ПРОГРАМНО-ТЕХНІЧНІ КОМПЛЕКСИ

А.Л. ЕРОХИН, А.С. БАБИЙ, А.С. НЕЧИПОРЕНКО, А.П. ТУРУТА

УДК 004.942:616-073.175

МЕТОД ПОСТРОЕНИЯ НЕЧЕТКОЙ РЕГРЕССИОННОЙ МОДЕЛИ НА ОСНОВЕ LARS ДЛЯ ВЫБОРА ЗНАЧИМЫХ ПРИЗНАКОВ

Аннотация. Предложен метод построения нечеткой регрессионной модели на основе LARS. Рассмотрены особенности использования нечеткого регрессионного анализа в задачах медицинской диагностики. Данный метод позволяет сократить число параметров модели, влияющих на прогнозируемую степень обструкции носового дыхания и избежать «перетренированности» модели.

Ключевые слова: риноманометрия, нечеткий регрессионный анализ, метод LARS, выбор значимых факторов, мультиколлинеарность, линейное программирование.

ВВЕДЕНИЕ

Нарушение функции носового дыхания всегда сопровождается ухудшением самочувствия пациента и во многих случаях является симптомом острых или хронических воспалительных заболеваний носа и околоносовых пазух, различных опухолевых процессов. В современной ринологии для оценки функции носового дыхания используются методы, изложенные в работах [1, 2]. Среди них методы томографии (КТ и МРТ), исследования воздушного потока, проходящего через носовую полость (риноманометрия, ринорезистометрия, исследование спектра звуковых характеристик носового воздушного потока и др.), акустическая риноманометрия [3], комплекс методов субъективной диагностики [4].

В настоящее время для оценки степени обструкции носового дыхания в клинической практике наиболее широко применяется метод передней активной риноманометрии (ПАРМ) [5]. При оценке риноманометрических исследований понятие нормы четко не определено и имеет множество интерпретаций. Результаты риноманометрических измерений зависят от расы, возраста, пола, индекса массы тела и роста [6–8]. Однако согласно рекомендациям Международного комитета по оценке носового дыхания основным диагностическим параметром ПАРМ принято считать величину носового сопротивления R_{150} [6], которая рассчитывается при фиксированном значении дифференциального давления 150 Па. Отметим, что по различным данным [9, 10] у 24–33 % людей дифференциальное давление не достигает 150 Па при спокойном дыхании, поэтому такой расчет для них неприемлем.

ПОСТАНОВКА ЗАДАЧИ ИССЛЕДОВАНИЯ

Для оценки результатов риноманометрических исследований рассчитываются следующие параметры: носовое сопротивление R_{150} (при четырехфазовой концепции расчет проводится для первой фазы на вдохе и четвертой фазы на вы-

дохе), R_{100}, R_{75}, коэффициенты k_1 и k_2 по формуле Рехрера [11], $R_2(V_2)$ по модели Бромса [12], коэффициент λ [13]. В ринологической практике в настоящее время превалирует концепция оценивания степени обструкции носового дыхания, основанная на расчете носового сопротивления R_{150}.

Как отмечалось ранее, для пациентов с порогом дифференциального давления, меньшим 150 Па, невозможно получить стандартных параметров носового дыхания. Для решения данной проблемы в [10] проведено оценивание степени обструкции носового дыхания для определенной категории пациентов, где расчет параметров R_{150} реализован с помощью экстраполяции. При этом анализируемая зависимость давление–поток аппроксимируется уравнением Рехрера [11].

В работах [14, 15] исследовалась оценка степени обструкции с помощью методов логистической, линейной и кусочно-линейной регрессий. В частности, по результатам расчетов в [14] сделаны выводы о влиянии на степень обструкции возраста пациента и аллергического компонента. При этом массив входных переменных регрессионных моделей содержал значения носовых сопротивлений $R_2(V_2)$, полученных из модели Бромса, а стандартные параметры R_{75}, R_{100}, R_{150} в расчетах не учитывались.

В работе [16] с использованием линейной регрессионной модели установлена взаимосвязь между параметрами R_{75}, R_{100}, возрастом и характеристиками поверхности тела для пяти возрастных групп детей, а также сделан вывод о значимости коэффициентов R_{75}, R_{100} при проведении оценки степени обструкции у детей. Однако в [5–16] исследуемые коэффициенты рассматривались как четкие параметры модели.

Анализируя данные расчета R_{150} для одного и того же пациента в различные промежутки времени, можно отметить разброс значений, обусловленный влиянием на измерения таких факторов, как температура и влажность в помещении, смещение маски и нарушение фиксации трубки давления, невыполнение рекомендаций по адаптации пациента к измерениям. Кроме того, для некоторых пациентов наблюдается смещение границ интервалов степеней обструкции. При оценке степени обструкции оперируют не точечными значениями, а интервальными, следовательно, не всегда отношения между регрессорами и возвращаемыми результатами соответствуют модели четкого регрессионного анализа. Для решения этой проблемы предлагается представлять значения коэффициента R_{150} в виде нечетких данных. Тогда задачу исследования можно отнести к классу задач нечеткого регрессионного анализа.

Задача нечеткого регрессионного анализа рассматривалась в [17], где предлагалось для ее решения использовать методы линейного программирования.

Построение модели регрессии с применением методов линейного программирования имеет следующие основные недостатки:

— слабое обоснование соотношения между решением задачи линейной оптимизации суммарной «нечеткости» возврата модели и минимизацией ее суммарной ошибки по сравнению с обучающей выборкой, например, в [18] предлагалось минимизировать расстояния между нечеткими числами на выходе модели и обучающей выборкой, что порождало решение нелинейной оптимизационной задачи;

— повышенная чувствительность модели к появлению аномальных уровней данных [19];

— присущая нечеткой линейной регрессии тенденция к мультиколлинеарности с увеличением количества влияющих факторов модели [20].

Для решения проблемы сокращения количества факторных переменных при построении нечетких моделей в [21] рекомендуется применять метод шагового регрессионного анализа. В качестве критерия выбора фактора для задачи построения нечеткой нелинейной регрессионной модели в [22] предлагается использо-

вать критерий Фишера, согласно которому осуществляется последовательное добавление и удаление признаков. Однако существенным недостатком данного метода является невозможность получения оптимального уравнения регрессии. Вследствие корреляций между предикатами значимую переменную можно не включать в уравнение, а вводить в него второстепенные переменные.

В настоящей работе для построения нечеткой регрессионной модели с четким входом и нечетким выходом предлагается последовательное использование метода регуляризации LARS и способа построения нечеткой регрессионной модели на основе линейного программирования.

МЕТОДИКА ИЗМЕРЕНИЙ И ОБРАБОТКА ДАННЫХ

При проведении ПАРМ измеряют дифференциальное давление и расход воздушного потока с помощью разработанного для риноманометрических исследований программно-аппаратного комплекса «Optimus» [23]. Комплекс сертифицирован в Украине (свидетельство государственной регистрации № 14777/2015 от 12.06.2015 г.). Измерительный модуль системы представляет собой микропроцессорное устройство, предназначенное для измерения физических величин малого дифференциального давления и двунаправленного потока воздуха с их первичной обработкой и дальнейшей передачей в ЭВМ. Функционально модуль состоит из первичных преобразователей давления и потока, цепей аналоговой и цифровой обработки сигнала, цепей питания и преобразования интерфейса. Сигналы дифференциального давления ΔP (Па) и расхода воздушного потока $\dot{V}$ (см^3/с) регистрируются синхронно (рис. 1).

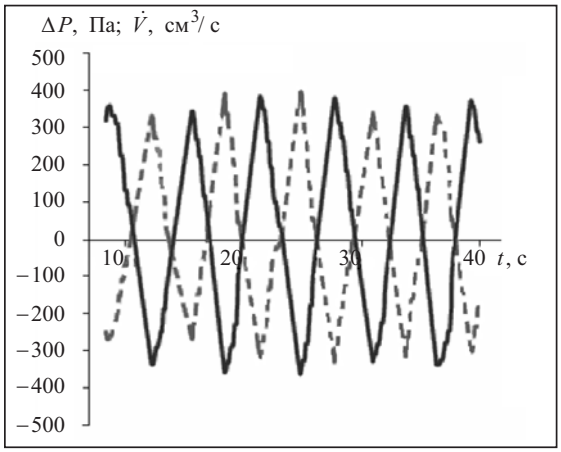

Рис. 1. Графики зависимости дифференциального давления (штриховая линия) и расхода воздуха (сплошная линия) от времени

Измерение давления в носоглотке осуществляется в одной обтурированной ноздре, которая исключается из акта дыхания, следовательно, проводится для каждой ноздри отдельно, исследуются вдох и выдох. На основании проведенных измерений строится график (ринограмма) зависимости расхода воздушного потока от дифференциального давления (рис. 2) и рассчитывается основной диагностический параметр — сопротивление R_{150} [9], согласно которому оценивается степень обструкции дыхания $R = \Delta p / \dot{V}$.

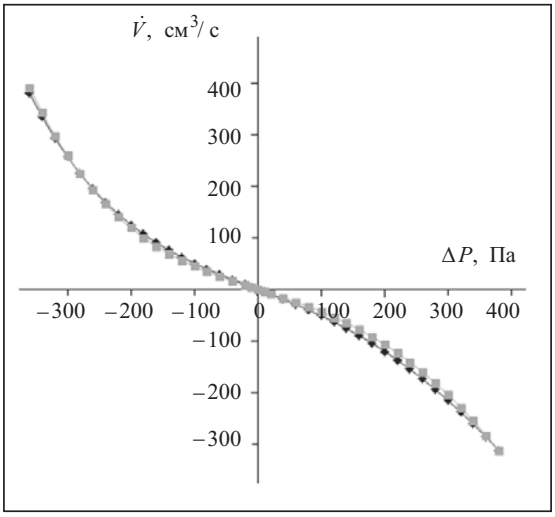

Рис. 2. Ринограмма

Таким образом, формируется массив входных данных, представленный в виде четких

значений $X_j = \{x_{ij}\}$, $j = 1, \ldots, n$, $i = 1, \ldots, m$, где X_1 — значение k_1 ($\text{Па} \cdot \text{с} / \text{м}^3$), X_2 — значение k_2 ($\text{Па} \cdot \text{с}^2 / \text{м}^6$), X_3 — значение R_{100} ($\text{Па} \cdot \text{с} / \text{м}^3$), X_4 — значение R_{75} ($\text{Па} \cdot \text{с} / \text{м}^3$), X_5 — значение $R_2(V_2)$, X_6 — диаметр ноздри (мм).

НЕЧЕТКАЯ РЕГРЕССИОННАЯ МОДЕЛЬ НА ОСНОВЕ LARS

Для реализации предлагаемой нечеткой регрессионной модели необходимо иметь функцию принадлежности, позволяющую представить параметры, характеризующие степень обструкции, в виде нечетких множеств. Воспользуемся симметричной треугольной функцией принадлежности согласно [20, 25], для построения которой используем набор исходных данных — массив значений сопротивления R_{150}, обозначим его $\widetilde{Y}_i$, $i = \overline{1, n}$. Тогда функция принадлежности i-го коэффициента (рис. 3) будет иметь вид

$$\mu_{\widetilde{Y}_i} = \max\left\{1 - \frac{y - y_i}{e_i}, 0\right\},$$

где y_i — центр нечеткой величины, e_i — разброс значений нечеткой величины.

Для уменьшения размерности модели необходимо выбрать факторы, которые существенно влияют на значение коэффициента R_{150} с помощью методики LARS [24], описанной далее.

Для входного набора данных X и значений коэффициентов R_{150} (центров y_i) необходимо выполнить следующую $L1$ регуляризацию.

Шаг 1. Задать начальную оценку $\hat{\mu}_A = 0$ вектора значений зависимой переменной y.

Шаг 2. Вычислить вектор корреляции $\hat{c} = X^{\mathrm{T}}(y - \hat{\mu}_A)$.

Шаг 3. Найти текущий набор индексов A, который соответствует признакам с наибольшими абсолютными значениями корреляции $A = \{j : |\hat{c}_j| = \hat{C}\}$, где

$$\hat{C} = \max_{j=1,\ldots,n} \{|\hat{c}_j|\}.$$

Шаг 4. Найти $s_j = \operatorname{sign}(\hat{c}_j)$ для $j \in A$. Рассчитать матрицы X_A и ψ_A: $X_A = [s_{j_1} x_{j_1}, \ldots, s_{j_{|A|}} x_{j_{|A|}}]$, $j = (j_1, \ldots, j_{|A|}) \in A$, $\psi_A = (1_A^{\mathrm{T}} \varsigma_A^{-1} 1_A)^{-1/2}$, где $s_j \in \{+1, -1\}$ и $|A|$ — мощность множества A (количество значений множества A), $\varsigma = X_A^{\mathrm{T}} X_A$, 1_A — единичная матрица размера $1 \times |A|$.

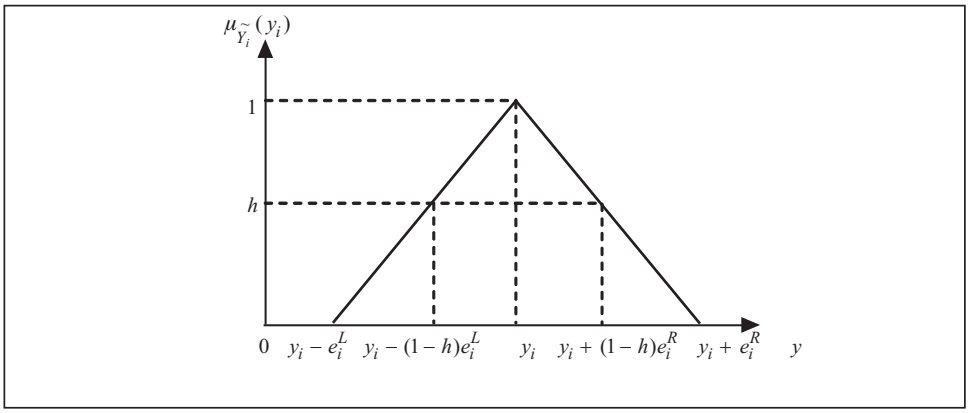

Рис. 3. Функция принадлежности

Шаг 5. Вычислить вектор $a = X^{\mathrm{T}} u_A$, где $u_A = X_A w_A$, $w_A = \psi_A \varsigma_A^{-1} 1_A$.

Шаг 6. Найти значение $\hat{\gamma} = \min\limits_{j \in A}{}^{+} \left\{ \dfrac{\hat{C} - \hat{c}_j}{\psi_A - a_j}, \dfrac{\hat{C} + \hat{c}_j}{\psi_A + a_j} \right\}$ (минимум определяется по всем положительным значениям для каждого j).

Шаг 7. Найти значения $\hat{\mu}_A$ для итерации $\hat{\mu}_{A+} = \hat{\mu}_A + \hat{\gamma} u_A$.

Шаг 8. Повторить процесс n раз (n — количество факторов), начиная с шага 2. Для каждого шага вычислить значение коэффициента Cp Маллоуза.

Для построения нечеткой регрессионной модели выберем набор коэффициентов, соответствующий минимальному значению коэффициента Cp. Тогда в общем виде зависимость коэффициента степени обструкции можно представить как нечеткую регрессионную модель

$$\widetilde{Y} = \widetilde{A}_0 + \widetilde{A}_1 X_1 + \ldots + \widetilde{A}_n X_n, \tag{1}$$

где $\widetilde{Y}_i = (y_i, e_i)$, $i = 1, \ldots, m$, — нечеткая величина с центром y_i и шириной e_i, $\widetilde{A}_j = (a_j, c_j)$, $j = 0, \ldots, n$, — нечеткая величина с центром a_j и шириной c_j.

Решение данной задачи согласно [17] сводится к решению задачи линейного программирования: минимизировать функцию

$$S = c_0 + \sum_{j=1}^{n} c_j \sum_{i=1}^{m} x_{ij}$$

с ограничениями

$$c_0 \geq 0, \ c_j \geq 0, \ j = 1, \ldots, n,$$

$$a_0 + \sum_{j=1}^{n} a_j x_{ij} + (1-h)\left[c_0 + \sum_{j=1}^{n} c_j |x_{ij}| \right] > y_i + (1-h)e_i,$$

$$a_0 + \sum_{j=1}^{n} a_j x_{ij} - (1-h)\left[c_0 + \sum_{j=1}^{n} c_j |x_{ij}| \right] < y_i - (1-h)e_i, \ i = 1, \ldots, m,$$

где $h \in \{0, 1]$ — коэффициент неопределенности. Таким образом, в результате решения этой задачи определяются нечеткие коэффициенты регрессионной модели $\widetilde{A}$.

РЕЗУЛЬТАТЫ

Для исследования использовалась выборка из 70 элементов, состоящая из двух одинаковых частей: обучающей и тестовой. При обучении модели выделялись два значимых фактора, а именно x_3 и x_5, которые в дальнейшем использовались для построения нечеткой регрессионной модели.

Согласно формуле (1) получена нечеткая регрессионная модель зависимости коэффициента R_{150}:

$$Y = (0.14, 0.001) X_3 + (0.001, 0.01) X_5 + (0.48, 0.2).$$

Количество ошибок для данной модели по обучающей и тестовой выборкам составило 1 % и 3.4 % соответственно. При создании модели, включающей все факторы, количество ошибок по обучающей и тестовой выборкам составило 1 % и 60 % соответственно.

ЗАКЛЮЧЕНИЕ

В работе предложен метод построения нечеткой регрессионной модели на основе метода LARS для выбора значимых признаков риноманометрических

исследований при диагностике степени обструкции носового дыхания. Предложенный метод позволяет уменьшить количество входных параметров модели, а значит, избежать ее «перетренированности». В отличие от методов выбора признаков на основе шаговой регрессии с использованием F-коэффициента уровень значимости задавать не нужно. Полученные результаты позволяют выделить два значимых коэффициента модели, влияющих на прогнозируемую степень обструкции.

СПИСОК ЛИТЕРАТУРЫ

1. Diagnosis and management of rhinitis: Complete guidelines of the Joint Task Force on practice parameters in allergy, asthma and immunology / M. Dykewicz, S. Fineman, D. Skoner, R. Nicklas, R. Lee, J. Blessing-Moore, J. Li, I. Bernstein, W. Berger, S. Spector, D. Schuller // Ann. Allergy Asthma Immunol. — 1998. — **81.** — P. 478–518.
2. Wheeler P., Wheeler S. Vasomotor rhinitis // Am. Fam. Physician. — 2005. — **72**, N 6. — P. 1057–1062.
3. Dadgarnia H., Baradaranfar M.H., Mazidi M., Azimi M.R. Assessment of septoplasty effectiveness using acoustic rhinometry and rhinomanometry // Iranian Journal of Otorhinolaryngology. — 2013. — 25 (71). — P. 71–78.
4. Thulesius H.L., Cervin A., Jessen M. Can we always trust rhinomanometry? // Rhinology. — 2011. — **49**, N 1. — P. 46–52.
5. Clement P.A., Gordts F. Standardisation committee on objective assessment of the nasal airway. Consensus report on acoustic rhinometry and rhinomanometry // Rhinology. — 2005. — **43.** — P. 169–179.
6. Canbay E.I., Bhatia S.N. A comparison of nasal resistance in white, caucasians and blacks // Rhinology. — 1997. — **11**, Iss. 1. — P. 73–75. — doi:10.2500/105065897781446801.
7. Samolinski B.K., Grzanka A., Gotlib T. Changes in nasal cavity dimensions in children and adults by gender and age // Laryngoscope. — 2007. — **117**, Iss. 8. — P. 1429–1433. — doi:10.1097/mlg.0b013e318064e837.
8. Crouse U., Laine-Alava M.T. Effects of age, body mass index, and gender on nasal airflow rate and pressures // Laryngoscope. — 1999. — **109**, Iss. 9. — P. 1503–1508. — doi:10.1097/00005537-199909000-00027.
9. Vogt K., Jalowayski A.A. 4-Phase-rhinomanometry basics and practice // Rhinology. — 2010. — **21.** — P. 1–50.
10. Naito K., Mamiya T., Mishima Y., Kondo Y., Miyata S., Iwata S. Comparison of calculated nasal resistance from Rohrer's equation with measured resistance at delta P 150Pa // Rhinology. — 1998. — **36**, N 1. — P. 28–31.
11. Rührer F. Der Stromungswiderstand in der menschlichen Atemwegen // Pflügers Arch Ges Physiology. — 1915. — **162.** — P. 225–295.
12. Broms P. Rhinomanometry. III. Procedures and criteria for distinction between skeletal stenosis and mucosal swelling // Acta Otolaryngol. — 1982. — **94.** — P. 361–370.
13. Mlynski G., Beule A. Diagnosis of respiratory function of the nose. Diagnostik der respiratorischen Funktion der Nase // HNO. Springer Medizin Verlag. — 2008. — **56**, Iss. 1. — P. 81–99. — http://link.springer.com/article/10.1007%2Fs00106-007-1655-0#page-2.
14. Thulesius H.L. Rhinomanometry in clinical use. A tool in the septoplasty decision making process: Doctoral dissertation. Clinical Sciences, 2012. — 67 p.
15. Malm L., v. Wijk R.G., Bachert C. Guidelines for nasal provocations with aspects on nasal patency, airflow, and airflow resistance. International Committee on Objective Assessment of the Nasal Airways, International Rhinologic Society. // Rhinology. — 2000. — **38**, N 1. — P. 1–6. — http:// www.rhinologyjournal.com/abstract.php?id=88.
16. Juliá J.C., Enriqueta M. Burchés Martorell A. Active anterior rhinomanometry in paediatrics. Normality criteria. // Allergologia et Immunopathologia. — 2011. — **39**, N 6. — P. 342–346.
17. Tanaka H., Uejima S., Asai K. Linear regression analysis with fuzzy model // IEEE Transactions on Systems, Man and Cybernetics. — 1982. — **12**, N 6. — P. 903–907.
18. Diamond P. Fuzzy least squares // Information Sci. — 1988. — **46**, N 3. — P. 141–157.
19. Peters G. Fuzzy linear regression with fuzzy intervals // Fuzzy Sets and Systems. — 1994. — **63**, N 1. — P. 45–55.

20. Kim K.J., Moskowitz H., Koksalan M. Fuzzy versus statistical linear regression // European Journal of Operational Research. — 1996. — **92**, N 2. — P. 417–434.
21. Milea V., Almeida R.J., Kaymak U., Frasincar F. A fuzzy model of a European index based on automatically extracted content information // Symposium on Computational Intelligence for Financial Engineering & Economics. IEEE Symposium, 2011, 11–15 April. — P. 1–8. — doi: 10.1109/CIFER.2011.5953571.
22. Chan K.Y., Lam H.A., Dillon T.S., Ling S.H. A stepwise based fuzzy regression procedure for developing customer preference models in new product development // Fuzzy Systems, IEEE Transactions on Fuzzy Systems. — 2013. — **23**, Iss. 5. — P. 1–18.
23. Нечипоренко А.С. Технические аспекты риноманометрии // Восточно-европейский журнал передовых технологий. — 2013. — **4**, № 9(64). — С. 11–14.
24. Efron B., Hastie T., Johnstone I., Tibshirani R. Least angle regression // Ann. Statist. — 2004. — **32**, N 2. — P. 407–499.
25. Shapiro A.F. Fuzzy regression models // ARC USA. — 2005. — P. 1–17.

Надійшла до редакції 03.12.2015

А.Л. Єрохін, А.С. Бабій, А.С. Нечипоренко, О.П. Турута
МЕТОД ПОБУДОВИ НЕЧІТКОЇ РЕГРЕСІЙНОЇ МОДЕЛІ НА ОСНОВІ LARS ДЛЯ ВИБОРУ ЗНАЧИМИХ ОЗНАК

Анотація. Запропоновано метод побудови нечіткої регресійної моделі на основі LARS. Розглянуто особливості використання нечіткого регресійного аналізу у задачах медичної діагностики. Цей метод дозволяє скоротити число параметрів моделі, які впливають на прогнозований ступінь обструкції носового дихання, а також уникнути «перетренованості» моделі.

Ключові слова: риноманометрія, нечіткий регресійний аналіз, метод LARS, вибір значущих факторів, мультиколінеарність, лінійне програмування.

A.L. Yerokhin, A.S. Babii, A.S. Nechyporenko, O.P. Turuta
THE METHOD TO CONSTRUCT FUZZY REGRESSION MODEL BASED ON LARS FOR SELECTION OF SIGNIFICANT FEATURES

Abstract. The paper proposes a method to construct a fuzzy regression model based on the LARS. The features of the use of fuzzy regression analysis for medical diagnosis are considered. The proposed method can reduce the number of model parameters affecting the projected degree of obstruction of nasal breathing and allows one to avoid “overtraining” of the model.

Keywords: rhinomanometry, fuzzy regression analysis, method LARS, selection of significant factors, multicollinearity, linear programming.

Ерохин Андрей Леонидович,
доктор техн. наук, профессор Харьковского национального университета радиоэлектроники, e-mail: ayerokhin@ukr.net.

Бабий Андрей Степанович,
аспирант Харьковского национального университета радиоэлектроники, e-mail: apratster@gmail.com.

Нечипоренко Алина Сергеевна,
кандидат техн. наук, доцент Харьковского национального университета радиоэлектроники, e-mail: alinanechiporenko@gmail.com.

Турута Алексей Петрович,
кандидат техн. наук, доцент Харьковского национального университета радиоэлектроники, e-mail: alexey.turuta@gmail.com.