УДК 531.37

О.В. ШКУЛІПА, Б.В. РУБЛЬОВ, Б.С. ШЕЙМАН

КОМБІНОВАНИЙ МЕТОД РОЗПІЗНАВАННЯ ПРИЧИНИ ВИНИКНЕННЯ ЗАХВОРЮВАННЯ

Анотація. *У роботі наведено підходи до побудови алгоритму розпізнавання причини виникнення захворювання. Запропоновано комбінований метод для визначення причини виникнення захворювання у хворого. Розглянуто роботу алгоритмів, що були використані, та їх покрокове застосування на практиці. Проведено оцінку можливості використання комбінованого методу на практиці.*
Ключові слова: *алгоритм кластеризації, еліпс мінімальної площі, метод опорних векторів, ForEll.*

Аннотация. *В работе представлены подходы к распознаванию причины возникновения болезни. Представлен комбинированный метод для определения причины возникновения болезни у больного. Продемонстрирована работа алгоритмов, которые были использованы, и их пошаговое применение на практике. Проведена оценка возможности ее использования для распознавания причины возникновения болезни.*
Ключевые слова: *алгоритм кластеризации, эллипс минимальной площади, метод опорных векторов, ForEll.*

Abstract*. Approaches to construction for recognition algorithm of the causes of diseases are shown in the article. A combined method to determine the causes of disease at a patient are proposed. A work of algorithms which have been used in model is considered, and step-by-step application of the algorithms in practice is demonstrated. Possibilities of using this method in practice are estimated.*
Key words: *clustering algorithm, minimal enclosing ellipse, support vector machine, ForEll.*

1. Вступ

У медицині одним із найбільш важливих компонентів, що впливають на ефективність лікування хворого в цілому, є заходи з діагностики етіологічного чинника захворювання (безпосередньої причини виникнення захворювання). Саме оперативна верифікація причини хвороби обумовлює правильний вибір терапії, скерованої на її знешкодження. Проте існуючі в медицині діагностичні підходи у більшості не є оперативними, і на визначення етіологічного чинника (віруси, бактерії тощо) витрачаються значні терміни часу.

Отримані результати досліджень на сучасному етапі розвитку знань з генетики та токсикології дозволяють фахівцям стверджувати про наявність зв'язків між етіологічним чинником, що потрапив до організму хворого, та токсикометричними характеристиками речовин, які накопичилися у кров'яному руслі в результаті дії цього чинника (токсикоз).

Таким чином, подальший пошук та розробка нових підходів оперативної діагностики етіологічного чинника є важливою міждисциплінарною проблемою, вирішення якої повинно сприяти підвищенню ефективності лікування хворих в цілому.

Вищезазначене дозволяє нам зробити гіпотетичне припущення, що побудова математичного алгоритму дослідження токсикометричних параметрів токсикозу для визначення етіологічного чинника захворювання у пацієнтів з різними захворюваннями є цікавим та перспективним напрямом сучасних математичних досліджень.

Метою досліджень є побудова математичного апарата для визначення причини виникнення захворювання у пацієнтів, тобто для заданого хворого $x=(x_1,...,\ x_{29})$ необхідно визначити чинник виникнення хвороби.

Матеріали та методи дослідження. Для побудови математичної моделі були використані токсикометричні характеристики, які супроводжували перебіг різних захворювань 548 пацієнтів. У всіх хворих за допомогою традиційних загальновизнаних методів дослідження

були встановлені етіологічні чинники (9 етіологічних чинників): аномалії печінки, аномалії нирок, автоімунні/автоалергічні реакції, бактеріальні та вірусні/паразитарні збудники, деструкція тканин різного походження, локальні гіперпластичні процеси в тканинах (новоутворення), системні запальні реакції (SIRS), отрути екзогенного походження.

Постановка задачі. Розглядається побудова алгоритму кластеризації для знаходження етіологічного чинника (причини виникнення захворювання). Важливою умовою для кластеризації даних, наприклад, з використанням дискримінантного аналізу [1], t-критерію Стьюдента [2], ForEll або ж інших методів, є те, що для їх використання розподіл даних має задовольняти нормальному закону. Так як для медичних даних ця умова часто не виконується, то виникає необхідність розробки нового алгоритму, позбавленого вищезазначеного обмеження. Алгоритм кластеризації з використанням побудови еліпсу мінімальної площі [3] дозволяє уникнути обмеження на вхідні дані. Метод опорних векторів дозволяє побудувати кластеризуючу гіперплощину у довільному просторі. Запропонований метод кластеризації є синтезом методів і використовує для кожної пари кластерів, що порівнюються, той метод, який дозволяє якнайкраще розпізнати обидва кластери.

2. Комбінований метод розпізнавання етіологічного чинника токсемії

Ідея визначення причини виникнення захворювання полягає в попарному порівнянні кожного з дев'яти кластерів з усіма іншими різними методами та у відборі для кожної пари кластерів того алгоритму, що найкраще їх розділяє. Для побудови алгоритму розпізнавання причини виникнення захворювання нами були використані такі алгоритми кластеризації: з застосуванням t-критерію Стьюдента; з застосуванням дискримінантного аналізу; ForEll; еліпса мінімальної площі та методу опорних векторів (SVM).

Побудова алгоритму класифікації з застосуванням дискримінантного аналізу

Знаходяться прості дискримінантні функції

$$d_{ik} = b_{k0} + b_{k1}x_{i1} + ... + b_{kp}x_{ip} + \ln q_k, \qquad (1)$$

де $k = 1,..., g$, $b_k = (b_{k1},..., b_{kp})$ і b_{k0} – коефіцієнти k-тої класифікуючої функції i-го об'єкта;

$$b_k = \bar{x}_k' \sum^{\wedge},$$

де $\sum^{\wedge}$ – коваріаційна матриця, $\bar{x}_k = \bar{x}_{k1},...,\bar{x}_{kp}, k = 1,...,g$.

$$b_{k0} = -\frac{1}{2}\bar{x}_k' \sum^{\wedge -1} \bar{x}_k; k = 1,...,g.$$

Коефіцієнти b_{ki} в даному випадку інтерпретуються як параметри, що характеризують нахил гіперплощини з координатними осями, а b_{k0} називається порогом і відповідає відстані від гіперплощини до початку координат.

При використанні простих класифікуючих функцій $x_i = (x_{i1},..., x_{ip})$ буде відноситися до етіологічного класу, в якого значення d виявиться більшим.

Важливим етапом дискримінантного аналізу є виявлення змінних, що входять в дискримінантну функцію. В нашому дослідженні використовували метод покрокового дискримінантного аналізу, в якому змінні вводяться послідовно, виходячи із їхньої здатності розділяти дискримінантні етіологічні класи. Тобто при покроковому аналізі з «включенням» змінних на кожному кроці переглядаються всі змінні і встановлюється одна з них, яка вносить найбільший вклад у розділення етіологічних класів. Ця змінна включається в

модель і переходить до наступного кроку. Покрокова процедура дискримінантного аналізу базується на F-статистиці.

Побудова алгоритму класифікації з застосуванням t-критерію Стьюдента

Використовується двовибірковий t-критерій для незалежних вибірок:

$$t_{i_{lm}} = \frac{\left| m_{i_l} - m_{i_m} \right|}{\sqrt{\frac{\sigma_{i_l}^2}{N_l} + \frac{\sigma_{i_m}^2}{N_m}}},$$

де m_{i_l}, m_{i_m} – математичні сподівання по i-тому фактору;

$\sigma_{i_l}^2$ та $\sigma_{i_m}^2$ – стандартні відхилення по i-тому фактору;

N_l та N_m – кількість хворих в l та m -етіологіях відповідно.

Тобто висуваються дві гіпотези:

H_0 – головна, нульова гіпотеза, в якій висувається припущення, що етіології неможливо відрізнити по i-тому фактору;

H_1 – альтернативна гіпотеза до H_0 – обидві вибірки відрізняються по i-тому фактору.

Так як для використання t-критерію Стьюдента необхідно, щоб вихідні дані задовольняли нормальному розподілу, треба провести перевірку нормальності даних за допомогою критерію Колмогорова-Смірнова з рівнем значущості, рівним 0,05, а також для зменшення розмірності вхідних даних провести кореляційний аналіз залежності x_i від кластера. Для подальшого аналізу використовуються лише ті фактори, які корелюють.

Побудова алгоритму класифікації з застосуванням алгоритму ForEll

Алгоритм ForEll (форель) є прикладом евристичного алгоритму класифікації. В основі його роботи лежить використання гіпотези компактності: схожі об'єкти набагато частіше лежать в одному класі, ніж в різних, або класи утворюють компактно локалізовані підмножини у просторі об'єктів. Ціль роботи алгоритму ForEll знати таке розбиття множини об'єктів на класи, щоб величина $P = \sum_{i=1}^{n} P_n$ була мінімальною, де P_n – відстань між центром n-го класу і всіма точками класу. Робота алгоритму заключається в переміщенні гіперсфери в геометричному просторі до тих пір, поки не отримаємо стійкий цент мас.

Побудова алгоритму класифікації з застосуванням еліпса мінімальної площі

Ідея полягає в знаходженні рівняння еліпса [3, 4], який найкращим чином буде кластеризувати два класи. Еліпс мінімальної площі будується в двовимірному просторі, тобто перебором по всіх x_i, x_j, $i \neq j, i, j = \overline{1,29}$.

Крок 1

Для $l, m \in I, l \neq m$ будуються випуклі оболонки для кожного з класів g_1, g_2 по x_{ikm}. Якщо випуклі оболонки не перетинаються або їх можна розділити кривою другого порядку, то l, m переходять до наступного кроку, в протилежному випадку l, m відкидаються.

Крок2

Для отриманих l, m будується еліпс мінімальної площі e_1 для g_1.

Крок 3

Для отриманих l, m будується еліпс мінімальної площі e_2 для g_2.

У результаті отримуються кластеризуючі еліпси, що дозволяють однозначно визначити, до якого з двох класів буде віднесено хворого за допомогою перевірки.

Якщо з показниками l та m виконується умова, що x_{ikm} лежить в середині e_1, то хворого буде віднесено до класу g_1, в іншому випадку виконується аналогічна перевірка для e_2.

Однією з основних переваг використання алгоритму з побудовою еліпса мінімальної площі є те, що при порівнянні двох класів існує кілька пар $x_i, x_j \; i \neq j, i, j = \overline{1,29}$, тобто будується декілька кластеризуючих еліпсів, що дозволяє збільшити точність.

Так як для таких пар показників l, m визначається кілька еліпсів, то запропонований алгоритм дозволяє з великою точністю визначити етіологічний чинник токсемії у хворого.

Побудова алгоритму класифікації з застосуванням методу опорних векторів

Навчання методу опорних векторів [5] полягає у знаходженні гіперплощини $\langle w, x \rangle - b = 0$, що найкращим чином розділяє два класи, які не перетинаються, де вектор $w = (w_1, w_2, ..., w_n) \in \Re^n$ та скалярний поріг $b \in \Re^n$ – параметри алгоритму, а $x = (x_1, x_2, ..., x_n) \in \Re^n$ – показники хворого.

Тобто фактично після перетворень отримаємо двоїсту задачу мінімізації квадратичного функціоналу:

$$-L(\lambda) = -\sum_{i-1}^{l} \lambda_i + \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \lambda_i \lambda_j y_i y_j (\langle x_i, x_j \rangle) \longrightarrow \min_{\lambda},$$

$$\lambda_i \geq 0, i = 1, ..., l,$$

$$\sum_{I=1}^{L} \lambda_I y_i = 0,$$

де $\lambda = (\lambda_1, \lambda_2, ..., \lambda_l)$ – вектор двоїстих змінних, а y_i – значення приналежності x_i до першого або другого класу. В результаті алгоритм класифікації може бути записаний у вигляді $\alpha(x) = sign(\sum_{i-1}^{l} \lambda_i y_i \langle x_i, x \rangle - b)$, тобто розв'язується задача з жорстким зазором. Так як гарантувати, що вибірка буде лінійно роздільна, в нашому випадку достатньо важко, тому використовуємо перехід вихідних даних до нового спрямляючого простору, що задається поліноміальним ядром:

$$K(x, x') = (\langle x, x' \rangle + 1)^2.$$

3. Застосування алгоритмів розпізнавання на прикладі етіологічних факторів, "Аномалії нирок" та «Екзогенних токсинів»

t-критерій Стьюдента

Першим кроком до застосування t-критерію Стьюдента є проведення кореляційного аналізу та відбору тих параметрів, що корелюють із зазначеними етіологіями: «Аномалії нирок» та «Екзогенні токсини». В результаті з 29 параметрів залишилось 10. Потім проводиться перевірка, чи розподілені дані за нормальним законом за допомогою критерію Колмогорова-Смірнова з достовірністю 95%. В табл. 1 наведено ті показники, що залишились після перших кроків та до яких застосовується t-критерій Стьюдента.

Таблиця 1. Значимі показники для t-критерію Стьюдента

НСТ-Гсп	НСТ-Мсп	АРоЛ-п	АРоЛ-г	АРоЛ-с	АРоЛс-СВ	ЦАЛа

Дискримінантний аналіз

Для застосування дискримінантного аналізу обов'язковими є перевірка розподілу даних та відбір тільки тих показників, що задовольняють нормальний розподіл. Далі проводиться «покроковий дискримінантний аналіз з включенням» для виявлення тих змінних, що вносять найбільший вклад у розділення двох класів. Отримано 8 показників, за якими і будувалися прості дискримінантні функції. В табл. 2 записані коефіцієнти простих дискримінантних функцій для кожного класу.

Таблиця 2. Коефіцієнти простих дискримінантних функцій

Показники	Аномалії нирок	Екзогенні токсини
ЦАЛа-МВ	0,2223	0,1398
АРоЛк-СВ	0,0869	0,1744
ЦАЛп	0,6002	0,4956
ЦИК	0,0570	0,0941
АРоЛс-МВ	0,2644	0,1891
НСТ-М-инд	3,1315	2,9115
НСТ-М-инд	0,1347	0,0743
АРоЛ-п	0,1581	0,2038
АРоЛ-п	0,1497	0,1707
Константа	-45,7252	-39,6512

ForEll

Алгоритм ForEll – знаходження локального максимума щільності точок – застосовується до точок, що знаходяться в діапазоні від 0 до 1. Саме тому спочатку проводиться нормування даних і надалі застосовується алгоритм ForEll до двох класів, повним перебором по всіх $x_i, i = \overline{1,29}$, тобто по два, три, чотири і т.д. показники, для виявлення комбінації тих змінних, що найкращим чином розділяють два обраних класи. В табл. 3 наведено ті 15 показників, що правильно розділяють класи з найбільшою точністю.

Таблиця 3. Значимі показники для алгоритму ForEll

АРоЛг-СВ	АРоЛа-СВ	АРоЛс-СВ	АРоЛк-БВ	АРоЛг-МВ	АРоЛа-МВ	АРоЛс-МВ	ЦАЛп	ЦАЛг
ЦАЛа	ЦАЛс	ЦАЛг-СВ	ЦАЛа-СВ	ЦАЛс-СВ	ЦАЛг-МВ	ЦАЛа-МВ	ЦАЛс-МВ	ЦИК

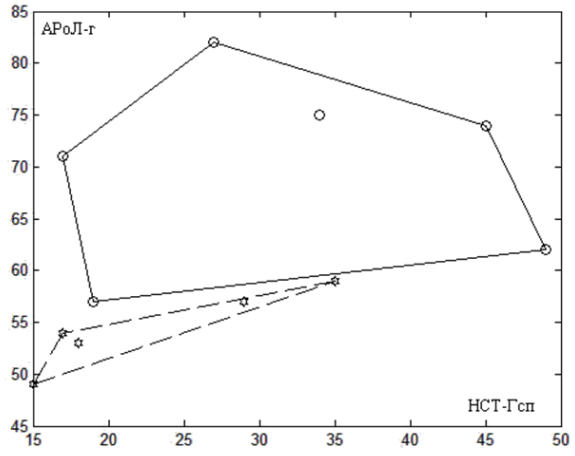

Рис. 1. Показники (АРОЛ-г та НСТ-Гсп), що розділяють випуклі оболонки

Примітка:

1) На рисунках знаком * позначені параметри ендотоксемії у хворих з етіопатогенетичним чинником "Аномалії нирок";

2) Знаком ° – параметри ендотоксемії у хворих з етіопатогенетичним чинником "Екзогенні токсини"

Еліпс мінімальної площі

Алгоритм на основі побудови еліпса мінімальної площі застосовувався до найхарактерніших хворих для кожного з класів. Тобто, з медичної точки зору, було відібрано найтяжчі форми, а далі для отриманих даних будувалися випуклі оболонки і знаходилися ті показники, в яких випуклі оболонки або не перетинались, або їх можна розділити за допомогою кривої другого порядку

$$a_{11}x^2 + 2a_{12}xy + a_{22}y^2 + 2a_{13}x + 2a_{21}y + a_{33} = 0$$

На рис. 1 зображено один із таких випадків.

На основі аналізу побудованих випуклих оболонок для кожного з класів відібрано 204 з 406 пар показників. На рис. 2, 3 зображено приклади побудови еліпсів мінімальної площі для кожного з класів та зображено хворих з контрольної вибірки для наочної демонстрації роботи алгоритму.

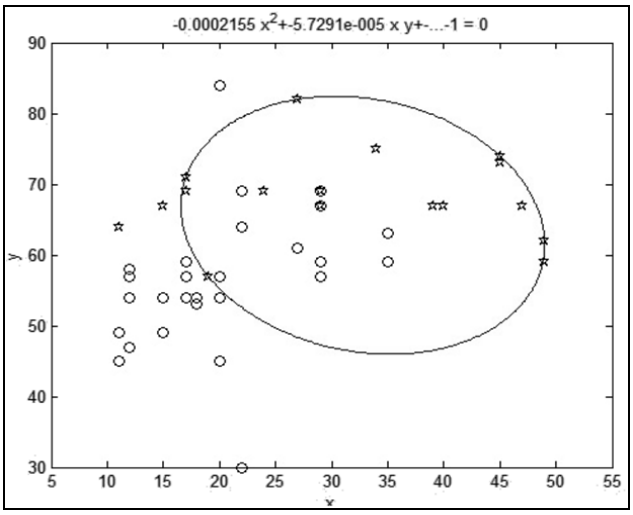

Рис. 2. Еліпс для «Аномалії нирок» за показниками АРОЛ-г та НСТ-Гсп

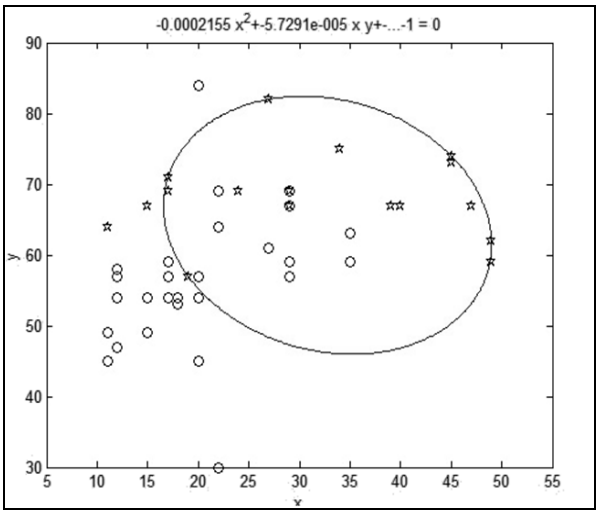

Рис. 3. Еліпс для «Екзогенних токсинів» за показниками АРОЛ-г та НСТ-Гсп

У результаті проведених досліджень отримано 6 рівнянь еліпсів мінімальної площі, що з достовірністю більше 65% правильно класифікують два класи. Один із рівнянь еліпса мінімальної площі, побудований для класу «Аномалії нирок», зображений нижче:

$$f(x,y) = -9{,}2824e-005x^2 - 0{,}00012723xy - 7{,}3637e-005y^2 + 0{,}018738x + 0{,}015551y - 1 = 0.$$

Тобто, якщо показники хворого попадають у середину еліпса, то хворого буде віднесено до класу «Аномалії нирок», у протилежному випадку буде виконуватися перевірка для класу «Екзогенних токсинів».

Важливим є встановлений результат, який указує на існування декількох пар показників за якими можна з імовірністю більше 65% визначити приналежність хворого до одного з класів простою перевіркою потрапляння показників хворого в середину побудованого еліпса мінімальної площі.

Метод опорних векторів

Для визначення тих токсикометричних параметрів, які є значимими при розділенні зазначених двох класів, розв'язується задача знаходження оптимальної розділяючої гіперплощини для токсикометричних показників, згрупованих по $2 - C_{29}^2$, по три – C_{29}^3 тощо. Тобто розв'язується задача з жорстким зазором, проте з застосуванням поліноміального ядра $K(x,x') = (\langle x,x' \rangle + 1)^2$. Це дає нам можливість вважати, що в деякому з просторів вибірка буде лінійно роздільною.

Для заданих класів найкраща кластеризуюча площина будується за всіма 29 показниками. Значення параметрів алгоритму відповідно дорівнюють $b = 18{,}0208$ та матриця $w \in \Re^{57*57}$.

Результати класифікації для класів «Аномалії нирок» та «Екзогенних токсинів»

У наступній таблиці наведено результати роботи кожного з алгоритмів для обраних двох класів.

Отже, аналізуючи отримані результати для кожної пари класів, вибирається алгоритм, що найкращим чином розділяє класи. В даному випадку для кластеризації «Аномалії нирок» та «Екзогенних токсинів» вибирається алгоритм побудови з використанням еліпса мінімальної площі.

Таблиця 4. Результати тестових досліджень розпізнавання етіологічного фактора за допомогою різних алгоритмів

	Розпізнавання етіології Аномалії нирок (%)	Розпізнавання етіології Екзогенні токсини (%
t -критерій Стьюдента	35,24	95,3
Дискримінантний аналіз	45,67	67,5
ForEl	56,4	25,4
Еліпс мінімальної площі	71,2	83,6
Метод опорних векторів	87,4	67,6

4. Комбінований алгоритм розпізнавання

У побудованому алгоритмі використовується синтез методів для розпізнавання причини виникнення захворювання. Це дозволяє значно збільшити точність розпізнавання та покращити розпізнавання тих пар, для яких один із методів не дає бажаних результатів розпізнавання (більше 70%). Тобто, ідея визначення причини захворювання у хворого полягає в попарному порівнянні кожного класу (9 етіологій) попарно з усіма іншими та у відборі того алгоритму, що найкращим чином розділяє вибрані класи. Застосування синтезу методів дозволило у 73 % випадків правильно визначити етіологічний чинник захворювання.

5. Висновки

У статті описано модель визначення причини виникнення захворювання. Розглянуто алгоритми, які лежать в основі побудованої моделі. Наведено приклади роботи алгоритму. Отримані результати наочно демонструють ефективність застосування запропонованих методів:

1. Застосування *t* -критерію Стьюдента для побудови алгоритму кластеризації дозволяє визначити етіологічний чинник захворювання нирок з імовірністю 20%.

2. Використання дискримінантного аналізу для побудови алгоритму кластеризації токсикометричних даних дозволяє визначити етіологічний чинник захворювання у 48% випадків.

3. Побудова алгоритму класифікації з застосуванням алгоритму ForEll обумовлює визначення етіологічного чинника захворювання з імовірністю 42%.

4. Застосування еліпса мінімальної площі для побудови алгоритму кластеризації дозволяє визначити етіологічний чинник захворювання з імовірністю 56%.

5. Використання методу опорних векторів (SVM) для побудови алгоритму кластеризації токсикометричних даних дозволяє визначити етіологічний чинник захворювання у 53% випадків.

6. Застосування синтезу двох методів – еліпс мінімальної площі та метод опорних векторів – є найбільш оптимальним та ефективним підходом для розпізнавання за токсикометричними параметрами токсикозу етіологічного чинника захворювання – 73%.

СПИСОК ЛІТЕРАТУРИ

1. Шкуліпа О.В. Застосування дискримінантного аналізу для визначення етіологічного чинника токсемії у хворого / О.В. Шкуліпа, Б.С. Шейман, Б.В. Рубльов // Вісник Київського національного університету імені Тараса Шевченка (Серія «Фізико-математична»). – 2010. – № 1. – С. 158 – 161.

2. Застосування *t* -критерію Стьюдента в кластеризації точок у багатовимірному просторі / Б.В. Рубльов, О.В. Шкуліпа, Б.С. Шейман [та ін.] // Вісник Київського національного університету імені Тараса Шевченка (Серія «Фізико-математична»). – 2009. – № 3. – С. 180 – 183.

3. Shkulipa O. Determining an etiological toxemia factor using smallest enclosing ellipses / O. Shkulipa // Матеріали 8-ї міжнар. міждисциплінарної наук.-практ. конф. молодих вчених «Шевченківська весна». – Київ, 2010. – С. 91 – 92.
4. Рубльов Б.В. Геометричні властивості еліпса мінімальної площі та деякі суміжні питання / Б.В. Рубльов, Ю.І. Петунін, Ю.Ю. Милейко. – К.: Київський університет, 2000. – 73 с.
5. Nayak A. Handbook of applied algorithm. Solving scientific, engineering and practical problem / A. Nayak, I. Stojmenovic. – Willey interscience, 2008. – 541 p.

Стаття надійшла до редакції 25.06.2010