Q_2}^{(2)}) \\ & & ... & \\ L_\varepsilon(S_{Q_1}^{(1)}, S_1^{(2)}) & L_\varepsilon(S_{Q_1}^{(1)}, S_2^{(2)}) & ... & L_\varepsilon(S_{Q_1}^{(2)}, S_{Q_2}^{(2)}) \end{pmatrix}. \quad (5)$$

За елементами матриць (5) обчислимо середню міжкласову відстань $\overline{L}_\varepsilon^{(1,2)}$, що залежить від ε:

$$\overline{L}_\varepsilon^{(1,2)} = \frac{1}{Q_1 Q_2} \sum_{\rho=1}^{Q_2} \sum_{\mu=1}^{Q_1} L_\varepsilon(S_\mu^{(1)}, S_\rho^{(2)}). \quad (6)$$

.

Використовуючи величини (3), (4) та (6) сформуємо критерій оптимальності

$$\eta(\varepsilon) = \frac{\overline{L}_\varepsilon^{(1)} + \overline{L}_\varepsilon^{(2)}}{\overline{L}_\varepsilon^{(1,2)}}, \quad (7)$$

який надає змогу визначити оптимальне значення

$$\varepsilon_0 = \arg\min_{0 \le \varepsilon \le \varepsilon_{\max}} \eta(\varepsilon), \quad (8)$$

що забезпечує мінімальні внутрішньокласові відстані $\overline{L}^{(1)}$, $\overline{L}^{(2)}$ та одночасно максимальну міжкласову відстань.

III. ВИРІШУВАЛЬНЕ ПРАВИЛО

Нехай у результаті експериментів зареєстровано Q_g спостережень класу $\Psi_g \in \{\Psi_1,..., \Psi_G\}$, які закодовані словами $S_1^{(g)}$, $S_2^{(g)}$,…, $S_{Q_g}^{(g)}$.

Сформуємо квадратну $Q_g \times Q_g$ матрицю

$$\Lambda^{(g)} = \begin{pmatrix} L(S_1^{(g)}, S_1^{(g)}) & L(S_1^{(g)}, S_2^{(g)}) & ... & L(S_1^{(g)}, S_{Q_g}^{(g)}) \\ L(S_2^{(g)}, S_1^{(g)}) & L(S_2^{(g)}, S_2^{(g)}) & ... & L(S_2^{(g)}, S_{Q_g}^{(g)}) \\ & & ... & \\ L(S_{Q_g}^{(g)}, S_1^{(g)}) & L(S_{Q_g}^{(g)}, S_2^{(g)}) & ... & L(S_{Q_g}^{(g)}, S_{Q_g}^{(g)}) \end{pmatrix} \quad (9)$$

між усіма парами кодових слів, що відповідають ЕКГ g-го класу з навчальної вибірки.

Еталон $S_0^{(g)}$ класу $\Psi_g \in \{\Psi_1,..., \Psi_G\}$ визначаємо за рядком матриці (9), сума елементів якого мінімальна, тобто.

$$S_0^{(g)} = \arg\min_{1 \le \nu \le Q_g} \sum_{\mu=1}^{Qg} L(S_\mu^{(g)}, S_\nu^{(g)}). \quad (10)$$

Аналогічним чином визначимо еталони $S_0^{(1)},...,S_0^{(G)}$ інших класів. Рішення про належність аналізованої ЕКГ класу Ψ_q, $q = 1,...,G$, приймаємо у разі, коли відповідне кодове слово задовольняє умову

$$L(S_t, S_0^{(q)}) = \min_{1 \le g \le G} L(S_t, S_0^{(g)}). \quad (11)$$

.

IV. ЕКСПЕРИМЕНТАЛЬНІ ДОСЛІДЖЕННЯ

Для оцінювання ефективності правила (11) використовувалися 100 записів ЕКГ верифікованих хворих на хронічну форму ішемічної хвороби серця (клас CAD) і 100 записів ЕКГ здорових добровольців (клас HEALTHY) [10]. Діагноз CAD був підтверджений за результатами

коронарографії, хоча на ЕКГ хворих були відсутні традиційні ознаки ішемії міокарда.

На рис. 1 представлені оцінки умовних розподілів $P(L(S_t, S_0^{(1)}))$ та $P(L(S_t, S_0^{(2)}))$ відстаней Левенштейна між кодовими словами навчальної вибірки по відношенню до еталонів $S_0^{(1)}$ та $S_0^{(2)}$. Перевірка за критерієм Колмогорова-Смирнова показала, що з високою статистичною значимістю $(p < 0{,}001)$ гіпотеза про однаковість розподілів $P(L(S_t, S_0^{(1)}))$ і $P(L(S_t, S_0^{(2)}))$ має бути відкинута [10]. Аналогічний факт підтвердив критерій Мана-Уітні для незалежних вибірок.

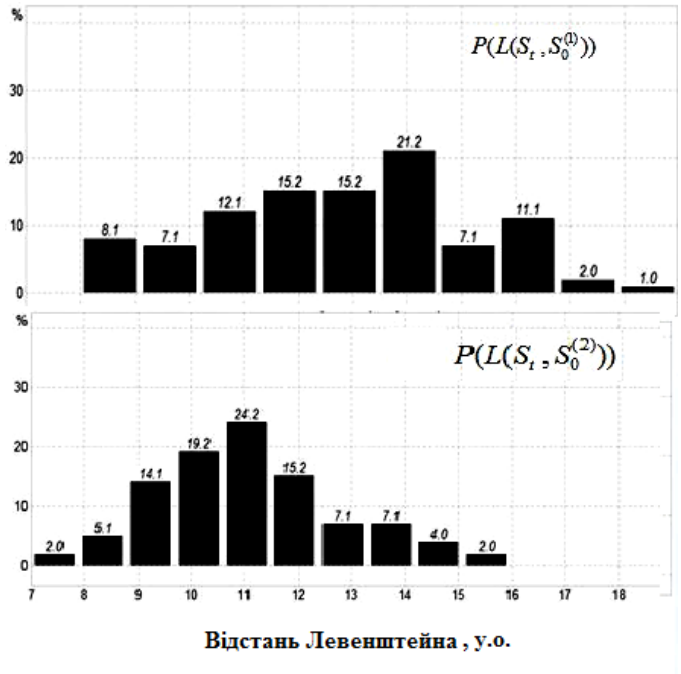

Рис. 1. Умовні розподіли відстаней Левенштейна в класах CAD и HEALTHY

Запропонований підхід надав змогу з чутливістю $S_E = 72\%$ та специфічністю $C_P = 79\%$ класифікувати реальні ЕКГ в складних випадках (коли відсутні традиційні ознаки ішемії міокарда) за правилом.

$$\text{CAD, якщо } L(S_t, S_0^{(1)}) \le L(S_t, S_0^{(2)}), \qquad (12)$$

$$\text{HEALTHY, якщо } L(S_t, S_0^{(1)}) > L(S_t, S_0^{(2)}), \qquad (13)$$

де S_t – кодове слово аналізованої ЕКГ, а $S_0^{(1)}$ й $S_0^{(2)}$ – побудовані еталони хворих на ІХС та здорових волонтерів.

Достовірність класифікації ЕКГ може бути збільшена, якщо додатково до відстані Левенштейна аналізувати появу домінантного патерну сигналу у вигляді послідовності трьох символів $\pi = \lambda\rho\vartheta$, $\lambda, \rho, \vartheta \in A$ у кодовому слові, яка найбільш характерна для даного класу ЕКГ [14]. Домінантний патерн g-го класу ($g = 1,...G$) обчислюють за формулою

$$\pi_0^{(g)} = \arg\max_{1 \le l \le L} \hat{P}^{(g)}(\pi_l) \qquad (14)$$

де

$$\hat{P}^{(g)}(\pi_l) = \frac{1}{Q_g} \sum_{\mu=1}^{Q_g} \frac{W_\mu^{(g)}(\pi_l)}{N-2}, \; l = 1,..., L \qquad (15)$$

– середні значення частот $\hat{P}^{(g)}(\pi_l)$ появи трьохсимвольних патернів $\pi_l, l = 1,..., L$, $W_\mu^{(g)}(\pi_l)$ – кількість входжень l-го патерну π_l в μ-те кодове слово класу $V_g \in \{V_1,...,V_G\}$, а L – кількість можливих варіантів трьохсимвольних патернів з елементів алфавіту A.

Для прискорення процедури обчислення $W_\mu^{(g)}(\pi_l)$ доцільно використовувати алгоритм Боєра–Мура [6].

Експериментально доведено, що додатковий аналіз домінантних патернів дозволяє на 4,2% підвищити достовірність рішень за правилом (12) (13)

IV. ВИСНОВКИ

У рамках лінгвістичного підходу розроблено конструктивні комп'ютерні процедури, які на основі аналізу відстаней Левенштейна між кодовими словами ЕКГ та виявлення на них характерних послідовностей символів надають змогу приймати діагностичні рішення у складних випадках, коли на ЕКГ відсутні традиційні діагностичні ознаки ішемії міокарда.

Література References

[1] V. Zvarich, B. Marchenko. (2011). "Linear autoregressive processes with periodic structures as models of information signals", in Radioelectronics and Communications Systems, vol. 54, no. 7, pp. 367–372.

[2] L. Fainzilberg. (2015). "Generalized Method of Processing Cyclic Signals of Complex Form in Multidimension Space of Patameters", in Journal of Automation and Information Sciences, vol. 47, issue 3, pp. 24-39. https://doi.org/10.1615/JAutomatInfScien.v47.i3.30

[3] L, Fainzilberg , Ju. Dykach. (2019). "Linguistic approach for estimation of electrocardiograms's subtle changes based on the Levenstein distance", in Cybernetics and Computer Engineering, no. 2 (196), pp. 3-26. https://doi.org/10.15407//kvt196.02.003

[4] L. Fainzilberg. (2022). "Cyclic signals classification by codegrams characterizing the dynamics of cycles shape changing", in International Scientific Technical Journal "Problems of Control and Informatics", no. 3, pp. 112-123. http://jnas.nbuv.gov.ua/article/UJRN-0001367046

[5] R. Wagner, M. Fischer. (1971). "The String-to-String Correction Problem", in Journal of the ACM, vol. 21, issue 1, pp.168-173, https://doi:10.1145/321796.321811

[6] R. Cole. (1994). "Tight bounds on the complexity of the Boyer-Moore string matching algorithm", in SIAM Journal on Computing, vol 23, no. 5, pp. 1075-1091. https://doi:10.1137/S0097539791195543