*Вісник Національного технічного університету "ХПІ", 2019, № 28 (1353)*
**ISSN 2079-0031 (Print) ISSN 2411-0558 (Online)**

**УДК 004.852** **DOI: 10.20998/2411-0558.2019.28.12**

***А. Ю. ІВАНИЦЬКА,*** канд. техн. наук, КНУ, Київ,
***Д. Є. ІВАНОВ,*** д-р. техн. наук, доц., ІПММ НАНУ, Слов'янськ,
***Л. В. ЗУБИК,*** канд. пед. наук, доц., КНУ, Київ

# РОЗРОБКА МЕТОДУ АНАЛІЗУ СКЛАДНИХ ДАНИХ НА ОСНОВІ ТЕХНОЛОГІЙ MACHINE LEARNING

Проведено аналіз сучасних підходів та методів інтелектуальної обробки даних, що застосовуються при аналізі складних даних. Запропоновано новий метод аналізу складних даних, який є комбінацією трьох відомих підходів machine learning: прогнозування, кластеризація та класифікація даних. Цей метод дозволяє виявляти нові характеристики та цінні знання за рахунок підвищення точності прогнозування часових рядів даних. Іл.: 3. Табл.: 1. Бібліогр.: 10 назв.

**Ключові слова**: аналіз; складні дані; кластеризація; machine learning; класифікація

**Постановка проблеми.** За останні роки стрімко розвиваються методи аналізу складних даних, що пов'язано із ускладненням структури таких даних та зі збільшенням обсягу інформації, що оброблюється. Складні дані представляють собою великі і неструктуровані дані, недосконалу інформацію, що зберігаються у різних форматах. Для виявлення цінних даних, нових закономірностей та їх інтерпретації, а також прогнозування результатів обробки і формування практичних рішень у бізнесі необхідно використовувати методи кластеризації та технології машинного навчання.

У зв'язку з цим, в роботі проводиться дослідження особливостей класифікації методів і технологій machine learning при обробці часових рядів даних. На теперішній час дослідниками запропоновано цілу низку методів обробки великих даних. Наприклад, у [1] запропоновано сучасні технології business intelligence на основі кластеризації даних для діагностики роботи підприємства. Також існують методи опрацювання та технології машинного навчання, що використовуються для аналізу і прогнозування великих даних (big data) зі супутників та сховищ даних [2, 3]. Вони дають практично прийнятні показники в термінах точності та швидкодії для конкретного завдання, що досліджується авторами. Проте, аспект виявлення цінних даних, алгоритмічна реалізація й експериментальне впровадження результатів висвітлюються авторами недостатньо. Таким чином, дослідження методів обробки складних даних з метою прогнозування та виявлення цінних сутностей є актуальним завданням. В даній роботі запропоновано новий вдосконалений метод

аналізу складних даних, який є комбінацією трьох відомих підходів machine learning: прогнозування, кластеризація та класифікація даних. А також розглянуто використання прикладних засобів при проведенні апробації методу.

**Аналіз літератури.** Сучасні інтелектуальні технології дозволяють реалізовувати моделі і методи різної складності для класифікації, прогнозування та кластеризації даних, а також розробки інформаційних технологій їх обробки. Наприклад, для кластеризації даних із супутника використано алгоритми максимізації k-means та expectation maximization (EM), що сформували розподілені кластерні рішення, а для обробки big data обрано хмарні технології (InterCloud) [2]. Також в [3] розроблено модель великих даних "сутність-характеристика", яка дає змогу створити метод інтеграції даних, що пов’язує дані з джерел з наперед невідомою структурою даних та дозволяє підвищити ефективність подальшого аналізу Великих даних. В [4] проведено аналіз технології big data та супутних методів класифікації. Для бінарної класифікації табличних даних використовується логічна регресійна модель, що дозволяє визначити кількість корисних даних та спаму для навчальної та перевірочної вибірок. Після проведення експерименту отримані наступні значення для кількісних показників прогнозу: SE = 0.89; SP = 0.86; AC = = 0.93 [5].

Основна архітектура нейромережі, що використовується на основі традиційного повністю зв'язаного багатошарового персептрону (MLP) та найбільш часто використовуваний підхід у спільноті RS на основі випадкових лісів (random forest), що порівнюються із згортковими NN (CNN) для класифікації сільськогосподарських культур та картографування земель зі супутникових знімків Sentinel 2 на основі згорткової нейронної мережі (CNN) [6 – 8].

У дослідженні [9] показано, що використання методів машинного навчання для мінімізації повернень товару в системах електронної комерції є ефективним та дає змогу зменшити витрати на обслуговування клієнтів. В [10] були описані та класифіковані пласкі та ієрархічні методи кластеризації.

В результаті аналізу існуючих підходів виявлено, що переважно математичний апарат використовується для конкретного завдання, яке вирішується авторами. Однак аспект виявлення цінних даних, алгоритмічна реалізація і експериментальне впровадження результатів висвітлюється авторами недостатньо. Тому запропоновано новий метод аналізу складних даних, який є комбінацією трьох відомих підходів machine learning: прогнозування, кластеризація та класифікація даних.

Розробка і апробація нового методу виконуються на основі реальних даних про біржову вартість акцій компанії Amazone.

**Мета дослідження** – провести аналіз методів обробки складних даних, що використовують технологію machine learning, на основі якого розробити новий метод аналізу даних з метою виявлення і формування цінних даних.

Для досягнення зазначеної мети слід вирішити наступні завдання:

– провести аналіз методів обробки складних даних на основі технологій machine learning;

– виділити етапи нового методу обробки даних;

– дослідити застосування розробленого методу на тестових даних.

**Матеріали дослідження.** Спочатку виконується аналіз методів обробки складних даних, що представлений вище, на основі огляду сучасних підходів до класифікації даних. На наступному кроці проведено аналіз методів кластеризації даних, а саме: k-середніх, expectation-maximization (EM) кластеризація, спектральна кластеризація (spectral), агломеративна кластеризація (agglom).

Проведена порівняльна характеристика на основі існуючих підходів до оцінки якості обробки методів кластеризації даних: однорідність (homogenity), функціональна повнота (completeness), V-міра, Adjusted Rand Index (ARI) оцінює, наскільки багато з тих пар елементів, які перебували в одному класі, і тих пар елементів, які перебували в різних класах, зберегли цей стан після кластеризації алгоритмом. Adjusted Mutual Information (AMI) – базове значення взаємної інформації між двома випадковими кластерами. Силует (Silhouette statistic) показує, наскільки об'єкт схожий на свій кластер у порівнянні з іншими кластерами (чим ближче отримана дана оцінка до 1, тим краще).

Для проведення аналізу методів кластеризації, що наведено вище, використано дані рукописних цифр MNIST. В результаті дослідження отримані значення оцінок якості методів кластеризації на основі технологій machine learning, які наведено в табл. 1.

Таблиця 1

Порівняльна характеристика методів кластеризації на основі якісних оцінок

| | Однорідність | Повнота | V-міра | Силует | ARI | AMI |
|---|---|---|---|---|---|---|
| k-середніх | 0.7354 | 0.743 | 0.7392 | 0.1821 | 0.6623 | 0.7328 |
| EM-алгор. | 0.959 | 0.4869 | 0.6459 | 0.1152 | 0.1752 | 0.4512 |
| Spectral | 0.8295 | 0.8764 | 0.8523 | 0.1822 | 0.7526 | 0.8278 |
| Agglom. | 0.8575 | 0.8791 | 0.8682 | 0.1785 | 0.794 | 0.8561 |

**ISSN 2079-0031 (Print) ISSN 2411-0558 (Online)**

Відповідно до результатів з табл. 1 можна сказати, що метод k-середніх показує вищі результати для показників силует та повнота (серед своїх значень). У свою чергу, метод EM кластеризації дає найвищі результати для всіх оцінок якості кластеризації. Тому, саме дані методи буде використано при розробці нового вдосконаленого методу аналізу складних даних.

**Розробка методу аналізу даних**

Після порівняльного аналізу методів кластеризації даних сформулюємо етапи (кроки) пропонованого вдосконаленого підходу обробки складних даних на основі технологій машинного навчання:

1. Завантажити файл даних в середовище програмування R.
2. Використати ARIMA модель для аналізу та прогнозування даних.
3. Виконати кластеризацію даних на основі методів k-середніх і EM кластеризації.
4. Проаналізувати отримані результати обробки даних.
5. Провести навчання та класифікацію даних на основі нейронної мережі типу персептрон.
6. Виконати оцінку ефективності розробленого методу аналізу складних даних на конкретних прикладах.

Об'єктом даного дослідження є дані по акціях компанії Amazone на біржі (тікер) за період з 01 травня 2019 по 30 вересня 2019 року.

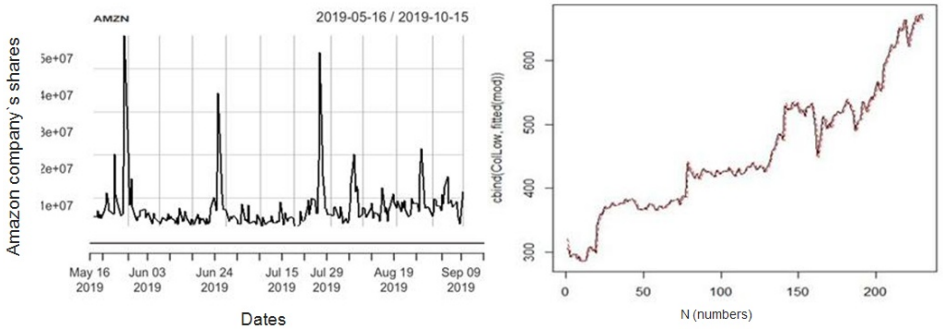
Для аналізу даних використовується ARIMA модель – інтегрована модель авторегресії для аналізу часових рядів. Вихідні дані по акціях компанії представлені на графіку (рис. 1, *а*).

ARIMA модель, що побудована на реальних даних, дозволяє отримати графік модельованих і реальних значень часового ряду (рис 1, *б*), для яких перераховано стандартні помилки коефіцієнтів та виконано оцінку дисперсії. За формулами (1) отримано наступні дані: значення критерію Акаіке AIC = 542.38, скоригованого критерію Акаіке AICc = 542.42 і критерію Шварца BIC = 545.04.

$$yt = 186.92 + 0.21yt - 1 + 0.66yt - 2 + \varepsilon t + 0.72\varepsilon t - 1,$$
$$BIC = -2 \cdot lnL + lnn \cdot \mathrm{k}. \quad (1)$$

Далі виконується прогнозування на 20 часових кроків уперед. Отримано значення по акціях компанії Amazone на біржі, що включають 80% і 95% довірчі інтервали для прогнозу, які представлені на рис. 2.

**ISSN 2079-0031 (Print) ISSN 2411-0558 (Online)**




*а*) *б*)

Рис. 1. *а*: Вихідні дані по акціях компанії; *б*: Графік співвідношення модельованих та реальних значень часового ряду

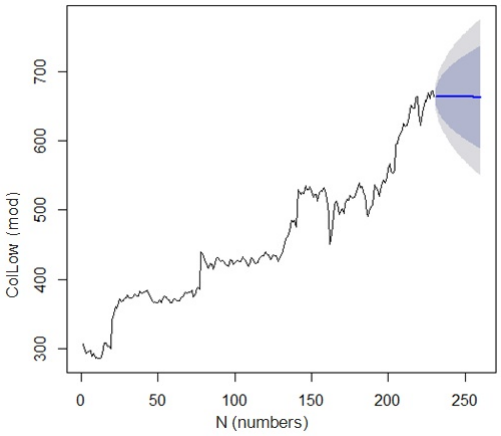



Рис. 2. Графік результату, що включає 80% та 95% довірчі інтервали для прогнозу даних

Далі використовується метод k-середніх, який виконує кластеризацію даних на 5 кластерів і впорядковує множину об'єктів у порівняно однорідні групи. Відомо, що в основі методу лежить мінімізація суми відхилення відстані між центром кластера і поточним показником, тобто зменшенням середньої похибки.

Після виконання всіх кроків алгоритму метод k-середніх прагне мінімізувати дисперсію даних окремого кластера від центру цього кластера:

$$V(x,\mu)=\sum_{i=1}^{k}\sum_{x_j\in S_i}\left(x_j-\mu_i\right)^2, \qquad (2)$$

де $k$ – кількість кластерів, $S_i$ – отримані кластери, $x_j$ – центр мас векторів,

**ISSN 2079-0031 (Print) ISSN 2411-0558 (Online)**

$\mu_i$ – середнє значення з множини $S_i$.

ЕМ кластеризація реалізує нескладну евристичну процедуру визначення класів і теж використовує Евклідовому відстань. На відміну від порогового алгоритму, виявляються найбільш віддалені кластери. "Типова" відстань визначається як середнє арифметичне всіх відстаней між знайденими центрами (три центри). На підставі описаної кластеризації над вхідними даними акцій компанії було обрано 3 кластери.

Після отримання результатів кластеризації даних методом k-середніх і ЕМ кластеризації виконано аналіз та обрано метод k-середніх для розробленого методу через високу швидкодію та відносну простоту кластеризації, а також через більше значення показника повноти інформації.

Наступним етапом розробленого методу аналізу даних акцій компанії Amazone на біржі є класифікація на основі персептроного методу. Для аналізу вибираємо з таблиці дані за такими стовпцями: AMZ.Open і AMZ.Close, які необхідно класифікувати відповідно вартості акцій стосовно днів тижня. Алгоритм навчання персептрона зводиться до ітераційного визначення вектора вагових коефіцієнтів за принципом "заохочення – покарання". У алгоритмі корекції абсолютна величина приросту обирається досить великою, щоб гарантувати правильну класифікацію кластера даних після корекції ваг. Іншими словами, якщо $\vec{w}^T(k)\vec{x}(k) \le 0$, то коефіцієнт виходить таким чином:

$$\vec{w}^T(k+1)\vec{x}(k) = \left[\vec{w}(k) + c\vec{x}(k)\right]^T \vec{x}(k) > 0. \quad (3)$$

Приріст $c$ повинен вибиратися з умови:

$$c = \left\lceil \left| \frac{\vec{w}^T(k)\vec{x}(k)}{x^T(k)\vec{x}(k)} \right| \right\rceil. \quad (4)$$

Після класифікації отримуємо розбиття на дані зі збільшенням акцій і з падінням мінімального значення помилки (*Err*) і мінімальною кількістю епох. Це свідчить про точність і швидкость запропонованого способу.

Результати класифікації і тестування на основі нейронної мережі типу "персептрон" показують, що точність АС = 94,6% підвищена у порівнянні з існуючими підходами приблизно на 2%. При цьому швидкість виконання алгоритму збільшилася в 10 разів. Принципово швидкість роботи методу вдалося збільшити за рахунок застосування нейронної мережі на етапі класифікації.

**Висновки.** Розглянуто сучасні підходи до інтелектуального аналізу даних, які застосовуються при обробці складних даних. Запропоновано новий вдосконалений метод прогнозування результатів обробки, що включає наступні етапи: 1) прогнозування даних на основі моделі ARIMA, 2) кластеризація на основі методу k-середніх, 3) класифікація на основі нейронної мережі типу персептрон. Побудова і апробація нового методу виконані на основі реальних даних про вартість акцій компанії Amazone на біржі. Отримані числові результати експериментів показують, що в порівнянні з існуючими підходами вдалося підвищити як точність прогнозу (на 2%), так і швидкість роботи методу. У подальших дослідженнях пропонується перевірити працездатність розробленого методу при обробці недосконалих даних.

**Список літератури:**

**1.** *Титова А.Ю.* Анализ технологий business intelligence при обработке диагностической информации / *А.Ю.* Титова // Материалы Регионального семинара Международного союза электросвязи для стран Европы и СНГ «Цифровое будущее на основе 4G/5G», г. Киев, 14-16 мая 2018. – 2018. – С. 84-85.

**2.** *Ayma V.* Mapping glacier changes using clustering techniques on cloud computing infrastruvture. ISPRS / *V. Ayma, B. Castañón , C. Happ, P Raul* // International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. – 2019. doi: XLII-2/W16.29-34.10.5194/isprs-archives-XLII-2-W16-29-2019.

**3.** *Болюбаш Ю.Я.* Методи опрацювання Великих даних у федеративному сховищі даних / *Ю.Я. Болюбаш* // Вісник Національного університету "Львівська політехніка". – 2016. – № 843: Комп'ютерні науки та інформаційні технології. – С. 356-365.

**4.** *Manal A.* Big data mining: A classification perspective / *A. Manal.* – 2016. doi: 10.1201/9781315375083-97.

**5.** Розробка моделі аналізу складних даних на основі класифікації machine learning / *А.Ю. Тітова, Д.Є. Іванов* // Вісник НТУ "ХПІ". Серія: Інформатика та моделювання. – Харків: НТУ "ХПІ". – 2018. – № 42. – С. 171–178.

**6.** *Kussul N.* Deep Learning Classification of Land Cover and Crop Types Using Remote Sensing Data / *N. Kussul, M. Lavreniuk, S. Skakun, A. Shelestov* // IEEE Geoscience and Remote Sensing Letters, vol. 14, no. 5, pp. 778-782, May 2017. doi: 10.1109/LGRS.2017.2681128

**7.** *Nijhawan R.* A Futuristic Deep Learning Framework Approach for Land Use-Land Cover Classification Using Remote Sensing Imagery / *R. Nijhawan, D. Joshi, N. Narang, A. Mittal* // In: Mandal J., Bhattacharyya D., Auluck N. (eds) Advanced Computing and Communication Technologies. Advances in Intelligent Systems and Computing. – Springer, Singapore, – 2019. – Vol. 702.

**8.** *Яременко В.С.* Огляд наявних мультиагентних систем для задач інтелектуального аналізу даних / *В.С. Яременко* // Вчені записки Таврійського національного університету імені В. І. Вернадського. Серія: технічні науки: Інформатика, обчислювальна техніка та автоматизація. – 2018. – Том 29 (68). – Ч. 2. – № 3.

**9.** *Нарушинська О.* Застосування методів машинного навчання для мінімізації повернень товару в системах електронної комерції / *О.О. Нарушинська, В.М. Теслюк, П.Ю. Денисюк* // Актуал. пробл. екон. : наук. економ. журнал. – 2017. – № 3. – С. 342-347.

**10.** *Якимець Р.В.* Методи кластеризації та їх класифікація / *Р.В. Якимець* // Міжнародний науковий журнал. – 2016. – № 6 (2). – С. 48-50. – Режим доступу: http://nbuv.gov.ua/UJRN/mnj_2016_6%282%29__12.

**References:**

**1.** Titova, A.Yu. (2018), "Analysis of business intelligence technologies in the processing of diagnostic information". *Materials of the Regional Seminar of the International Telecommunication Union for Europe and the CIS "Digital Future Based on 4G / 5g"*, Kiev, 2018, pp. 84-85.
**2.** Ayma, V., Castañón, B., Happ, C., Raul, P. (2019), "Mapping glacier changes using clustering techniques on cloud computing infrastruvture". *ISPRS. International Archives of the Photogrammetry, Remote Sensing and Spatial Information.* doi: XLII-2/W16.29-34.10.5194/isprs-archives-XLII-2-W16-29-2019.
**3.** Bolyubash, Yu.Ya. (2016), "Methodi oprazyuvannya Great tributes to the federal monastic tribute". *News of the National University of Lviv Polytechnic*, No. 843: Computer Science and Information Technology, pp. 356-365.
**4.** Manal A. (2016) *Big data mining: A classification perspective*. doi: 10.1201/9781315375083-97.
**5.** Titova, A.Yu., Ivanov, D.E. (2018), "Development of a model of complex data analysis based on the classification of machine learning". *Bulletin of NTU "KPI"*. Series: Informatics and Modeling. - Kharkiv: NTU "KPI", 2018, No. 42, pp. 171-178.
**6.** Kussul, N., Lavreniuk, M., Skakun, S., Shelestov, A. (2017), "Deep Learning Classification of Land Cover and Crop Types Using Remote Sensing Data". *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 5, pp. 778-782, May 2017. doi: 10.1109/LGRS.2017.2681128.
**7.** Nijhawan, R., Joshi, D., Narang, N., Mittal, A. (2019), "A Futuristic Deep Learning Framework Approach for Land Use-Land Cover Classification Using Remote Sensing Imagery". *In: Mandal J., Bhattacharyya D., Auluck N. (eds) Advanced Computing and Communication Technologies. Advances in Intelligent Systems and Computing*, 2019, Vol. 702, Springer, Singapore.
**8.** Yaremenko, V.S. (2018), "Review of existing multiagent systems for data mining problems". *Series: Engineering Sciences: Computer Science, Computer Engineering and Automation*, Volume 29 (68), Part 2, No. 3, 2018.
**9.** Narushinskaya, O., Teslyuk, V., Denysyuk, P. (2017), "Application of Machine Learning Methods for Minimizing Product Returns in Electronic Systems of commerce". *Actual. prob. econom. : Sciences. economy. magazine*, 2017, No. 3, pp. 342-347.
**10.** Yakimets, R.V. (2016), "Methods of clustering and their classification". *International scientific journal*, 2016, No. 6 (2), pp. 48-50. - Access mode: http://nbuv.gov.ua/ UJRN/mnj_2016_6%282%29__12.

*Статтю представив д-р техн. наук, проф. КНУ Сайко В.Г.*

*Надійшла (received) 08.12.2019*

Ivanytska Anastasiia, PhD Tech
Taras Shevchenko National University of Kyiv
Str. Bohdan Hawrylyshyn, 24, Kyiv, 04116
Tel.: (095) 333-51-01, e-mail: a.titova.wk@gmail.com
ORCID ID: 0000-0002-4803-2090

Ivanov Dmitriy, Dr.Sci.Tech, Ass. Professor,
Institute of Applied Mathematics and Mechanics,
Str. Gen. Batyuka, 19, Slavyansk, 84100
Tel: (063) 559-51-90, e-mail: dmitry.ivanov.iamm@gmail.com
ORCID ID: 0000-0001-9956-6589

Zubyk Liudmyla, PhD Ped, Ass. Professor,
Taras Shevchenko National University of Kyiv
Str. Bohdan Hawrylyshyn, 24, Kyiv, 04116
Tel.: (066) 60-42-556, e-mail: l.v.zubyk@univ.kiev.ua
ORCID ID: 0000-0002-2087-5379

УДК 004.852

**Розробка методу аналізу складних даних на основі технологій machine learning / Іваницька А.Ю., Іванов Д.Є., Зубик Л.В.** // Вісник НТУ "ХПІ". Серія: Інформатика та моделювання. – Харків: НТУ "ХПІ". – 2019. – № 28 (1353). – С. 131 – 140.

Проведено аналіз сучасних підходів та методів інтелектуальної обробки даних, що застосовуються при аналізі складних даних. Запропоновано новий метод аналізу складних даних, який є комбінацією трьох відомих підходів machine learning: прогнозування, кластеризація, класифікація даних. Цей метод дозволяє виявляти нові характеристики та цінні знання за рахунок підвищення точності прогнозування часових рядів даних. Іл.: 3. Табл.: 1. Бібліогр.: 10 назв.

**Ключові слова**: аналіз; складні дані; кластеризація; machine learning; класифікація.

УДК 004.852

**Разработка метода анализа сложных данных на основе технологий machine learning / Иваницкая А.Ю., Иванов Д.Е., Зубик Л.В.** // Вестник НТУ "ХПИ". Серия: Информатика и моделирование. – Харьков: НТУ "ХПИ". – 2019. – № 28 (1353). – С. 131 – 140.

Выполнено анализ современных подходов и методов интеллектуальной обработки данных, которые применяются при анализе сложных данных. Предложен новый метод анализа сложных данных, который является комбинацией трех известных подходов machine learning: прогнозирование, кластеризация, классификация данных. Этот метод позволяет выявлять новые характеристики и ценные знания за счет повышения точности прогнозирования временных рядов данных. Ил.: 3. Табл.: 1. Библиогр.: 10 назв.

**Ключевые слова:** анализ; сложные данные; кластеризация; machine learning; классификация.

УДК 004.852

**Development of the method of complex data analysis based on machine learning techniques / Ivanytska A.Yu., Ivanov D.E., Zubyk L.V.** // Herald of the National Technical University "KhPI". Series of "Informatics and Modeling". – Kharkov: NTU "KhPI". – 2019. – № 28 (1353). – P. 131 – 140.

The analysis of modern approaches and methods that used in intellectual processing of complex data are considered. A new method for complex data analysis is proposed that combines three well-known machine learning approaches: forecasting, clustering and classification of data. This method allows to identify new characteristics and valuable knowledge by improving the accuracy of forecasting time series data. Fig.: 3. Tabl.: 1. Refs.: 10 titles.

**Keywords**: analysis; complex data; clustering; machine learning; classification.