Науковий вісник НЛТУ України
Scientific Bulletin of UNFU
https://nv.nltu.edu.ua
https://doi.org/10.15421/40290626
Article received 28.05.2019 p.
Article accepted 27.06.2019 p.
УДК 004.[9+89]
ISSN 1994-7836 (print)
ISSN 2519-2477 (online)
@ Correspondence author
V. R. Verhun
vverhun@gmail.com

***В. Р. Вергун***
*Національний університет "Львівська політехніка", м. Львів, Україна*

# ХАРАКТЕРИСТИКА МЕТОДІВ РОЗВ'ЯЗАННЯ ЗАДАЧІ КЛАСИФІКАЦІЇ В ІНТЕЛЕКТУАЛЬНОМУ АНАЛІЗІ ДАНИХ НАВЧАЛЬНИХ ПРОГРАМ

Досліджено публікації останніх років у галузі інтелектуального аналізу даних навчальних програм. Кількість досліджень у цій галузі зростає, проте здебільшого це однотипні дослідження, що використовують однакові вибірки даних. Розроблено критерії, відповідно до яких було отримано вибірку з публікаціями для проведення аналізу використання методів інтелектуального аналізу даних навчальних програм. Найбільше досліджень у галузі Інтелектуального Аналізу Даних у навчанні стосуються вирішення задачі кластеризації, класифікації та асоціації. Для створення вибірки до уваги обрано дослідження з використанням методів та алгоритмів, що вирішують задачу класифікації. Вибірка статей включає дослідження, що аналізують продуктивність методів класифікації та представляють результати та порівняння показників. За результатом аналізу вибрано алгоритми, що показують найкращі результати продуктивності серед інших алгоритмів з вибірки. Згідно із встановленими критеріями, кожна публікація повинна вирішувати конкретну наукову задачу. У цій галузі методи інтелектуального аналізу даних отримують застосування для вирішення різних прикладних задач у навчальному процесі. Відповідно до контексту та типу прикладної задачі залежить вибір конкретного методу та точність вибраних алгоритмів. Тому категоризація прикладних завдань дає змогу отримувати якісніші підходи до розв'язання наукової задачі. Встановлено категорії проблематики, яких стосуються найбільше наукових досліджень з використанням методів класифікації.

***Ключові слова:*** інтелектуальний аналіз навчальних програм; класифікація; продуктивність; алгоритми; навчальні програми; прогнозування; порівняння алгоритмів.

**Вступ.** У сучасних умовах використання інформаційних технологій значно покращують якість навчального процесу. Водночас основним завданням будь якого навчального закладу є високі оцінки учасників навчання. Беручи до уваги різноманітні стилі та методи навчання, поведінку студентів та різноманітні підходи до викладання не можна однозначно стверджувати, що саме інформаційні технології мають прямий вплив на остаточні результати навчання. Впровадження технологій дає змогу значно розширити інструментарій, методи навчання, урізноманітнити спосіб доставки знань. Також інформаційні системи допомагають якісно управляти навчальним процесом. Впродовж останнього десятиліття інформаційні системи стали невід'ємною частиною навчання та нерозривно інтегровані в процеси прийняття рішень у будь-яких навчальних закладах.

Унаслідок використання будь-яких інформаційних систем нагромаджується велика кількість даних. Структура на склад таких даних може бути найрізноманітнішою. Системи зберігають інформацію про активність студентів, вподобання, засоби комунікації, певні персональні дані тощо. Ці дані можуть бути об'єктом для аналізу та можуть бути використані для прогнозування успішності, створення індивідуальних навчальних планів, визначення поведінки студента та стилю навчання, створення моделей мотивації. З таких наборів даних можливо виявити певні шаблони та закономірності. Таку задачу розв'язують за допомогою Інтелектуального Аналізу Даних (англ. *Data Mining*, ІАД) – процесу видобутку знань з даних. ІАД – це міждисциплінарна галузь, що використовує різні методи та алгоритми для розв'язання задач кластеризації, класифікації, асоціації, прогнозування. Залежно від моделей, що використовуються, задачі можуть бути прогнозуючими або дескриптивними. Пошуком закономірностей та знань у даних, що генеруються різноманітними системами, що інтегровані в навчальний процес, займається більш вузька галузь, що похідна від ІАД – Інтелектуальний Аналіз Даних Навчального Процесу (англ. *Educational Data Mining*). Методи та моделі, розроблені в межах цієї галузі, дають змогу краще розуміти поведінку студентів та прогнозувати їхню успішність.

Прогнозування поведінки студентів та стилів навчання є одним із найвагоміших завдань у навчальному процесі, таким як і прогнозування успішності. Розв'язання такої задачі дає змогу гнучкіше керувати індивідуальними навчальними планами. Різноманітні чинники мають вплив на успішність та стиль навчання, такі як: вік, стать, освіта батьків, економічні чинники. Вирішення задачі класифікації під час процесу ІАД дає змогу викладачу індивідуально підходити до потреб конкретного студента, що приводить до вищих фінальних результатів та значно підвищує мотивацію під час навчання.

Методи задачі класифікації є найпопулярнішими в

**Інформація про авторів:**
**Вергун Володимир Ростиславович**, аспірант, кафедра автоматизованих систем управління. **Email:** vverhun@gmail.com
**Цитування за ДСТУ:** Вергун В. Р. Характеристика методів розв'язання задачі класифікації в інтелектуальному аналізі даних навчальних програм. Науковий вісник НЛТУ України. 2019, т. 29, № 6. С. 136–139.
**Citation APA:** Verhun, V. R. (2019). Review of classification methods in educational data mining. *Scientific Bulletin of UNFU, 29*(6), 136–139. https://doi.org/10.15421/40290626

інтелектуальному аналізі даних навчального процесу згідно з аналізом наукових досліджень останніх років (Bishop, 2006).

*Метою цього дослідження* є:

- Дослідити останні публікації в галузі Educational Data Mining.
- Створити вибірку публікацій, що включають дослідження з використанням методів, що розв'язують задачу класифікації.
- Визначити, які конкретні задачі вирішують автори за допомогою методів класифікації в цих дослідженнях.
- Дослідити вибрані авторами методи у вибірці публікацій та визначити методи з найкращими показниками продуктивності.

**Аналіз літературних джерел.** Кількість статей, що стосуються дослідження у цій галузі, щороку збільшується (Singh, 2017), що свідчить про зростаючий інтерес та великі можливості застосування інтелектуального аналізу даних задля покращення процесу навчання. Водночас існує багато досліджень поточного стану галузі та огляд літературних джерел з аналізом вже опублікованих досліджень. Більшість досліджень зосереджені на таких запитаннях:

- Знаходження визначальних чинників, які мають вплив на успішність у навчальному процесі.
- Пошук оптимальних методів та алгоритмів для прогнозування успішності.
- Визначення точності та продуктивності вибраних методів та підходів.

З використанням методів інтелектуального аналізу даних прогностичне моделювання зазвичай використовують у прогнозуванні успішності студентів. Загалом кількість досліджень у сфері інтелектуального аналізу даних у освітніх програмах швидко зростає, а також збільшується різноманітність використовуваних методів (Hellas et al., 2018). Відповідно в дослідженні (Muthukrishnan et al., 2017) було розглянуто методи, що використовують для створення моделей, та прогнозування успішності. Ці методи було поділено на 4 категорії: дерево прийняття рішень, регресію, кластеризацію та всі решта.

У дослідженні (Manjarres et al., 2018) було проаналізовано понад 100 публікацій і встановлено, що, розпочинаючи з 2010 р., значний інтерес у дослідженнях приділено аналізу причин відрахувань з навчальних програм та побудови прогностичних моделей. В одному з останніх оглядів публікацій було класифіковано чинники, що брались до уваги найчастіше в дослідженні причин відрахувань: персональні, академічні, економічні, соціальні та інституційні. І найбільш досліджуваними є персональні чинники, такі як: вік, стать, національність (Alban et al., 2019).

Проте традиційні алгоритми та підходи інтелектуального аналізу даних не можуть бути безпосередньо застосовані до вирішення проблем у навчальному процесі, оскільки вони можуть мати специфічну мету та функцію. Це означає, що спочатку повинен бути застосований алгоритм попереднього оброблення і тільки тоді можуть бути застосовані деякі специфічні методи аналізу даних. Одним із таких алгоритмів попередньої обробки є кластеризація (Dutt et al., 2017). Систематичні огляди літератури підтверджують загальну тенденцію у використанні методів лінійної регресії та класифікації, які є найбільш популярними до використання в різноманітних дослідженнях (Hellas et al., 2018). Проте все ще залишаються запитання до якості досліджень, зосереджених у проблематиці прогнозування успішності. Тільки 33 % досліджень мають чітку постановку задачі й тільки 8 % досліджень перевіряли результати в декількох наборах даних навчальних програм (Hellas et al., 2018).

**Метод дослідження.** Більшість наукових робіт, що використовують методи класифікації у своїх дослідженнях, можна поділити за такими категоріями:

- дослідження швидкодії вибраних методів та алгоритмів;
- дослідження чинників, що впливають на вирішення певної проблеми;
- нові методи, що можна застосовувати в задачі класифікації.

Для більш якісного аналізу наукових робіт було вибрано статті, що досліджують швидкодію методів та алгоритмів, які розв'язують задачі класифікації. Було проаналізовано вибірку з більше, ніж 100 статей, та вибрано 20 публікацій, що були опубліковані впродовж останніх 5–7 років, та які підпадали під критерії вибірки. Основними критеріями вибірки були:

- чітка постановка наукової задачі з окресленою проблемою;
- використання алгоритмів класифікації;
- продемонстровані результати продуктивності вибраних методів.

Оскільки дуже часто автори використовують у дослідженні декілька алгоритмів, до уваги брали тільки перші 4 алгоритми з найкращим показником точності в конкретній публікації. Також, беручи до уваги результати інших публікацій на тему літературних оглядів (Hellas et al., 2018), у вибірку попадали тільки ті дослідження, що використовували унікальну неповторювану вибірку даних.

**Результати дослідження.** У табл. 1 наведено список завдань, що вирішували автори за допомогою методів класифікації.

**Табл. 1. Список завдань**

| Завдання | Кількість статей з вибірки |
|---|---|
| Прогнозування успішності | 15 |
| Виявлення студентів, які перебувають під ризиком | 2 |
| Прогнозування відчислень за неуспішність | 3 |
| Класифікація стилів навчання | 1 |

У табл. 2 наведено список усіх методів та алгоритмів, які використовували авторами для вирішення поставлених завдань.

**Табл. 2. Список усіх методів та алгоритмів**

| Алгоритм | Кількість статей з вибірки |
|---|---|
| J48 | 11 |
| Naive Bayes | 12 |
| Zero R, PART, COMP, Decision Stump, MLP, C-Support Vector Classification, | 1 |
| Logistic Regression, Random Forest, Multilayer perceptron, KNN, Support Vector Machines | 5 |
| RBFNetwork, RepTREE, NBTree | 2 |
| JRip | 3 |

Як видно з табл. 2, найчастіше в цій вибірці досліджень використовували методи J48 та Naive Bayes. Загалом у вибірці статей було проаналізовано 19 алгоритмів. Достовірність отриматих математичних моделей автори досліджень здебільшого оцінювали методом перехресної перевірки (10-fold cross-validation).

Середні значення точності кожного з алгоритмів представлено у табл. 3. Середню цифру обраховували без урахування мінімального та максимального значень.

**Табл. 3. Середні значення точності алгоритмів, %**

| | |
|---|---|
| **J48** (Kabakchieva, 2013) (Kaur et al., 2015) (Osmanbegović et al., 2015) (Bhavesh Patel et al., 2017) (Umar Bawah al., 2018) (Shakeelet al., 2016) (Al Luhaybi et al., 2018) (Kabakchieva, Dorina., 2013) (Maaliw et al., 2017) (Kaur et al., 2017) (Jovanovic et al., 2012) | 78 |
| **Naive Bayes** (Kabakchieva, 2013) (Singh et al., 2017) (Asif, Raheela, et al., 2017) (Kaur et al., 2015) (Bhavesh Patel et al., 2017) (Lopez, Manuel Ignacio, et al., 2012) (Shakeelet al., 2016) (Al Luhaybi et al., 2018) (Kabakchieva, Dorina., 2013) (Veena, 2017) (Maaliw et al., 2017) (Jovanovic et al., 2012) | 77 |
| **Zero R** (Singh et al., 2017) | 76 |
| **Logistic Regression** (Osmanbegović et al., 2015) (Bhavesh Patel et al., 2017) (Bucos et al., 2017) (Jovanovic et al., 2012)) (Marbouti et al., 2016) | 90 |
| **Support Vector Machines** (Kaur et al., 2015) (Agarwal et al., 2012) (Soni, Astha, et al. et al., 2018) (Lopez, Manuel Ignacio, et al., 2012) (Veena, 2017) (Kaur et al., 2017) (Marbouti et al., 2016) | 89 |
| **RBFNetwork** (Agarwal et al., 2012) (Lopez, Manuel Ignacio, et al., 2012) | 88 |
| **JRip** (Osmanbegović et al., 2015) (Mobasheret al., 2017) (Kabakchieva, Dorina., 2013) | 69 |
| **PART** (Mobasheret al., 2017) | 69 |
| **COMP** (Abu-Oda et al., 2015) | 98 |
| **Decision Stump** (Mobasheret al., 2017) | 70 |
| **MLP** (Umar Bawah al., 2018) | 84 |
| **C-Support Vector (**Bucos et al., 2017) | 85 |
| **Random Forest** (Asif, Raheela, et al., 2017) (Osmanbegović et al., 2015)(Shakeelet al., 2016) (Bucos et al., 2017) (Jovanovic et al., 2012) | 82 |
| **Multilayer perceptron** (Kaur et al., 2015) (Agarwal et al., 2012) (Lopez, Manuel Ignacio, et al., 2012) (Kaur et al., 2017) (Marbouti et al., 2016) | 84 |
| **KNN** (Kabakchieva, 2013) (Umar Bawah al., 2018) (Kabakchieva, Dorina., 2013) (Veena, 2017) (Marbouti et al., 2016) | 74 |
| **NBTree** (Abu-Oda et al., 2015) (Maaliw et al., 2017) | 87 |
| **RepTREE** (Mobasheret al., 2017) (Kaur et al., 2017) | 70 |
| | |

На рисунку зображено значення точності тих алгоритмів та методів, які трапляються в більше ніж 5 статтях. Ці значення точніші, оскільки значення точності має велику залежність від вибраних даних, тому для коректнішого аналізу потрібно враховувати більшу кількість статей.

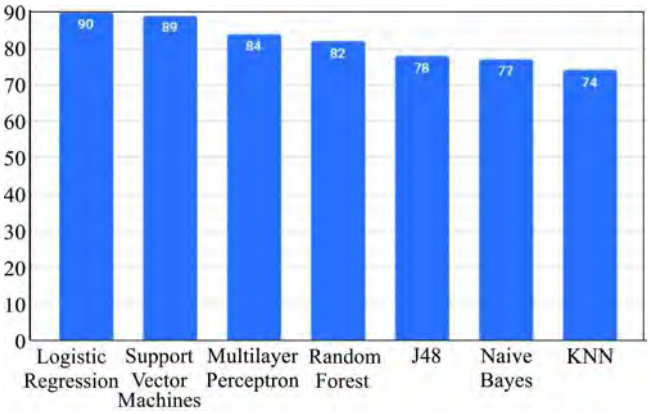


**Рисунок.** Значення точності найбільш вживаних алгоритмів, %

Відповідно до проаналізованих результатів з вибірки статей, алгоритмами з найбільшою точністю є Logistic, Multilayer Perceptron, Support Vector Machines (Functions-based algorithms), та Random Forest (Trees-based algorithms), точність яких перевищує 80 %. Водночас потрібно зазначити, що алгоритми, які найчастіше використовували в дослідженнях, не показали високих результатів продуктивності. Проте велика залежність існує від мети дослідження та від конкретної вибірки даних. Наприклад, метод Logistic Regression показує різні результати точності в дослідженнях, де метою було дослідити можливості прогнозування успішного завершення навчання (Bucos et al., 2018), та прогнозування відрахування з навчання (Jovanovic et al., 2012). Конкретна вибірка даних може містити різноманітні чинники, що тим чи іншим чином впливають на точність дослідження. Наприклад, під час прогнозування значень фінальних оцінок використовують чинник участі та активності студентів у форумах та дискусіях (Marbouti et al., 2016). Автор дослідження (Luhaybi et al., 2018) також звертає увагу на великий вплив контексту та мети на остаточні результати.

**Висновки.** У цьому дослідженні проаналізовано 120 публікацій, та за заданими критеріями було вибрано 21 публікацію, опублікованих за останні 5–7 років, де автори досліджують різноманітні методи, що використовуються для розв'язання задачі класифікації в галузі інтелектуального аналізу даних навчальних програм. Було визначено 4 алгоритми, що отримали найкращі середні показники продуктивності в цих дослідженнях. Встановлено, що найвищі показники продуктивності отримали алгоритми Logistic Regression та Support Vector Machines. Результати продуктивності методу KNN є найнижчими. Всі дані наведено в табличних та графічному представленні.

## Перелік використаних джерел


Abu-Oda, Ghadeer S., & Alaa M. El-Halees. (2015). Data mining in higher education: university student dropout case study. *Data mining in higher education: university student dropout case study, 5*(1), 13–19.

Agarwal, Sonali, Pandey, G. N., & Tiwari, M. D. (2012). Data mining in education: data classification and decision tree approach. *International Journal of e-Education, e-Business, e-Management and e-Learning, 2*(2), 140–145.

Al Luhaybi, Mashael, Tucker, Allan, & Yousefi, Leila. (2018). The Prediction of Student Failure Using Classification Methods. *A Case study*, 79–90. https://doi.org/10.5121/csit.2018.80506

Alban, Mayra, & David Mauricio. (2019). Predicting University Dropout through Data Mining: A Systematic Literature. *Indian Journal of Science and Technology, 12,* 4–9.

Asif, Raheela, et al. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers & Education, 113,* 177–194.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. NY: Springer, 260 p.

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5–32.

Bucos, Marian, & Bogdan Drăgulescu. (2018). Predicting student success using data generated in traditional educational environments. *TEM Journal, 7*(3), 617–620.

Dutt, A., Ismail, M. A., & Herawan, T. (2017). A Systematic Review on Educational Data Mining. *IEEE Access, 5,* 15991–16005. https://doi.org/10.1109/ACCESS.2017.2654247

Haykin, S. (2009). Neural networks and learning machines. *Upper Saddle River, 3,* 938. NJ, USA: Pearson.

Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V. V., et al. (2018). Predicting Academic Performance: A Systematic Literature Review. *In Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education*, (pp. 175–199). (ITiCSE 2018 Companion). New York, NY, USA: ACM. https://doi.org/10.1145/3293881.3295783

Jovanovic, Milos, et al. (2012). Using data mining on student behavior and cognitive style data for improving e-learning systems: a case study. *International Journal of Computational Intelligence Systems, 5*(3), 597–610.

Kabakchieva, D. (2013). Predicting Student Performance by Using Data Mining Methods for Classification. *Cybernetics and Information Technologies, 13*(1), 61–72. https://doi.org/10.2478/cait-2013-0006

Kabakchieva, Dorina. (2013). Predicting student performance by using data mining methods for classification. *Cybernetics and information technologies, 13*(1), 61–72.

Kaur, Parneet, Manpreet Singh, & Gurpreet Singh Josan. (2015). Classification and prediction based data mining algorithms to predict slow learners in education sector. *Procedia Computer Science, 57,* 500–508.

Lopez, Manuel Ignacio, et al. (2012). Classification via clustering for predicting final marks based on student participation in forums. *International Educational Data Mining Society*, 234 p.

Maaliw III, Renato R., & Melvin A. Ballera. (2017). Classification of Learning Styles in Virtual Learning Environment Using J48 Decision Tree. *International Association for Development of the Information Society*, 420 p.

Manjarres, Andrés Villanueva, Luis Gabriel Moreno Sandoval, & Martha Salinas Suárez. (2018). Data mining techniques applied in educational environments: Literature Review. *Digital Education Review, 33,* 235–266.

Marbouti, Farshid, Heidi A. Diefes-Dux, & Krishna Madhavan. (2016). Models for early prediction of at-risk students in a course using standards-based grading. *Computers & Education, 103,* 1–15.

Mobasher, Ghadeer, Ahmed Shawish, & Osman Ibrahim. (2017). Educational Data Mining Rule based Recommender Systems. *CSE-DU, 1.*, 34–39

Mr. Bhavesh Patel, & Dr. Jyotindra Dharwa. (2017). Selection of Optimal Classification Algorithms in Education Data Mining. *Imperial Journal of Interdisciplinary Research, 3*(1), 43–49.

Muthukrishnan, S. M., Govindasamy, M. K., & Mustapha, M. N. (2017). Systematic mapping review on student's performance analysis using big data predictive model. *Journal of Fundamental and Applied Sciences, 9*(4), 730–758.

Osmanbegović, Edin, Mirza Suljić, & Hariz Agić. (2015). Determining dominant factor for students performance prediction by using data mining classification algorithms. *Tranzicija, 16*(34), 147–158.

Parneet Kaur, Manpreet Singh, & Gurpreet Singh Josan. (2015). Classification and prediction based data mining algorithms to predict slow learners in education sector. *3rd Int. Conf. on Recent Trends in Computing.* (Vol. 57, pp. ), 134–141).

Pirotti, F., Sunar, F., & Piragnolo, M. (2016). Benchmark of machine learning methods for classification of a Sentinel-2 image. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences, 41,* 335–340.

Sapiton, M. (2019). How the IT industry of Ukraine and Eastern Europe works: a report. Retrieved from: https://ain.ua/en/2019/02/15/it-industry-of-ukraine-and-eastern-europe/

Shakeel, Khawar, & Naveed Anwer Butt. (2019). *Educational Data Mining to Reduce Student Dropout Rate by Using Classification*, 420 p.

Singh, Manmohan, Amit Dutta, & Ramesh Prasad Aharwal. Analysis: Classification Data Mining Process in Primary Education System. *An e-Journal of RBIMS, 1*(1), 34–39.

Soni, Astha, et al. (2018). Predicting student performance using data mining techniques. *International Journal of Pure and applied Mathematics, 119*(12), 221–227.

Umar Bawah, Faiza, & Ussiph, Najim. (2018). Appraisal of the Classification Technique in Data Mining of Student Performance using J48 Decision Tree, K-Nearest Neighbor and Multilayer Perceptron Algorithms. *International Journal of Computer Applications, 179*, 39–46. https://doi.org/10.5120/ijca2018916751

Veena, N., & Guruprasad, S. (2017). Comparative Analysis of Classification Algorithms for Student Performance. *International Journal of Science Technology & Engineering, 4*(2), 134–141.

***V. R. Verhun***
*Lviv Polytechnic National University, Lviv, Ukraine*

## REVIEW OF CLASSIFICATION METHODS IN EDUCATIONAL DATA MINING

Educational data mining is an emerging discipline, concerned with developing methods for exploring the unique and increasingly large-scale data that come from educational settings and using those methods to better understand students, and the settings which they learn in. Currently there is an increasing interest in data mining and educational systems, making educational data mining a new growing research community. The number of studies in this industry is constantly increasing; however, it is mainly the same type of research that uses the same data sets. The research of publications of recent years in the field of educational data mining has been conducted. In order to analyze the use of methods of the educational data mining in recent publications and for more precise selection of published articles for this analysis the criteria for selecting specific articles has been developed. According to developed criteria a data set with publications has been created. Most studies in educational data mining are devoted to solving the problem of clustering, classification and association. In this research the data set of publications has been created by taking into consideration the articles which includes the methods and algorithms for solving the classification problem. All the selected articles include researches that analyze the performance of classification methods and present results and benchmarking of those methods. As a result of the analysis of created data set of articles the algorithms that show the best performance results among other are highlighted. According to the established criteria, each publication from data set should address a specific scientific problem. Methods of educational data mining are used to solve various applied problems in the learning process. Depending on the context and type of application, the choice of a specific method and the accuracy of the selected algorithms highly depend on a specific application problem. Therefore, categorization of applied tasks allows obtaining more qualitative approaches to solving a scientific problem. The categories of problems most often solved by educational data mining methods have been established.

***Keywords:*** data mining; educational data mining; classification; performance; educational programs; performance prediction.