**УДК 004.032.26** **https://doi.org/10.33082/td.2017.1-1.11**

# ОПТИМІЗАЦІЯ ОЦІНЮВАННЯ ІНФОРМАТИВНОСТІ МЕДИЧНИХ ПОКАЗНИКІВ НА ОСНОВІ ГІБРИДНОГО ПІДХОДУ

# ОПТИМИЗАЦИЯ ОЦЕНКИ ИНФОРМАТИВНОСТИ МЕДИЦИНСКИХ ПОКАЗАНИЙ НА ОСНОВЕ ГИБРИДНОГО ПОДХОДА

# OPTIMIZATION OF EVALUATION OF THE INFORMATIVITY OF MEDICAL INDICATORS ON THE BASIS OF THE HYBRID APPROACH

**Є.В. БОДЯНСЬКИЙ,** докт.техн.наук,
**І.Г. ПЕРОВА,** канд.техн.наук, **Г.В. СТОЙКА**
*Харківський національний університет радіоелектроніки, Україна*

*У роботі пропонується новий підхід до оцінки інформативності медичних показників, відмінною рисою якого є оптимальне поєднання методів компресії вихідного простору ознак (Feature Extraction) і методів вибору найбільш інформативних показників з набору наявних (Feature Selection). Таким чином, досягається лінгвістична інтерпретовність простору ознак і оптимальний вибір скороченого набору ознак.*

***Ключеві слова:*** *метод головних компонент, інформативність показників, медичне діагностування, власний вектор, власне значення.*

*В работе предлагается новый подход к оценке информативности медицинских показателей, отличительной особенностью которого является оптимальное сочетание методов компрессии исходного пространства признаков (Feature Extraction) и методов выбора наиболее информативных показателей из набора имеющихся (Feature Selection). Таким образом, достигается лингвистическая интерпретируемость пространства признаков и оптимальный выбор сокращенного набора признаков.*

***Ключевые слова:*** *метод главных компонент, информативность показателей, медицинское диагностирование, собственный вектор, собственное значение.*

*Feature Selection task is one of most complicated and actual in Data Mining area. Any approaches for it solving are based on non-mathematical and presentative hypothesis. New approach for evaluation of medical features information quantity, based on optimal combination of Feature Selection and Feature Extraction methods. This approach permits to produce optimal reduced number of features with linguistic interpreting of each ones. Hybrid system of Feature Selection/Extraction is proposed.*

*This system is numerically simple, can produce Feature Selection/ Extraction with any number of features using standard method of principal component analysis and calculating distance between first principal component and all medical features.*

***Keywords:*** *Principal Component Analysis, Feature Selection, Feature Extraction, Medical Data Mining, Eigen Vector, Eigen Value.*

**Вступ.** В цей час задача оцінки інформативності вхідних показників, як і задача стиснення інформації при проведенні діагностування в медицині є однією з найбільш актуальних і складних.

Необхідність такої обробки обумовлена тим, що медичні вибірки даних досить часто містять занадто велику кількість ознак при малому числі спостережень (пацієнтів), що істотно обмежує можливості існуючих методів для проведення подальшого діагностування.

**Аналіз основних досягнень і літератури.** Необхідним елементом загальної задачі Data Mining є етап попередньої обробки даних, в якому реалізуються такі завдання як Data Reduction (очищення і заповнення пропусків), Feature Selection (вибір найбільш інформативних ознак з набору наявних) та Feature Extraction (стиснення або компресія вихідного простору ознак) [1-7].

Одним з найважливіших етапів є завдання Feature Selection / Feature Extraction бо від вибору конкретних ознак і від їх кількості часто залежить як якість класифікації (діагностування), так і в принципі можливість її проведення.

На сьогоднішній день найбільш чітке математичне обґрунтування отримали методи Feature Extraction такі, як метод головних компонент (Principal Component Analysis) [8-10], дискримінантний аналіз, Principal Manifolds Analysis [11].

Для вирішення завдання компресії даних (Feature Extraction) часто використовується нейромережевий підхід, що особливо актуально в завданнях глибинного навчання (енкодери типу Bottle Neck, Restricted Boltzmann Machine та інші) [11-12]. Однак, при проведенні медичного діагностування надзвичайно важливим є можливість інтерпретації результатів на рівні вхідних факторів. І тут на перший план виходять завдання виділення найбільш інформативних ознак з набору наявних (Feature Selection). Найчастіше завдання Feature Selection будуються на інтуїтивних неформальних припущеннях [13], тому актуальним завданням є формалізація та оптимізація цього процесу.

Вибір найбільш інформативних ознак з набору наявних (Feature Selection) є процесом вибору відображення вигляду $x^R = f(x)$, в якому вхідний вектор-образ $x(k) = (x_1(k),..., x_n(k))^T$, ($k = 1,2,..., N$ – загальна кількість векторів-образів) належить простору $R^n$, а трансформований вектор належить простору $R^{n^R}$, причому $n^R < n$. Скорочений простір

ознак повинен складатися з найбільш інформативних показників вихідного простору $R^n$ [14]. Таким чином, головною метою такого перетворення є зменшення розмірності простору ознак так, щоб були збережені оптимальні характеристики даних, необхідні для здійснення подальшого процесу медичного діагностування.

Застосування гібридних систем для різних завдань при інтелектуальній обробці даних медико-біологічних досліджень (Medical Data Mining) має безперечний інтерес і підвищує якість діагностування, класифікації (кластеризації), розпізнавання образів, оскільки гібридизація дозволяє об'єднати переваги різних систем для досягнення поставленої мети. Актуальним завданням є застосування таких систем і для пошуку найбільш інформативних ознак в різних завданнях, особливо в задачах медичного діагностування, які виникають у зв'язку з дефіцитом (наприлад, нечіткість, незавершеність, наявність викидів та пропусків в даних) інформації.

**Мета дослідження, постановка задачі.** У роботі пропонується інтегрувати переваги систем, заснованих на Feature Extraction і Feature Selection та створити єдину гібридну систему Feature Extraction-Selection оцінки інформативності показників із виділенням найбільш інформативних без втрати фізичного сенсу (лінгвістичної інтерпретовності) скороченого простору ознак.

Таким чином, пропонується новий підхід, який полягає в тому, що на основі використання методу головних компонент (Principal Component Analysis) найбільш інформативними визначаються ті ознаки, що мають найменші відстані (у сенсі манхеттенської метрики) до першої головної компоненти вихідного набору даних.

**Матеріали досліджень.** На першому етапі реалізації підходу, що пропонується, всі виміряні у пацієнта ознаки формують матрицю ознак, представлену у вигляді таблиці «об'єкт-властивість». Якщо в ній є пропуски, то такі пацієнти повинні бути або видалені, або їх необхідно заповнити виходячи з припущень до заповнення пропусків, описаних в [15].

Далі вхідні дані повинні бути центровані щодо середнього $\bar{x}_i$ за допомогою співвідношення

$$x\_norm_i(k) = x_i(k) - \bar{x}_i \tag{1}$$

та закодовані в інтервал або $[0;1]^n$, або $[-1;1]^n$ згідно із формулами (2) і (3) відповідно

$$x\_kod_i(k) = \frac{x\_norm_i(k) - x_{i\min}}{x_{i\max} - x_{i\min}} \tag{2}$$

$$x\_kod_i(k) = \frac{2 \cdot x\_norm_i(k) - x_{i\max} - x_{i\min}}{x_{i\max} - x_{i\min}} \quad (3)$$

Далі обчислюється перший власний вектор і формується $(N \times 1)$ вектор перших головних компонент. При цьому перша головна компонента системи показників

$$x\_kod(k) = (x\_kod_1(k), ..., x\_kod_n(k))^T \in R^n$$

визначається як

$$\hat{y}^{(1)}(x\_kod) = l_1 \cdot x\_kod, \quad (4)$$

де $l_1$ – перший рядок матриці $L = \begin{pmatrix} l_{11} & \dots & l_{1n} \\ \dots & \dots & \dots \\ l_{n^R 1} & \dots & l_{n^R n} \end{pmatrix}$, власний вектор коваріаційної матриці ознак $\sum$, який відповідає найбільшому власному числу цієї матриці;

$l_{11}$ – проекція першої головної компоненти на вісь першої ознаки;

$l_{n^R n}$ – проекція $n^R$-ї головної компоненти на вісь $n$-ї ознаки.

Рядки матриці $L$ задовольняють умові ортогональності $LL' = L'L = I$,

$$\sum = (\sigma_{zj}), \quad (5)$$

$$\sigma_{zj} = \frac{\sum_{i=1}^{N} \left(x\_kod_i^{(z)} - \bar{x}^{(z)}\right)\left(x\_kod_i^{(j)} - \bar{x}^{(j)}\right)}{N}, z, j = 1, ..., n, \quad (6)$$

Наступним кроком є визначення вектору-ознак найближчого в сенсі манхеттенської метрики до першої головної компоненти (визначення ознаки-«переможця»), тобто відшукується вектор-ознак з мінімальною відстанню

$$d\left(x\_kod(z), \hat{y}^{(1)}\right) = \sum_{i=1}^{N} \left| x\_kod_i(z) - \hat{y}_i^{(1)} \right|. \quad (7)$$

Далі з вихідної матриці даних виключається ознака-«переможець» і робота системи триває на скороченій матриці ознак до тих пір, поки всі ознаки або будуть перебрані, або поки не виникне необхідність зупинити роботу системи.

На рисунку представлена гібридна система оцінювання інформативності медичних показників, що складається з блоку нормування та центрування вхідних ознак, на який надходить вхідний вектор

$$x(k)=(x_1(k),...,x_n(k))^T .$$

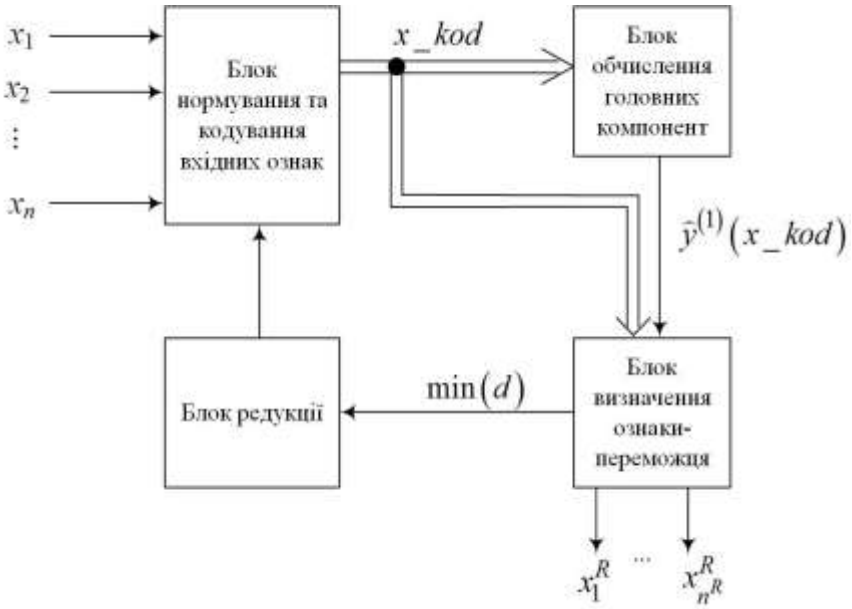


*Рис. 1. Гібридна система оцінювання інформативності медичних показників*

На виході цього блоку з'являється вектор закодованих ознак згідно із (1), (2), (3). Далі за допомогою метода головних компонент виділяється перша головна компонента $\hat{y}^{(1)}(x\_kod)$ згідно із (4-6) та у блоці визначення ознаки-«переможця» визначається та ознака, відстань до якої у сенсі манхеттенської метрики до першої головної компоненти є найменшою згідно із (7).

На заключному етапі у блоці редукції ця найбільш інформативна ознака вилучається з таблиці «об'єкт-властивість» і система починає пошук наступної на інформативністю ознаки.

В результаті роботи гібридної системи оцінки інформативності медичних показників з вихідного набору даних

$$x(k)=(x_1(k),...,x_n(k))^T \in R^n$$

буде сформований скорочений набір найбільш інформативних ознак

$$x^R(k)=(x_1^R(k),...,x_{n^R}^R(k))^T \in R^{n^R}, n^R < n.$$

**Результати досліджень.** Для перевірки роботи запропонованої гібридної системи оцінки інформативності медичних показників були використані медичні вибірки даних з репозиторію UCI Каліфорнійського університету, а саме dermatology.data (має 34 ознаки) [16], breastcancer.data (має 9 ознак) [17], pima-indian-diabetes.data [18], parkinsons.data [19]. В кожній із вибірок була проведена процедура пошуку найбільш інформативних ознак, результати представлені у таблиці.

*Таблиця*

*Перелік ознак за інформативністю у досліджуваних вибірках даних*

| Вибірка даних | Перелік ознак, починаючи з найбільш інформативних |
|---|---|
| dermatology.data | 21, 16, 33, 20, 29, 2, 19, 4, 28, 27, 22, 6, 9, 12, 10, 34, 5, 32, 17, 25, 3, 14, 26, 1, 18, 11, 8, 15, 24, 7, 23, 31, 13, 30 |
| breast-cancer.data | 6, 2, 3, 1, 7, 5, 8, 4, 9 |
| pima-indians-diabetes.data | 6, 3, 2, 1, 8, 4, 7, 5 |
| parkinsons.data | 19, 17, 20, 16, 14, 21, 11, 18, 9, 12, 10, 5, 4, 1, 2, 7 |

У вибірці dermatology.data найбільш інформативними виявились такі ознаки: подовження епідермальних гребнів, екзоцитоз, наявність смугового інфільтрату, потовщення епідермальних гребнів, поява пилковидних гребнів, поява псоріатичних лусочок та інші. У вибірці breastcancer.data: наявність немодифікованих ядр, однорідність розміру клітин, однорідність форми клітин, скупчення ущільненої маси та інші. У вибірці pima-indians-diabetes.data – це індекс маси тіла, діастолічний кров'яний тиск, концентрація глюкози в плазмі за 2 години та інші. У вибірці parkinsons.data – це сигнальний фрактальний показник масштабування, нелінійне динамічне вимірювання складності (RPDE), нелінійний вимір фундаментальної частотної варіації (spread1), виміри співвідношення шуму до тональних компонентів у голосі (HNR), вимір варіації в амплітуді голосового сигналу (Shimmer_DDA).

**Висновки.** Таким чином, у роботі запропонована гібридна система оцінювання інформативності медичних показників Feature Extrac-tion-Selection, що дозволяє виділяти найбільш інформативні показники без втрати фізичного сенсу скороченого простору ознак. Актуальним питанням залишається вибір кількості найінформативніших ознак. Для нього доцільним є порівняльний аналіз діагностування на повній вибірці даних та на скороченій із контролем якості діагностування.

## ЛІТЕРАТУРА

1. *Bezdek J.C. Prototype classification and feature selection with fuzzy sets / J.C. Bezdek, P. Castelaz // IEEE Trans. on Systems, Man and Cybernetics. – 1977. – № 7. – P. 87-92.*
2. *Pal S.K. Fuzzy set theoretic measures for automatic feature evaluation: II / S.K. Pal // Information Sciences. – 1992. – № 64. – P. 165-179.*
3. *Pal S.K. Fuzzy set theoretic measures for automatic feature evaluation / S.K. Pal, B. Chakraborty // IEEE Trans. on Systems, Man and Cybernetics. – 1986. – № 16. – P. 754-760.*
4. *Priddy K.L. Bayesian selection of important features or feed-forward neural networks / K.L. Priddy, S.K. Rogers, D.W. Ruck, G.L. Tarr, M. Kabrisky //Neurocomputing. – 1993. – № 5. – P. 91-103.*
5. *Steppe J.M. Improved feature screening in feedforward neural networks / J.M. Steppe, K.W. Bauer // Neurocomputing. – 1996. – № 13. – P. 47-58.*
6. *De R.K. Feature analysis: neural network and fuzzy set theoretic approaches / R.K. De, N.R. Pal, S.K Pal // Pattern Recognition. – 1997. – № 30. – P. 1579-1590.*
7. *Pregenzer M. Automated feature selection with a distinctive sensitive learning vector quantizer / M. Pregenzer, G. Pfurtscheller, D. Flotzinger // Neurocomputing. – 1996. – № 11. – P. 19-29.*
8. *Rao C.R. The use and interpretation of principal component analysis in applied research / C.R. Rao // Sankhya. – 1964. – Vol. 26. – № 4. – P. 329-358.*
9. *Okamoto M. Optimality principal components multivariate analysis / M. Okamoto // Proc. 3 Int. Symp. Dayton. – 1967.*
10. *Okamoto M. Minimization of eigenvalues of a matrix and optimality of principal components / M. Okamoto, M. Kanazawa // Ann. Math. Statist. – 1968. – Vol. 39. – № 3. – P. 1-20.*
11. *Bishop C.M. Neural Networks for Pattern Recognition / C.M. Bishop. − Oxford: Clarendon Press, 1995. – 482 p.*
12. *Cichocki A. Neural Networks for Optimization and Signal Processing / A. Cichocki, R. Unbehauen. – Stuttgart: Teubner, 1993. – 526 p.*
13. *Han J. Data Mining: Concepts and Techniques / J. Han, M. Kamber. – Amsterdam: Morgan Kaufman Publ., 2006. – 743 p.*
14. *Файнзильберг Л.С. Математические методы оценки полезности диагностических признаков: Монография / Л.С. Файнзильберг. – К.: Освіта України. 2010. – 152 с.*

15. *Mulesa P. Fuzzy Spacial Extrapolation Method Using Manhattan Metrics for Tasks of Medical Data Mining / P. Mulesa, I. Perova // Computer Science and Information Technologies CSIT'2015. – Lviv, 2015. – P. 104-106.*
16. *Dermatology dataset. – Available from: http://archive.ics.uci.edu/ml/machine-learning-databases/dermatology/dermatology.data. – Access Date: 2008.*
17. *Breast Cancer in Wisconsin dataset. – Available from: http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.data. – Access Date: 2008.*
18. *Pima Indians Diabetes dataset. – Available from: http://archive.ics.uci.edu/ml/machine-learning-databases/pima-indians-diabetes/pima-indians-diabetes.data. – Access Date: 2008.*
19. *Parkinson dataset. Available from: http://archive.ics.uci.edu/ml/machine-learning-databases/parkinsons/parkinsons.data. – Access Date: 2008.*

*Стаття надійшла до редакції 12.09.2017*