УДК 004.032.26 DOI: 10.31673/2412-4338.2019.013136

Тушич А.М., Сторчак К.П., Бондарчук А.П., Макаренко А.О.
Державний університет телекомунікацій, Київ

ВИМОГИ ДО ІНТЕЛЕКТУАЛЬНИХ СИСТЕМ АНАЛІЗУ ДАНИХ ТА ЇХ КЛАСИФІКАЦІЙ

У статті розглянуто основні існуючі системи аналізу даних та їх класифікації та проаналізовано вимоги, що до них ставляться. На основі аналізу приходимо до висновку, що найповніше справляється із поставленою задачею система на основі нейронних мереж. Запропонований підхід задовольняє більшість вимог до систем, а саме обробка великих об'ємів даних, які до того ж можуть бути зашумленими, а також містить єдиний математичний апарат, що не потребує спеціальних знань користувача.

Ключові слова: *інтелектуальний аналіз даних, система інтелектуального аналізу даних, нейронна мережа.*

Tushych A.M., Storchack K.P., Bondarchuk A.P., Makarenko A.O.
State University of Telecommunications, Kyiv

REQUIREMENTS FOR INTELLIGENT DATA ANALYSIS SYSTEMS AND THEIR CLASSIFICATIONS

The article deals with the main existing diverse data analysis systems and their classification, which are used as a mass product for business applications or for unique research. Today, the application of not all systems is optimal, especially when it comes to the speed of data processing. The work analyzes the requirements relating to the systems of data mining, namely, support for work with data of large volumes, support for work with data that are heterogeneous in quality composition, providing work with noisy data, ensuring the availability of one mathematical algorithm for solving problems. the tasks that belong to various problem areas and to ensure the ease of work with a program of specialists without additional mathematical knowledge. Intelligent processing of data was formed at the junction of areas such as applied statistics, pattern recognition, database theory, artificial intelligence, and others, which explains a large number of methods and algorithms implemented in various existing systems of data mining, in addition some of them integrate several approaches at a time. On the basis of the analysis, we arrive at the conclusion that the system with the given problem is the most complete with the help of neural networks. Such a system can work with noisy data, data of a large volume. In addition, the system based on neural networks can be applicable to a sufficiently wide range of tasks and does not require the user to have special knowledge, as the network setup process is replaced by the learning phase. Also, the system contains a single mathematical device that does not require special knowledge of the user. Such systems have one significant minus - the results are often not easily interpreted.

Key words: *intelligent data analysis, data mining system, neural network.*

Тушич А.Н., Сторчак К.П., Бондарчук А.П., Макаренко А.А.
Государственный университет телекоммуникаций, Киев

ТРЕБОВАНИЯ К ИНТЕЛЛЕКТУАЛЬНЫМ СИСТЕМАМ АНАЛИЗА ДАННЫХ И ИХ КЛАССИФИКАЦИЯМ

В статье рассмотрены основные существующие системы анализа данных и их классификации, проанализированы требования, которые к ним относятся. На основе анализа

приходим к выводу, что наиболее полно справляется с поставленной задачей система на основе нейронных сетей. Предложенный подход удовлетворяет большинство требований к системам, а именно обработка больших объемов данных, которые к тому же могут быть зашумленными, а также могут содержать единственный математический аппарат, который не требует специальных знаний пользователя.

Ключевые слова: *интеллектуальный анализ данных, система интеллектуального анализа данных, нейронная сеть.*

Вступ

Масштабний потік даних, спричинений техногенним розвитком усіх сфер від виробництва до повсякденної діяльності людини спричинив виникнення проблеми аналізу цього потоку, виявлення певних закономірностей у ньому тощо з метою отримання знань, що будуть корисними для прийняття майбутніх рішень. Без продуктивної обробки ці дані представляють собою сховища нікому не потрібної інформації.

Наразі існують різноманітні системи аналізу даних, що застосовуються як масовий продукт для бізнес-додатків або для проведення унікальних досліджень. Але під час їх використання стикаються із рядом проблем, основною з яких є те, що їх застосування не є оптимальним, коли мова йде про швидкість опрацювання даних та роботу із зашумленими даними.

Виходом є застосування штучних нейронних мереж задля створення системи інтелектуального аналізу даних.

У цій статті буде проаналізовано основні вимоги до інтелектуальних систем аналізу даних на основі нейронних мереж. На основі аналізу буде обрано оптимальну систему.

Вимоги до систем

Сьогодення вимагає від систем інтелектуального аналізу даних виконання наступних вимог, а саме:

- підтримка роботи з даними великих об’ємів;
- підтримка роботи з даними, що є різнорідними за якісним складом;
- забезпечення роботи із зашумленими даними;
- забезпечити наявність одного математичного алгоритму для розв’язування задач, що належать до різних проблемних областей;
- простота роботи програми – забезпечення простоти архітектури програми для роботи з нею фахівців без додаткових математичних знань.

Методи аналізу даних

Інтелектуальна обробка даних утворилась на стику таких областей як прикладна статистика, розпізнавання образів, теорія баз даних та штучний інтелект та інші (рис.1), що пояснює значну кількість методів та алгоритмів, що реалізовані в різноманітних діючих системах інтелектуального аналізу даних, крім того деякі з них інтегрують у собі кілька підходів одночасно.

Системи інтелектуального аналізу даних можуть бути засновані на одному із методів або поєднанням кількох. Розглянемо відомі системи та дамо їм коротку характеристику:

- предметно-орієнтовані аналітичні системи: такі системи є дуже різноманітними, проте вибір конкретного типу залежить від класу задачі, що розглядається; прості у розумінні та інтерфейсі, оскільки оперують термінами предметної області, що є зрозумілою для потенційного користувача (MetaStock, Super Charts, Candlestick Forecaster тощо);

- статистичні пакети: засновані на класичних методиках кореляційного, регресійного, факторного аналізу тощо, що потребує наявності достатніх для користування статистичних знань, тобто попереднього курсу підготовки; засоби автоматизації процесу відсутні та потребують додаткового програмування у випадку необхідності (SPSS Statistics, SAS тощо);

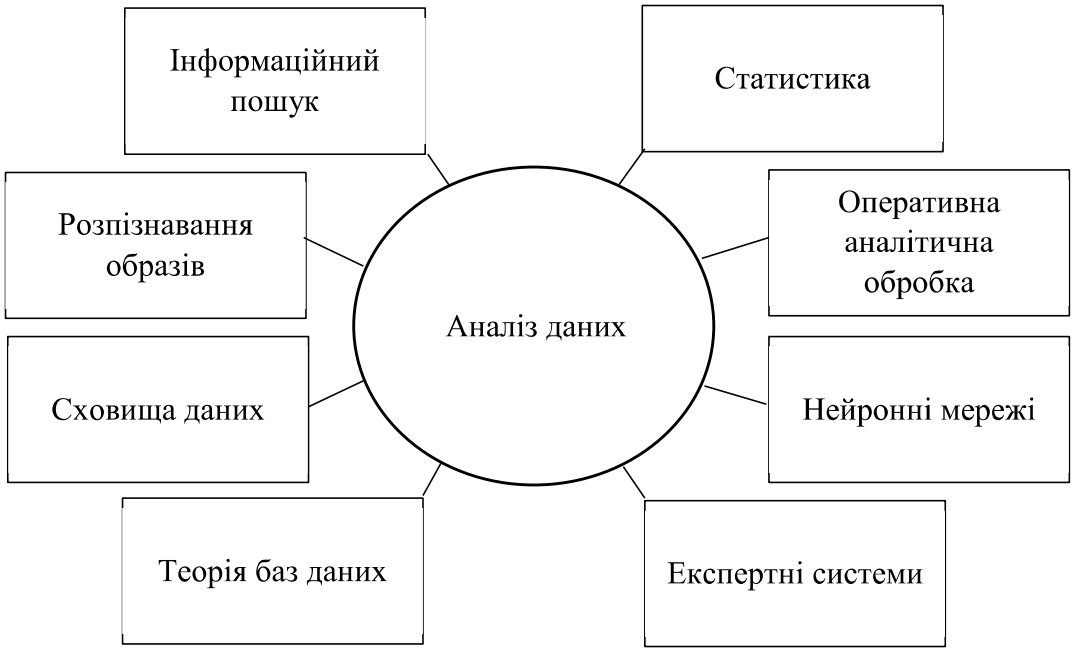

Рис. 1. Методи інтелектуального аналізу даних

- нейронні мережі: система, що імітує роботу головного мозку людини, здатна до навчання, за допомогою чого така мережа піддається коректуванню свого алгоритму у процесі роботи; для коректної роботи потрібно мати велику вибірку даних для тренування або навчання, що достатньою мірою характеризує систему, що буде вивчатись; система стійка до шуму; недоліком є те, що отримані результати зазвичай важко інтерпретувати (BrainMaker, Hiperlogic, NeuroShell тощо);
- дерева рішень: метод, що використовують для задач класифікації; аналіз проводиться за алгоритмом, що має структуру дерева з використанням правила «якщо» – «то»; недоліками є обмеженість сфер використання та можливість отримання статистично не обґрунтованого результату під час великої кількості умов (PS CLEMENTINE PRO, IDIS, See5/C5.0 тощо);
- еволюційне програмування: гіпотези цільової змінної як залежність від інших змінних формуються у вигляді програм; коли система знаходить програму, яка достатньо точно виражає залежність, яку необхідно знайти (хоча часом це зробити дуже складно – в чому і полягає недолік системи), тоді починає вносити в її структуру певні зміни та обирає серед утворених програм ті, що підвищують точність, вибудовуючи при цьому кілька генетичних програмних ліній, що конкурують за точністю залежності; інтерпретація результатів здійснюється у зрозумілому для користувача вигляді (Poly Analyst тощо);
- алгоритми обмеженого перебору: здійснення пошуку логічних закономірностей в даних за рахунок розрахунку частоти комбінацій простих логічних подій, що містяться

у підгрупах даних; недоліками є необхідність пошуку усіх комбінацій «якщо» – «то», а також залежність часу від розміру даних, що обробляються (WizWhy тощо);

- метод найближчого сусіда: використовується для задач класифікації об'єктів, при цьому об'єкт присвоюється тому класу, що є найбільш розповсюдженим серед сусідів даного об'єкта; недоліком є те, що такі системи є чутливими до зашумлених даних, а також помітні труднощі при роботі із великими об'ємами даних (Apache Commons, Apache Spark тощо).

Порівняльний аналіз можливостей систем

До систем інтелектуального аналізу даних поставлено ряд вимог задля забезпечення виконання ними поставленої задачі. У таблиці 1 наведено результати порівняльного аналізу можливостей систем.

Таблиця 1. Аналіз можливостей аналітичних систем.

Аналітична система	Дані великих об'ємів	Зашумлені дані	Єдиний математичний апарат	Знання математичного апарату	Ясність результату
Статистичні пакети	+	+	+	–	–
Предметно-орієнтовані системи	+	–	–	+	+
Нейромережеві пакети	+	+	+	+	–
Системи на основі методу дерева рішень	+	–	–	+	+
Системи на основі методу найближчого сусіда	–	–	+	+	–
Системи обмеженого перебору	–	–	–	+	+
Системи на основі методів еволюційного програмування	+	+	+	–	–

Результат аналізу

Згідно наведеного аналізу систем приходимо до висновку, що найбільш повно вимоги задовольняє система на основі нейронної мережі. Переваги використання такої мережі є наступними:

- можливість розв'язання однією мережею з порівняно невеликою кількістю нейронів одночасно декількох задач класифікації та прогнозу;
- можливість навчатись, а також «перенавчатись» під час отримання даних із іншої предметної сфери та «довчатись» під час надходження нових даних без втручання у програмний код;

- можливість використання будь-якої кількості незалежних та залежних ознак, кількість прикладів для різних класів у задачі класифікації;
- нейромережеві алгоритми локальні, а нейрони здатні функціонувати паралельно, що забезпечує високоефективну паралельно-послідовну обробку інформації, необхідну для ефективної обробки образів; кожен нейрон реагує лише на локальну інформацію, що надходить до нього від зв'язаних з ним аналогічних нейронів;
- здатність до навчання;
- здатність до узагальнення;
- висока стійкість та надійність до відмов у окремих елементах, що формують нейронну мережу;
- обчислення проводяться локально у нейронах, які змінюють свої адаптивні параметри у відповідності з інформацією про ефективність роботи всієї мережі загалом;
- місце програмування займає процес навчання з принципом мінімізації емпіричної помилки.

Висновки

У роботі представлено вимоги до сучасних систем інтелектуального аналізу даних. Показано переваги та недоліки систем. На основі аналізу приходимо до висновку, що система інтелектуального аналізу даних на основі нейронних мереж найповніше задовольняє поставлені вимоги до систем такого типу, оскільки така система може працювати із зашумленими даними та даними великого об'єму. До того ж система на основі нейронних мереж може бути застосовною для достатньо широкого кола задач та не потребує від користувача наявності особливих знань, адже процес налаштування мережі заміняється етапом навчання. Такі системи мають один суттєвий мінус – результати дуже часто не легко інтерпретуються. Тому задача побудови інтелектуальної системи на основі нейронних мереж потребує подальшого удосконалення.

Список використаної літератури

1. Сторчак К.П. Інтелектуальний аналіз даних з використанням нейронних мереж / Сторчак К.П., А.М. Тушич, К.С. Козелкова, М.М. Степанов // Зв'язок. – 2018. – № 4. – С. 17-19.

2. Тушич А.М. Аналіз доцільності використання автоматизованої системи інтелектуального аналізу даних на основі штучних нейронних мереж / А.М.Тушич // VII Всеукраїнська науково-практична конференція студентів, аспірантів та молодих вчених з автоматичного управління. – Херсон. – 10-12 квітня 2019 р. – С. 76-77.

3. Кулаков П.А. Основные классы нейронных сетей в задаче диагностики технологического оборудования / П.А. Кулаков // Электронный научно-практический журнал «Молодежный научный вестник». – 2016. – № 10. – С. 74-77.

4. Дюк В.А. Data Mining – интеллектуальный анализ данных // Информационные технологии: сайт. – URL: http://www.inftech.webservis.ru/it/database/datamining/ar2.html (дата звернення 03.12.2018)

5. Назаров А.В. Нейросетевые алгоритмы прогнозирования и оптимизации систем / А.В. Назаров, А.И. Лоскутов. – СПб.: Наука и Техника, 2015. – 384 с.

6. Лєнков С.В. Концептуальна схема системи інтелектуальної обробки даних / С.В. Лєнков, В.М. Джулій, О.М. Горбатюк, Н.М. Берназ // Збірник наукових праць Військового інституту Київського національного університету імені Тараса Шевченка. – 2014. – № 46. – С. 181-190.

7. Xianjun N. Research of Data Mining Based on Neural Networks / N. Xianjun // World Academy of Science, Engineering and Technology. – 2008. – Vol. 15. P. 381-384.

8. Yang J. Joint unsupervised learning of deep representations and image clusters / J. Yang, D. Parikh, D. Batra // In CVPR. – 2016. P. 5147–5156.

9. Yang B. Towards k-means-friendly spaces: Simultaneous deep learning and clustering / B. Yang, X. Fu, N. D. Sidiropoulos, M. Hong // arXiv preprint arXiv: 1610.04794. – 2016.

10. Li G. Data warehouse and data mining / G. Li, L. Hongjun // Microcomputer Applications. – 1999. – Vol. 15(9). P. 17-20.

11. Hand D. Principles of Data Mining / D. Hand, H. Mannila, P. Smyth // The Massachusetts Institute of Technology Press, 2001. – 546 p.

References (MLA)

1. Storchak K.P., Tushych A.M., Kozelkova K.S., Stepanov M.M. “Intelligent analysis of data using neural networks.” *Communication* 4 (2008): 17-19. Print.

2. Tushych A.M. “Analysis of the feasibility of using automated data mining system based on artificial neural networks.” *VII All-Ukrainian Scientific and Practical Conference of Students, Postgraduates and Young Scientists on Automatic Control. Kherson*. (April 10-12, 2019).:76-77.

3. Kulakov P.A. “Basic classes of neural networks in the problem of diagnostics of technological equipment.” *Electronic scientific and practical magazine "Youth scientific journal"* 10 (2016): 74-77. Print.

4. Dyuk V.A. “Data Mining - Intelligent Analysis of Data.” *Information technology: website.* URL: http://www.inftech.webservis.ru/it/database/datamining/ar2.html (date of request 03.12.2018)

5. Nazarov A.V. Loskutov A.I. *Neural network algorithms for prediction and optimization of systems.* Science and Technology, 2015. Print.

6. Lienkov S.V., Dzhulii B.M., Horbatiuk O.M., Bernaz H.M. “Conceptual scheme of the system of intellectual data processing.” *Collection of scientific works of the Military Institute of Kyiv National Taras Shevchenko University* 46 (2014): 181-190. Print.

7. Xianjun N. “Research of Data Mining Based on Neural Networks.” *World Academy of Science, Engineering and Technology* 15 (2008): 381-384. Print.

8. Yang J., Parikh D., Batra D. “Joint unsupervised learning of deep representations and image clusters.” *In CVPR* (2016): 5147–5156. Print.

9. Yang B., Fu X., Sidiropoulos N. D., Hong M. “Towards k-means-friendly spaces: Simultaneous deep learning and clustering.” *arXiv preprint arXiv:1610.04794*. (2016). Print.

10. Li G., Hongjun. L. “Data warehouse and data mining.” *Microcomputer Applications* 15(9). (1999): 17-20. Print.

11. Hand D., Mannila H., Smyth P. *Principles of Data Mining.* The Massachusetts Institute of Technology Press, 2001. Print.

Автори статті (Authors of the article)

Тушич Аліна Миколаївна – старший викладач кафедри інформаційних систем та технологій (Tushych Alina Mykolaivna – Senior Lecturer of the Department of Information Systems and Technologies). Phone.: +38(099) 602 81 68. E-mail: alinatushych@gmail.com.

Сторчак Каміла Павлівна – д.т.н., завідувач кафедри інформаційних систем та технологій (Storchack Kamila Pavlivna – D.Sc. in Technics, Head of the Department of Information Systems and Technologies). Phone.: +38(044) 249 25 42. E-mail: kpstorchak@ukr.net.

Бондарчук Андрій Петрович – д.т.н., директор інституту інформаційних технологій (Bondarchuk Andrii Petrovych – D.Sc. in Technics, Director of the Institute of Information Technologies). Phone.: +380974086131. E-mail: dekan.it@ukr.net.

Макаренко Анатолій Олександрович – д.т.н., професор кафедри мобільних та відеоінформаційних технологій (Makarenko Anatolii Oleksandrovych – D.Sc. in Technics, professor of the Department of mobile and video information technologies). Phone.: +38(097) 509 00 33. E-mail: makarenkoa@ukr.net.

Рецензент: канд. техн. наук, доцент **В. Б. Каток**, ПАТ "Укртелеком", Київ.