УДК 004.048+004.94

МЕТОД ОЦІНКИ СТУПЕНЯ БЛИЗЬКОСТІ СКЛАДНИХ ОБ'ЄКТІВ НА ОСНОВІ МОДИФІКОВАНОГО ІНДЕКСУ МАКСИМІЗАЦІЇ ВЗАЄМНОЇ ІНФОРМАЦІЇ

Л. М. Ясінська-Дамрі

Українська академія друкарства,
вул. Під Голоском, 19, Львів, 79020, Україна

Наведено метод оцінки ступеня близькості об'єктів зі складною природою на основі модифікованого індексу взаємної інформації, застосування якого передбачає використання ансамблю методів розрахунку ентропії Шеннона для оцінки взаємної інформації об'єктів, що досліджуються. Остаточне рішення щодо ступеня близькості відповідних об'єктів ухвалюється на основі функції бажаності Харрінгтона, яка містить як компоненти результати застосування окремих методів розрахунку ентропії Шеннона. Оцінка ефективності запропонованого методу здійснювалася за допомогою класифікації об'єктів, що вивчалися, на основі критеріїв якості класифікації даних. Запропоновано структурну блок-схему алгоритму покрокової процедури формування інформативних атрибутів даних за модифікованим критерієм взаємної інформації. Проведено апробацію запропонованого методу із застосуванням даних експресій генів пацієнтів, у яких досліджували можливий рак легенів.

Ключові слова: *ентропія Шеннона, максимізація взаємної інформації, функція бажаності Харрінгтона, експресія генів, бінарна класифікація.*

Постановка проблеми. Сучасні моделі обробки складних даних, що характеризуються великою кількістю атрибутів, наявністю шумової компоненти тощо, здебільшого ґрунтуються на застосуванні ансамблів методів реалізації відповідних процесів. Під час кластеризації атрибутів високорозмірних даних оцінка якості групування відповідних атрибутів або об'єктів на основі використання внутрішніх критеріїв якості кластеризації має великий відсоток суб'єктивізму. Навіть більше: сучасним методам кластеризації даних властива похибка відтворюваності: задовільні результати, отримані на основі певних даних, не повторюються, якщо застосовувати інші подібні. Підвищити об'єктивність обробки даних у цьому разі можливо завдяки комплексному використанню методів інтелектуального аналізу даних та машинному навчанню, водночас виникає необхідність вибору оптимальних критеріїв оцінки якості реалізації відповідних етапів.

Аналіз останніх досліджень та публікацій. Застосуванню ансамблю методів інтелектуального аналізу даних і машинного навчання для обробки складних даних сьогодні присвячено чимало наукових праць. Так, у статті [1] наведено результати досліджень щодо застосування взаємної інформації під час кластеризації білків,

водночас її якість оцінювали за результатами класифікації відповідних об'єктів. Результати аналізу комплексного використання методів кластеризації та класифікації під час вибору профілів експресій генів, що дають змогу із найбільшою точністю ідентифікувати відповідну хворобу, подані у науковій статті [2]. В огляді [3] автори провели порівняльний аналіз сучасних гібридних моделей вибору інформативних атрибутів для розв'язання проблеми подальшої класифікації об'єктів, які досліджували щодо різних типів раку, вивчили ефективність різних моделей редукції атрибутів від фільтрації за допомогою оцінки рівня інформативності відповідного атрибута на основі статистичного аналізу до комплексного застосування кластерного аналізу та методів класифікації даних. Основним критерієм в усіх випадках були оцінки якості класифікатора. Водночас проаналізовано різні комбінації методів інтелектуального аналізу даних та машинного навчання. Однак потрібно зазначити, що зараз проблема об'єктивного вибору атрибутів, що дають змогу з високою точністю класифікувати відповідні об'єкти, однозначного розв'язання не має.

Мета статті — комплексне застосування модифікованого індексу максимізації взаємної інформації та методів бінарної класифікації для вибору найбільш інформативних із погляду роздільної здатності атрибутів, які дають змогу з високою точністю ідентифікувати досліджувані об'єкти.

Виклад основного матеріалу дослідження. Нехай вихідні дані експресій генів подано у вигляді матриці:

$$G = \left\{ e_{ij} \right\}, i = \overline{1,n}; j = \overline{1,m},$$

де m — кількість атрибутів, що визначають стан відповідного об'єкта; n — кількість об'єктів або зразків, що досліджуються.

У цьому разі критерієм формування підмножин атрибутів може бути цільова функція:

$$C\left(e_s, e_p\right) = \min f\left(e_s, e_p\right), \tag{1}$$

де $e_s e_p$ — вектори атрибутів s і p відповідно; $f(\cdot)$ — функція подібності, що використовується для оцінки ступеня близькості векторів атрибутів s і p.

Очевидно, що у цьому разі вибір методу визначає функція подібності, яка властива такому методу. Формально взаємну інформацію двох векторів дискретних змінних e_s *і* e_p можна оцінити у такий спосіб [4]:

$$I\left(e_s, e_p\right) = \sum_{x \in e_s} \sum_{y \in e_p} p(x,y) \log_2 \left(\frac{p(x,y)}{p(x)\,p(y)} \right), \tag{2}$$

де $p(x,y)$ — функція спільного розподілу ймовірностей у векторах e_s *і* e_p; $p(x)$ і $p(y)$ — функції розподілу ймовірностей векторів e_s *і* e_p відповідно. Під час застосування ентропії взаємна інформація може бути виражена так:

$$I\left(e_s, e_p\right) = H\left(e_s\right) + H\left(e_p\right) - H\left(e_s, e_p\right), \tag{3}$$

де $H(e_s)$, $H(e_p)$ і $H(e_s,e_p)$ — ентропії векторів e_s, e_p і спільна ентропія даних векторів відповідно. Ентропія у цьому разі встановлюється як міра невизначеності відповідного стану системи і розраховується за формулою Шеннона [5]:

$$H(e) = -\sum_{i=1}^{m} p_i(e) \log_2 p_i(e), \tag{4}$$

де m — кількість досліджуваних зразків (довжина профілю експресій відповідного гена); $p_i(e)$ — ймовірність реалізації відповідного дискретного значення експресії i-ї змінної. У цьому разі спільна ентропія Шеннона розраховується за формулою:

$$H\left(e_s, e_p\right) = -\sum_{i_s=1}^{m}\sum_{i_p=1}^{m} p_i\left(e_s, e_p\right)\log_2 p_i\left(e_s, e_p\right) \quad , \tag{5}$$

де $p_i(e_s,e_p)$ — спільна ймовірність появи і-го значення атрибутів e_s і e_p.

Як засвідчує аналіз літературних джерел [6], наявні методи розрахунку ентропії Шеннона різняться за способом розрахунку ймовірності реалізації відповідного стану системи і загалом можуть бути поділені на дві групи. Перша група методів ґрунтується на оцінці частот виникнення відповідного стану системи. Методи другої групи передбачають розрахунок ентропії вектора змінних безпосередньо, без застосування частот виникнення відповідних станів системи. Структурну блок-схему найбільш поширених методів оцінки ентропії Шеннона зображено на рис. 1. Потрібно зазначити, що вибір методу оцінки ентропії Шеннона, який використовується для розрахунку індексу взаємної інформації, може впливати на результат у процесі подальшої класифікації об'єктів. Підвищити об'єктивність у цьому разі можливо, якщо застосовувати ансамбль методів розрахунку ентропії Шеннона. Для обчислення комплексного критерію у цій статті запропоновано функцію бажаності Харрінгтона [7], яка зараз успішно використовується у різних галузях наукових досліджень для розв'язання багатокритерійних задач. Функцію бажаності Харрінгтона задає таке рівняння:

$$d=exp(-exp(-Y)), \tag{6}$$

де значення безрозмірного параметра Y змінюється в діапазоні від -2 (d = 0) до 5 (d = 1). Тобто значення бажаності (d) під час використання цього методу змінюється в інтервалі від 0 до 1.

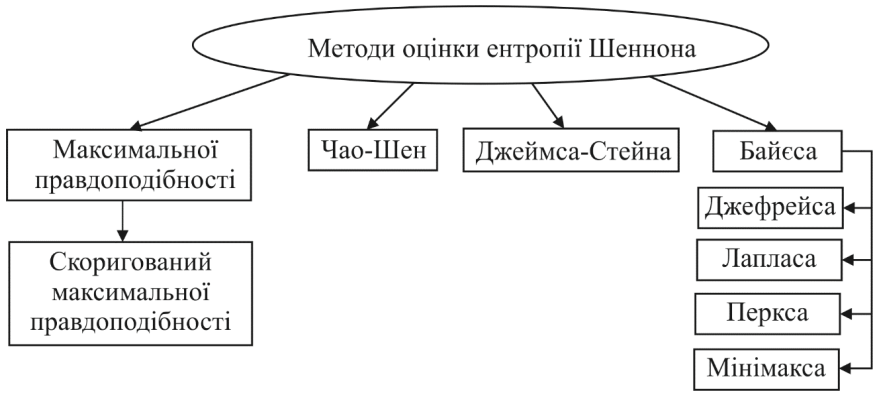

Рис. 1. Структурна блок-схема методів оцінки ентропії Шеннона

На рис. 2 зображено графік функції бажаності Харрінгтона. Точки перетину функції на графіку 0,63=1-1/e і 0,37=1/e відповідають граничним значенням бажаності всередині інтервалу зміни функції. Значення бажаності 0,37 зазвичай відповідає

межі допустимих значень, значення від 0,37 до 0,63 — задовільній бажаності, від 0,63 до 0,8 — добрій бажаності, вище 0,8 — відмінній бажаності.

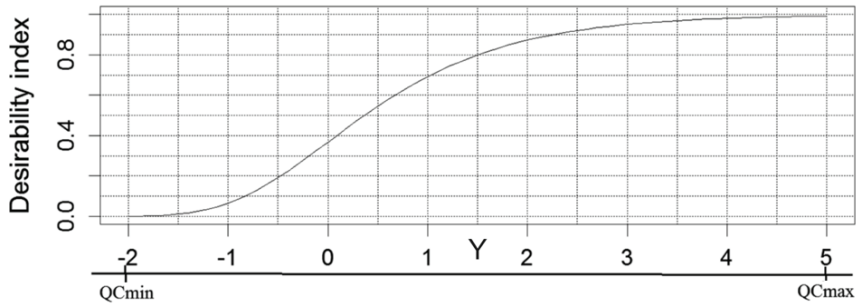

Рис. 2. Функція бажаності Харрінгтона

Алгоритм розрахунку комплексного критерію на основі функції бажаності Харрінгтона передбачає такі кроки:

1. Трансформація шкал відповідних критеріїв, що застосовуються в межах досліджень, у шкалу безрозмірного показника Y, значення якого змінюється у інтервалі від -2 до 5, відповідно до лінійного рівняння:

$$Y = a + b \cdot QC, \tag{7}$$

де a і b — коефіцієнти, що визначаються емпірично, із застосуванням граничних значень даних, що досліджуються:

$$\begin{aligned} Y_{\min} &= a + b \cdot QC_{\min} \\ Y_{\max} &= a + b \cdot QC_{\max} \end{aligned} \tag{8}$$

2. Розрахунок параметра Y для кожного значення відповідного критерію:

$$Y_i = a + b \cdot QC_i. \tag{9}$$

3. Розрахунок приватних бажаностей для кожного значення відповідних критеріїв:

$$d_i = exp(-exp(-Y_i)), \tag{10}$$

4. Розрахунок узагальненого індексу як середнє геометричне усіх приватних бажаностей критеріїв якості, що використовуються у процесі розв'язання цієї проблеми:

$$GMI = \sqrt[n]{\prod_{i=1}^{n} d_i}. \tag{11}$$

Найбільш високі значення узагальненого індексу бажаності відповідають максимальному значенню взаємної інформації за групою методів оцінки ентропії Шеннона, що використовуються у процесі моделювання.

Практична реалізація цього алгоритму проводилася на основі програмного забезпечення R [8] із застосуванням даних профілів експресій генів GSE19188 пацієнтів, у яких досліджували можливу ранню стадію раку легенів [9]. Інформацію отримали за допомогою використання ДНК-мікрочипів, і вона містила 156 мікрочипів. Відповідно до анотації даних, 65 мікрочипів охоплювали дані експресій генів здорових пацієнтів, а на 91 мікрочипі були ідентифіковані експресії генів

пацієнтів, хворих на рак легенів (легка форма). У статті [2] автори навели результати дослідження щодо комплексного застосування методів ієрархічної кластеризації та бінарної класифікації для виокремлення найбільш інформативних профілів експресій генів, що дають змогу з високою точністю ідентифікувати об'єкти, що вивчаються. Із 54 675 генів було виділено 401 із точністю класифікації 93,5 %. Ця підмножина профілів експресій генів була використана в процесі моделювання.

Якість класифікації даних під час застосування відповідних методів оцінки ентропії Шеннона проводилася із використанням критеріїв, що містять як компоненти похибки першого та другого роду. Матриця похибок (confusion matrix), яка охоплює похибки як першого, так і другого роду, наведена у табл. Оцінка результатів класифікації об'єктів здійснювалася на основі застосування таких критеріїв:

– Точність (Accuracy (AC) — визначається як загальна ймовірність прогнозування класифікатором правильних результатів:

$$AC = \frac{TP + TN}{TP + FP + TN + FN}. \tag{12}$$

Таблиця

Матриця похибок під час діагностики та класифікації пухлини у пацієнтів, у яких перевіряли можливий рак легенів

Реальний стан пацієнта за результатами діагностування	Результат класифікації об'єктів	
	Хворий (пухлина)	Здоровий
Пухлина	Правильно прийнята (True positives TP)	Неправильно знехтувана (False negatives FN)
Здоровий	Неправильно прийнята (False positives FP)	Правильно знехтувана (True negatives TN)

F-міра (F-measure (F) — визначається як гармонійне середнє влучності (Precision (PR) — positive predicted values) та повноти (Recall or Sensitivity (RC):

$$F = \frac{2 \cdot PR \cdot RC}{PR + RC}, \tag{13}$$

де

$$PR = \frac{TP}{TP + FP}; \; RC = \frac{TP}{TP + FN}.$$

– Коефіцієнт кореляції Метьюса (Matthews correlation coefficient (MCC) — використовується у машинному навчанні як міра ефективності бінарного класифікатора [10]:

$$MCC = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}}. \tag{14}$$

Вищі значення критеріїв (12)–(14) відповідають вищій ефективності класифікатора.

На рис. 3 зображено блок-схему алгоритму вибору групи інформативних профілів експресій на основі комплексного застосування методу класифікації об'єктів, що досліджуються, і модифікованого критерію взаємної інформації, який

розраховується за формулами (7)–(11). Практична реалізація алгоритму передбачає такі етапи:

Етап І. Формування вихідних даних профілів експресій генів та вектора методів розрахунку ентропії Шеннона.

1.1. Формування матриці профілів експресій генів, де рядки репрезентовані об'єктами, що досліджуються, а стовпці — це гени з відповідними значеннями експресії, які формують відповідний профіль.

1.2. Формування вектора методів розрахунку ентропії Шеннона, які використовуються у процесі розрахунку критерію взаємної інформації. Вибір першого методу.

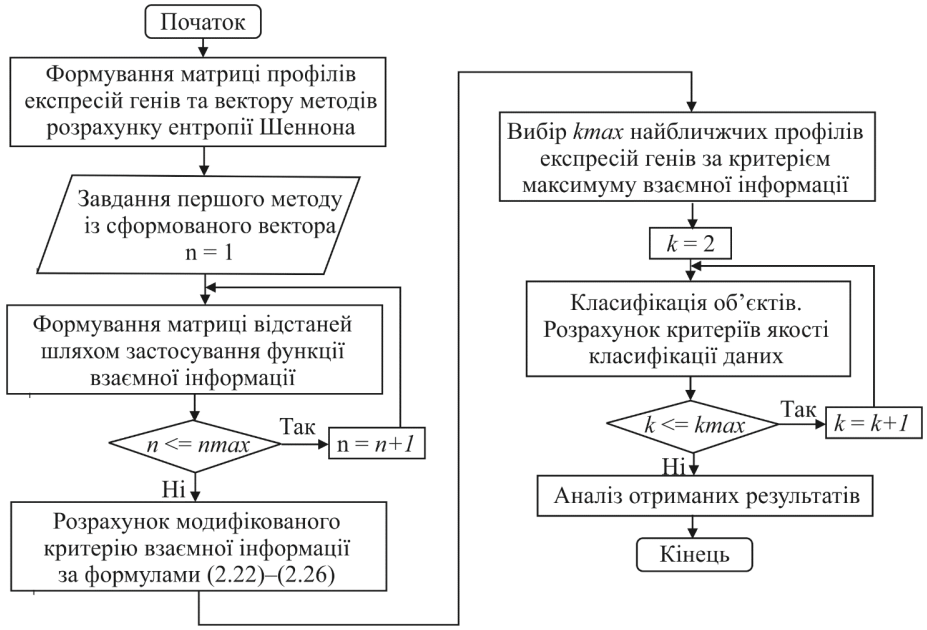

Рис. 3. Алгоритм формування матриці найбільш інформативних профілів експресій генів за модифікованим критерієм максимізації взаємної інформації

Етап ІІ. Формування таблиці значень критеріїв взаємної інформації, які розраховані під час застосування різних методів оцінки ентропії Шеннона.

2.1. Вибір першого профілю експресій генів із вихідної таблиці даних.

2.2. Розрахунок значення взаємної інформації між першим та подальшими профілями експресій генів за послідовного застосування усіх методів розрахунку ентропії Шеннона.

Етап ІІІ. Розрахунок модифікованого узагальненого значення взаємної інформації.

3.1. Формування трансформованої шкали безрозмірного показника Y за допомогою розрахунку коефіцієнтів a і b відповідно до системи рівнянь (8).

3.2. Розрахунок показника Y для усіх значень взаємної інформації, що утворюють таблицю даних, отриманих на етапі 2 за формулою (9).

3.3. Розрахунок приватних бажаностей для кожного методу розрахунку ентропії Шеннона за формулою (10).

3.4. Формування вектора узагальненого модифікованого критерію взаємної інформації за формулою (11).

Етап IV. Класифікація даних. Розрахунок критеріїв якості класифікації даних.

4.1. Вибір двох найближчих за критерієм (11) профілів експресій генів.

4.2. Застосування алгоритму класифікації даних. Розрахунок критеріїв якості класифікації даних на відповідному етапі.

4.3. Збільшення кількості найближчих до першого профілів експресій генів на одиницю. Якщо кількість профілів експресій генів менша за максимальну кількість, відбувається перехід на крок 4.2 цієї процедури. У протилежному разі відбувається перехід на наступний етап.

Етап V. Аналіз отриманих результатів.

5.1. Побудова діаграм залежності критеріїв якості класифікації об'єктів від кількості генів, виділених за модифікованим критерієм максимуму взаємної інформації.

5.2. Аналіз отриманих результатів.

На рис. 4 наведено діаграми залежності відповідних критеріїв якості класифікації об'єктів від кількості профілів експресії генів, виділених за модифікованим критерієм максимізації взаємної інформації. Кількість генів у процесі моделювання змінювалася від 2 до 100. Класифікатор «Випадковий ліс» (Random Forest (RF) [11], ефективність якого для бінарної класифікації профілів експресій генів доведена у статті [2], був використаний у процесі моделювання.

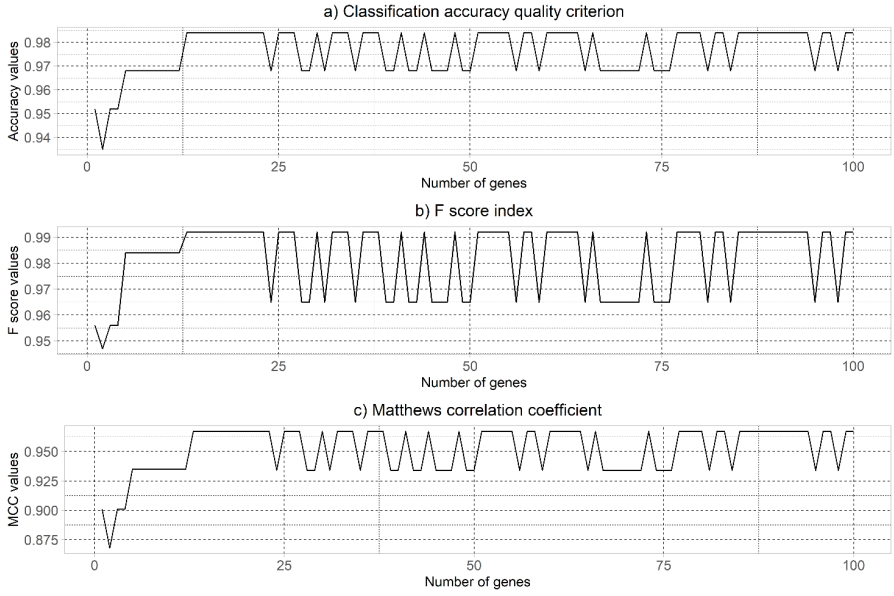

Рис. 4. Діаграми розподілу значень критеріїв якості класифікації даних при різних кількостях профілів експресій генів

Аналіз отриманих результатів дає змогу зробити висновок, що запропонований метод сприяє об'єктивному виділенню груп профілів експресій генів, що допомагають адекватно, з високою точністю (99 %), класифікувати об'єкти, які досліджуються. Підвищення об'єктивності у цьому разі зумовлене коректним використанням ансамблю методів розрахунку ентропії Шеннона, значення яких застосовується під час оцінки взаємної інформації відповідних профілів експресій генів.

Висновки. Запропоновано метод виділення інформативних атрибутів із високорозмірних даних за модифікованим індексом взаємної інформації, особливість якого полягає у вищій об'єктивності процесу формування підмножин атрибутів завдяки коректному застосуванню ансамблю методів оцінки ентропії Шеннона. Проведено апробацію методу із застосуванням цих профілів експресій генів пацієнтів, у яких досліджували можливий рак легенів. Ефективність запропонованого методу оцінювали за допомогою бінарної класифікації об'єктів із використанням алгоритму «Випадковий ліс». Результати моделювання засвідчили високу результативність запропонованого методу. Точність класифікації об'єктів, що містили виділені профілі експресій генів, сягала 99 %.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Trivial and nontrivial error sources account for misidentification of protein partners in mutual information approaches / Pontes C., Andrade M., Fiorote J., Treptow W. *Scientific Reports*. 2021. Vol. 11 (1), art. no. 6902.
2. Babichev S., Škvor J. Technique of Gene Expression Profiles Extraction Based on the Complex Use of Clustering and ClassificationMethods. *Diagnostics*. 2020. Vol. 10 (8), art. no. 584.
3. Almugren N., Alshamlan H. A survey on hybrid feature selection methods in microarray gene expression data for cancer classification. *IEEE Access*. 2019. Vol. 7, art. no. 8736725. Pp. 78533–78548.
4. Thomas M. C., Joy A. T. Elements of Information Theory. Wiley, 2nd Edition, 2006. 792 p.
5. Shannon C. E. A mathematical theory of communication. *Bell System Technical Journal*. 1948. Vol. 27. Pp. 379–423, 623–656.
6. Hausser J., Strimmer K. Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. *Journal of Machine Learning Research*. 2009. Vol. 10. Pp. 1469–1484.
7. Harrington J. The desirability function. *Industrial Quality Control*. 1965. Vol. 21 (10). Pp. 494–498.
8. Ihaka R., Gentleman R. R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics*. 1996. Vol. 5 (3). Pp. 299–314.
9. Hou J., Aerts J., den Hamer B. et al. Gene expression-based classification of non-small cell lung carcinomas and survival prediction. *PLoS ONE*. 2010. Vol. 5, art. no. e10312.
10. Matthews B. W. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *BBA–Protein Struct*. 1975. Vol. 405. Pp. 442–451.
11. Breiman L. Random forests. *Mach. Learn*. 2001. Vol. 45. Pp. 5–32.

REFERENCES

1. Pontes, C., Andrade, M., Fiorote, J., & Treptow, W. (2021). Trivial and nontrivial error sources account for misidentification of protein partners in mutual information approaches: *Scientific Reports*, 11 (1), art. no. 6902 (in English).

2. Babichev, S., & Škvor, J. (2020). Technique of Gene Expression Profiles Extraction Based on the Complex Use of Clustering and ClassificationMethods: Diagnostics, 10 (8), art. no. 584 (in English).
3. Almugren, N., & Alshamlan, H. (2019). A survey on hybrid feature selection methods in microarray gene expression data for cancer classification: IEEE Access, 7, art. no. 8736725, 78533–78548 (in English).
4. Thomas, M. C., & Joy, A. T. (2006). Elements of Information Theory. Wiley, 2nd Edition (in English).
5. Shannon, C. E. (1948). A mathematical theory of communication: Bell System Technical Journal, 27, 379–423, 623–656 (in English).
6. Hausser, J., & Strimmer, K. (2009). Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks: Journal of Machine Learning Research, 10, 1469–1484 (in English).
7. Harrington, J. (1965). The desirability function: Industrial Quality Control, 21 (10), 494–498 (in English).
8. Ihaka, R., & Gentleman R. (1996). R: a language for data analysis and graphics: Journal of Computational and Graphical Statistics, 5 (3), 299–314 (in English).
9. Hou, J., Aerts, J., & den Hamer, B. et al. (2010). Gene expression-based classification of non-small cell lung carcinomas and survival prediction: PLoS ONE, 5, art. no. e10312 (in English).
10. Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of t4 phage lysozyme: BBA–Protein Struct, 405, 442–451 (in English).
11. Breiman, L. (2001). Random forests: Mach. Learn, 45, 5–32 (in English).

doi: 10.32403/1998-6912-2021-1-62-42-51

METHOD OF THE PROXIMITY DEGREE OF COMPLEX OBJECTS EVALUATION ON THE BASIS OF THE MODIFIED INDEX OF MUTUAL INFORMATION MAXIMIZATION

L. M. Yasinska-Damri

Ukrainian Academy of Printing,
19, Pid Holoskom St., Lviv, 79020, Ukraine
Lm.yasinska@gmail.com

The paper presents a method to estimate the proximity degree of complex objects based on a modified index of mutual information, the use of which involves an application of a set of methods for calculating Shannon's entropy to assess the mutual information of examined objects. The final decision concerning the proximity degree of the respective objects was done based on the Harrington desirability function, which contains, as the components, the results of various methods applied to calculate Shannon's entropy. The evaluation of the effectiveness of the proposed method was carried out by classifying

the studied objects using the classification quality criteria. The random forest binary classifier was applied for data classification during the simulation procedure. The structural block chart of the step-by-step algorithm to form informative data attributes according to the modified index of mutual information has been offered. The proposed method has been tested using the data of gene expressions of patients studied for lung cancer. The application of the proposed technique assumed the stepwise increasing the nearest gene expression profiles from 2 to 100 with the classification of the examined objects at each step of this procedure implementation with calculation classification quality criteria. The accuracy, F-score and Matthews correlation coefficient were used as the classification criteria. The diagrams of these criteria values variation versus the number of gene expression profiles were created as the simulation results. The analysis of the obtained results has shown the high effectiveness of the proposed method since the accuracy of the data classification is achieved by more than 99%. The increase of objectivity, in this case, is due to the correct application of a set of methods for calculating Shannon's entropy, the value of which was used for assessing the mutual information of the respective gene expression profiles.

Keywords: *Shannon entropy, maximization of mutual information, Harrington desirability function, gene expression, binary classification.*

Стаття надійшла до редакції 28.04.2021.
Received 28.04.2021.