УДК 37:004; 303.4.025

Ковальчук Юрій Олексійович
доцент, кандидат фізико-математичних наук, декан факультету
Ніжинський державний університет імені Миколи Гоголя, м. Ніжин, Україна
yu.kovalchuk@i.ua

ПОШУК, ОТРИМАННЯ Й АНАЛІЗ ДАНИХ В ОСВІТІ: СУЧАСНИЙ СТАН І ПЕРСПЕКТИВИ РОЗВИТКУ

Анотація. Розглянуто основні задачі (класифікація і регресія, пошук асоціативних правил, кластеризація) і принципи роботи базових алгоритмів Data Mining у контексті їх використання для різноманітних досліджень у галузі освіти, що є предметом відносно нового самостійного напряму Educational Data Mining. Наведені дані про найпопулярніші теми досліджень у рамках даного напряму, окреслені перспективи його розвитку.
Стаття розрахована на читачів, які займаються дослідженнями в галузі освіти на різних рівнях, особливо тих, хто причетний до використання систем електронного навчання, але мало знайомий із цим напрямом аналізу даних.

Ключові слова: Educational Data Mining; класифікація; регресія; пошук асоціативних правил; кластеризація; система електронного навчання.

1. ВСТУП

Постановка проблеми. Значна частина досліджень у галузі освіти спирається на аналіз емпіричних даних, що мають вигляд кількісних або якісних характеристик масових сутностей чи процесів. Починаючи з кінця XIX століття, основним інструментом аналізу даних була (і значною мірою залишається) математична статистика. У класичному розумінні це наука про: моделювання випадкових величин, представлених вибірковими даними, у вигляді параметричних розподілів на метричних шкалах; виявлення залежностей між випадковими величинами; перевірка гіпотез щодо значущості рівностей розподілів, їх форм та мір зв'язку. У другій половині XX століття помітної популярності набули методи непараметричної статистики, які дозволяють досліджувати дані, виміряні у порядкових шкалах.

Поступово термін «математична статистика» почав витіснятися з ужитку більш загальним поняттям «аналіз даних». Це було пов'язано передусім з появою ефективних комп'ютерних алгоритмів пошуку закономірностей у даних, часто представлених у категоріальних шкалах, які, утім, не завжди вдавалося обґрунтувати математично. Наразі, уже наприкінці минулого століття стало очевидним, що дані, які можуть представляти інтерес для прикладних досліджень, усе частіше знаходяться у цифровому вигляді на носіях комп'ютерних систем, і обсяг цих даних швидко переходить за принципово важливу для людини межу: усі дані в сегменті, який цікавить дослідника, *ніколи не будуть ним переглянуті у звичайний, «фізичний», спосіб*. З'являється важливий клас задач і методів їх розв'язання, який називають *пошуком знань у базах даних* (KDD – Knowledge Discovery in Databases), або, частіше, *видобутком (розкопкою) даних* (DM – Data Mining).

Звичайно ж, метафора бурхливого дрейфу даних у цифровий океан є доречною і для галузі освіти. Тут потенційним джерелом знань для дослідника все частіше є як адміністративні бази освітніх даних рівня навчального закладу, регіону чи держави, так і Web, і, звичайно ж, бази даних і лог-файли різноманітних систем комп'ютерної підтримки навчання – CMS, LMS, ITS, системи комп'ютерного адаптивного навчання, тестування рівня навчальних досягнень тощо. Тому, закономірно, з поступовим

розповсюдженням DM, особливо після того, як навчальні курси з даної дисципліни почали вивчатися студентами IT спеціальностей (у тому числі й в Україні), а також унаслідок специфічності галузі освіти, виокремився напрям, який у світі отримав назву Educational Data Mining (EDM). Специфіка EDM порівняно із загальним DM полягає головним чином у складності визначення моделей і цільових змінних, максимізація яких приводила б до поліпшення якості освіти. У цьому EDM істотно відрізняється від застосування DM, скажімо, в економічній сфері, де цільова змінна, зазвичай, вимірюється у грошових одиницях на числовій шкалі. Указані відмінності часом призводять до необхідності використання специфічних технологій DM (особливо в галузі освітніх вимірювань), хоча у більшості випадків вистачає стандартного арсеналу алгоритмів і відповідного програмного забезпечення.

На жаль, дані вітчизняних наукових джерел свідчать про те, що українським дослідникам, які працюють в царині освіти, цей важливий напрям аналізу й осмислення освітньої інформації все ще мало відомий. Більшість українських дослідників нічого не знають ні про задачі, які можна ефективно розв'язувати за допомогою DM, ні про наявний інструментарій. У такій ситуації, очевидно, нашим науковцям слід передусім ознайомитися зі сферою можливого застосування EDM, наявним програмним забезпеченням та суттю і базовими параметрами настроювання роботи алгоритмів, тоді як розуміння тонкощів функціонування програм є бажаним, але не необхідним. Застереження ж тут можуть бути подібними до випадку використання популярних статистичних пакетів (новітні версії яких, до речі, зазвичай включають модулі DM) – використовуючи технологію як «чорну скриньку», без належного рівня розуміння специфіки задачі, дослідник ризикує дійти неправильних висновків.

Аналіз останніх досліджень і публікацій. DM як окремий напрям комп'ютеризованого аналізу даних почав формуватися з кінця 1980-х років. У середині першої декади нинішнього століття утворилася міжнародна робоча група з EDM, яка переросла у 2011 році в International Educational Data Mining Society (*http://www.educationaldatamining.org*). Цікаво, що хоча у складі цієї організації найбільше науковців із США, нині її президентом є випускник Харківського національного університету радіоелектроніки Микола Печенізький (Mykola Pechenizkiy), який працює зараз у Технологічному університеті Ейндховена. Організація опікується щорічними міжнародними конференціями (International Conference on Educational Data Mining), які проводяться з 2008 року, а також видає журнал JEDM – Journal of Educational Data Mining.

З основами DM можна ознайомитися за книгами [1; 10]. Основи теорії і практики байєсівських мереж у доступній для нефахівців формі викладено в книзі [5], а використання цієї перспективної технології в освітніх вимірюваннях детально висвітлено в [3].

Існує кілька англомовних книг з тематики EDM [6; 7].

У 2010 році вийшов чудовий огляд літератури з даної тематики [9], який містить аналіз 306 наукових публікацій. Зокрема, усі публікації згруповані за типом джерела освітніх даних, які використовувалися їх авторами (звертає на себе увагу, що приблизно кожна десята стаття спирається на дані, отримані з неелектронного, традиційного освітнього процесу), наведено діаграми розподілу кількості публікацій за роками (які вказують на прискорене, фактично експоненціальне зростання), а також проаналізовано зміст цих публікацій за 11 тематичними категоріями. Стаття завершується переліком перспективних напрямів EDM.

Іншими авторами у 2009 році також опубліковано огляд сучасного стану і перспектив розвитку EDM [4]. Ними запропонована дещо відмінна класифікація напрямів EDM, проаналізовано зміст восьми найбільш цитованих статей, а також

наведено розподіл кількості статей у матеріалах згаданої вище міжнародної конференції 2008 і 2009 років за кількістю публікацій з шести виділених авторами напрямів. Варто відзначити, що порівняно з більш раннім оглядом [8], можна дійти висновку про істотне зростання інтересу до такого напряму EDM, як прогнозування, тоді як більш ранні дослідження тяжіли до виявлення взаємозв'язків між даними.

Усі три згадані вище оглядові статті орієнтовані на читачів, уже обізнаних з технологіями DM. У них майже не розглядається суть методів, формалізація задач та форми очікуваних результатів, достатніх для того, щоб самостійно використовувати спеціалізоване програмне забезпечення тими науковцями, які не є фахівцями з комп'ютеризованого аналізу даних, проте бажають використовувати математико-статистичні методи і технології у своїх дослідженнях.

В Україні, як зазначалося, цей напрям досліджень розвинений мало. Так, термін Data Mining серед усіх 47 випусків журналу «Інформаційні технології і засоби навчання» згадується лише один раз та й то побіжно, у статті на іншу тему. Проте пошук у Google Scholar за запитом «data mining освіта» виявив 850 публікацій, що свідчить про наявність вітчизняних науковців, обізнаних з даною тематикою і діючих у ній. Більш детальний аналіз україномовних джерел виконати важко, оскільки вони розкидані по різноманітних університетських збірниках, матеріалах конференцій тощо.

Мета статті. Метою статті є ознайомлення читачів з типовими задачами DM, прикладами, у тому числі й авторськими, їх застосування для різноманітних досліджень у галузі освіти, доступним програмним забезпеченням. Автор сподівається на те, що для шанувальників журналу «Інформаційні технології і засоби навчання» ця оглядова стаття послужить поштовхом до застосування методів DM у їх майбутніх дослідженнях.

2. РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

2.1. Суть і задачі Data Mining

Data mining (DM) – це галузь аналізу даних, метою якої є пошук машиною (алгоритмом) взаємозв'язків між змінними у великих масивах цифрових даних. Такими масивами даних є зазвичай, бази даних або тексти. Для алгоритмів аналізу текстів використовується окремий термін – *Text Mining*, чим підкреслюється значна відмінність задач і відповідних технологій порівняно з класичним Data Mining, який ще називають часом пошуком знань у базах даних (*KDD – Knowledge Discovery in Databases*). Ми не будемо тут розглядати Text Mining через його специфіку, зауважимо лише, що чи не найголовнішою проблемою машинного аналізу текстів є їх попередня підготовка, що полягає у їх перетворенні до деякого канонічного виду.

Що стосується класичного DM, то, не дивлячись на велике розмаїття алгоритмів, які у ньому використовуються, ці алгоритми спрямовані на виконання однієї з трьох основних класів задач:

- класифікація і регресія;
- пошук асоціативних правил і секвенційний аналіз;
- кластеризація.

Деякі фахівці заперечують проти включення до арсеналу DM регресії, тому що тут технології зазвичай обмежуються застосуванням апарату регресійного аналізу як розділу «традиційної» математичної статистики.

Наведена вище проста класифікація задач є цілком достатньою для уведення в суть DM, а різні класифікації за практичним застосуванням DM в освіті розглянемо пізніше.

Базу даних, з якою працює деякий конкретний алгоритм, можна уявляти як просту плоску (двовимірну) таблицю на зразок тих, що зберігаються у файлах табличних процесорів.

Приклад такої таблиці:

Таблиця 1

Гіпотетичні дані про студентів 1 курсу

ID	Прізвище	Бал сертифікату ЗНО з математики	Бал сертифікату ЗНО з української мови та літератури	Рівень успішності за середнім балом атестату	Чи отримує стипендію у II семестрі
1	Богун	177	180	Високий	Так
2	Непитайло	160	170	Високий	Ні
3	Сірошапка	159	145	Низький	Так
…	…	…	…	…	…
100	Тягнирядно	165	173	Середній	Ні

Ця таблиця могла бути сформованою з інформації, що міститься у базах даних ЄДЕБО і деканату.

Кожен рядок таблиці містить дані про один об’єкт дослідження, у нашому прикладі це студент. Кожен стовпець, крім чисто «технічного» першого, містить значення про певну змінну-характеристику (атрибут) об’єкта, який становить інтерес для дослідника. Припустимо, що всі клітинки таблиці заповнені, хоча в реальних базах даних це далеко не завжди так.

Кожен атрибут є змінною певного типу. Якщо змінна може набувати будь-яких значень з деякого числового діапазону, то її називають неперервною, хоча реальні дані завжди дискретні. Тут йдеться не про строго математичне розуміння неперервності, а швидше про те, що дані у відповідному стовпці ніколи не повторюються (або повторюються відносно рідко). Для роботи деяких алгоритмів вимагається, щоб усі змінні були дискретними у тому розумінні, щоб кожна змінна мала малу кількість можливих значень, як у двох останніх стовпцях таблиці 1: змінна *Рівень успішності за середнім балом атестату* набуває одного із трьох значень *Високий*, *Середній*, *Низький*; змінна *Чи отримує стипендію у II семестрі* набуває одного із двох можливих значень. У таких випадках про окремі значення змінних говорять як про класи. Неперервні змінні можна легко перетворити на дискретні, розбивши діапазон зміни значень змінної на проміжки. Саме так могла бути отримана змінна *Рівень успішності за середнім балом атестату*, на що натякає сама її назва.

Укажемо на дві важливі проблеми, які можуть виникнути під час аналізу даних подібних таблиць. Перша проблема пов’язана з пропусками – можливою відсутністю даних у деяких клітинках таблиці. Є два основних способи боротьби з цим явищем: вилучити об’єкти з пропусками з розгляду, або замінити пропущені дані деякими правдоподібними (усередненими) значеннями. Оптимальне розв’язання проблеми залежить від конкретного випадку. Друга проблема може виникнути тоді, коли якесь конкретне значення атрибуту рідко зустрічається. Зрозуміло, що з рідкісних випадків неможливо робити статистично значимі висновки. Алгоритми, які не враховують цю проблему, називають *надчутливими* (*overfitting*). Для таких алгоритмів можна зарадити проблемі, об’єднавши класи з малою кількістю значень у більш крупні класи.

Розглянемо класи задач DM у розрізі їх застосування до розв’язання різних задач в дослідженнях у галузі освіти.

2.2. Задача класифікації

Суть. Припустимо у контексті наведеного прикладу (таблиця 1), що умови набору нових студентів на навчання в даний університет на дану спеціальність не змінилися, і контингент випускників шкіл України за розподілом успішності з різних предметів і середнім балом атестату подібний до минулорічного (точніше, випускники обох років належать до однієї популяції за розподілом успішності). Для кожного нового студента ми отримаємо новий рядок у таблиці 1, значення якого в останньому стовпці (стипендія у II семестрі) є все ще невідомим. З різних причин університет може бажати наперед спрогнозувати, чи буде такий студент отримувати стипендію (наприклад, для того, щоб визначити плановий обсяг стипендіального фонду для нового набору студентів). Іншими словами, потрібно спрогнозувати, до якого з двох класів (*Так* чи *Ні*) за цим атрибутом буде належати новий студент, тобто *розв'язати задачу класифікації*.

Зрозуміло, що спрогнозувати значення невідомого атрибуту (залежної змінної) за значеннями відомих атрибутів (незалежних змінних) можна, спираючись на попередній досвід, за умови, якщо залежна змінна дійсно залежить від незалежних змінних. Наш попередній досвід, тобто повні дані про студентів, містяться у таблиці 1. Таку таблицю з відомими даними залежної змінної називають *навчаючою вибіркою*, або просто *учителем*. Суть назви полягає у тому, що алгоритм, який ми збираємося застосувати для класифікації нового об'єкта, повинен спочатку навчитися це робити на основі повних відомих даних про інші об'єкти. Зауважимо, що у цьому випадку DM перетинається з іншим напрямом комп'ютерної науки – *машинним навчанням* (*Machine Learning*).

Форми представлення результатів і основні класи алгоритмів. Ми хотіли б, щоб результатом роботи алгоритму було не просто визначення класу для одиничного об'єкта, а правило чи набір правил, за якими можна було б класифікувати об'єкти вже без потреби наново запускати відповідну комп'ютерну програму. Найпростіша форма результату – це так зване 1-правило вигляду *якщо…то…*, яке дозволяє класифікувати об'єкт лише за значенням однієї незалежної змінної. Наприклад: Якщо *Бал сертифікату ЗНО з математики* більший ніж 172, то *Стипендія* дорівнює «так». Але все ж ми хочемо отримати більш детальну, а значить, більш точну «інструкцію» для класифікації. Така інструкція вже має вигляд ієрархічного набору вкладених правил зазначеного виду, зображуваних часто у вигляді графа, яке називають **деревом рішень**.

На рис. 1 зображено один із варіантів дерева рішень, отриманого у програмі Deductor Studio для задачі виявлення причин рішення випускників школи навчатися на фізико-математичному факультеті Ніжинського університету (раніше не публікувалося).

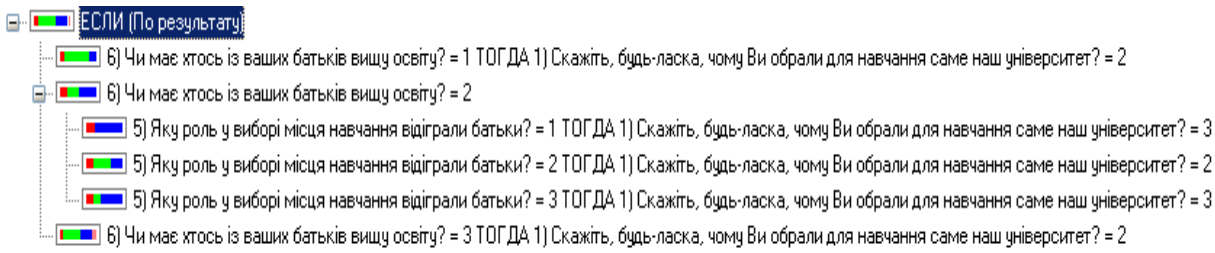

Рис. 1. Приклад дерева рішень, отриманого у середовищі Deductor Studio

Зображене на рисунку дерево рішень має одночасно вигляд багаторівневого списку. Його вузли – кольорові прямокутники, які містять у вигляді кольорів

інформацію про важливі характеристики – *підтримку*, *достовірність* та *значимість* кожного правила. Загалом, чим лівіше знаходиться вузол, тим більш значимим для визначення класу об'єкта є відповідний атрибут. Раніше ми вказували на важливість того, щоб кожне зі значень змінної зустрічалося достатньо часто. Це є змістом поняття підтримки правила. Також корисно контролювати, наскільки об'єкти самої навчаючої вибірки відповідають виведеним правилам. Адже одні й ті ж реальні об'єкти нерідко класифікуються по-різному. Наприклад, автору відомий випадок, коли один із першокурсників мав найвищу суму балів ЗНО і балу атестату серед однокурсників факультету, й одночасно – найнижчий середній бал за результатами першої сесії.

Дерева рішень добре допомагають зрозуміти структуру даних і взаємозв'язки між змінними. Найвідомішим алгоритмом побудови дерева рішень є алгоритм під назвою *C4.5*. Недоліком цього алгоритму є його *жадібність* (термін з теорії алгоритмів, який означає, що алгоритм на кожному кроці роботи видає оптимальне рішення саме для цього кроку, а не для кінцевого результату, тобто не вміє жертвувати вигодою заради більшої вигоди в майбутньому). Можна будувати дерево лише для якогось конкретного значення класу залежної змінної, використовуючи так званий *алгоритм покриття*. Зауважимо, що для кращого розуміння взаємозв'язків між даними дерева рішень зазвичай можна *обрізати*, тобто усувати ті гілки, які не є достатньо значимими.

Інша група алгоритмів реалізує так званий *байєсівський підхід* до розуміння ймовірності подій. Згідно з цим підходом, за навчаючою вибіркою визначаються спочатку апріорні ймовірності належності об'єкта до кожного з класів. Якщо 30% студентів отримували минулого року стипендію, то апріорі для нового студента, про якого ми ще нічого не знаємо, існує ймовірність 0,3 отримати стипендію. Далі, коли ми дізнаємося значення незалежних змінних для цього студента (його успішність за ЗНО й атестатом), це є подія, яка містить додаткову інформацію, що дозволяє *a posteriori*, тобто з врахуванням нових даних, переобчислити цю ймовірність. Відомий алгоритм *Naïve Bayes* (наївний алгоритм Байєса) працює за припущення, що незалежні змінні є незалежними також між собою у ймовірнісному сенсі. У нашому прикладі із студентами це не так, оскільки завжди існує додатна кореляція між оцінками ЗНО з різних предметів (і балом атестату). Результатом роботи алгоритму є не тільки прогноз класу об'єкта, а й імовірність цього прогнозу. Якщо ми хочемо враховувати залежності між змінними, потрібно користуватися технологією *байєсівських мереж*. На сьогодні цей напрям аналізу даних бурхливо розвивається й має вигляд одного з найперспективніших. Достатньо зауважити, що байєсівські мережі вже використовуються в освітніх вимірюваннях наряду з класичними методами психометрії, у рамках нової парадигми під назвою *Evidence-centered Design* [3].

До арсеналу DM відносять також технології *нейронних мереж* і метод *SVM* ––– *Support Vector Machine*, які ми тут не розглядаємо.

2.3. Задача регресії

Суть. У пункті 2.1 зазначалося, що задача регресії не всіма сприймається як сфера DM. З іншого боку, ми позначили там регресію як альтернативу класифікації. Чому? Припустимо, що всі незалежні атрибути в таблиці 1 неперервні, тобто атрибути з оцінками ЗНО залишені без змін, а атрибут *Рівень успішності за середнім балом атестату* замінений істинними числовими значеннями середнього балу атестату. Також припустимо, що залежною змінною також є неперервна величина, а саме, середній бал студента за підсумками навчання на першому курсі. Тепер нас може цікавити питання: як будуть навчатися на першому курсі студенти нового набору, якщо відомі їхні оцінки ЗНО і середній бал атестату?

Приклад відрізняється від попереднього тим, що всі змінні тепер числові і неперервні, і замість визначення класу для нового студента ми хочемо спрогнозувати його середню оцінку за 1 курс як точку на деякому відрізку числової прямої. Як бачимо, за ознакою можливого застосування задача регресії мало чим відрізняється від задачі класифікації. Однак у цьому випадку ми можемо скористатися класичним регресійним аналізом. Оскільки всі необхідні відомості можна знайти у будь-якому підручнику з математичної статистики, обмежимося тут лише деякими зауваженнями.

Форми представлення результатів і алгоритми. За даними навчаючої вибірки будується модель – лінія регресії. У більшості випадків залежність між залежною змінною і незалежними змінними добре моделюється прямою лінією, яка представляється простою формулою. Якщо позначити незалежні змінні як X_1, X_2,..., X_n, а залежну – як Y, то завдання полягає у відшуканні коефіцієнтів a_i і b у рівнянні

$$Y = a_1X_1 + a_2X_2 + \cdots + a_nX_n + b.$$

Усі популярні статистичні пакети і табличні процесори «уміють» це робити за даними навчаючої вибірки. Тепер, коли всі коефіцієнти знайдені, для прогнозування нового значення Y за відомими значеннями X_i достатньо підставити їх у рівняння і знайти результат. У нашому прикладі про студентів достатньо підставити в отримане рівняння дані про оцінки ЗНО і середній бал атестата нового студента, щоб спрогнозувати його успішність на першому курсі. Зауважимо, що прогноз швидше за все не справдиться, проте середній бал для групи студентів може бути прогнозований достатньо точно.

Коефіцієнти при незалежних змінних містять коефіцієнти кореляції, які вказують на так звану *прогностичну валідність* результатів ЗНО (і балу атестата).

Інше важливе застосування множинної регресії для даних, описаних у нашому прикладі, викладено у статті [2]: знайдені коефіцієнти регресії вказують на те, які *вагові коефіцієнти* для сертифікатів ЗНО й атестату слід підбирати університету для вступників на дану спеціальність, щоб ці дані якомога краще прогнозували успішність майбутніх студентів.

Зауважимо, що множинну регресію ні в якому разі не можна зводити до розгляду окремих одновимірних регресій (по одній для кожної незалежної змінної), оскільки тоді не будуть враховані ймовірнісні залежності між самими незалежними змінними. Результати множинної регресії завдяки врахуванню «внутрішніх» залежностей можуть виявитися досить несподіваними. Скажімо, цілком природно для вступу на поєднану спеціальність «математика і фізика» вимагати сертифікати ЗНО з математики і фізики, але після регресійного аналізу може виявитися, що краще прогнозує успішність студентів інша пара сертифікатів (наприклад, з математики і географії), оскільки між балами з математики і фізики, зазвичай, існує сильний кореляційний зв’язок, тобто інформація про оцінку з фізики є не такою вже й істотною, якщо вже відома оцінка з математики. Загалом, чим сильніший зв’язок між парою предикторів, тим менший ваговий коефіцієнт отримає один з них.

В арсеналі DM є дуже простий алгоритм, який дозволяє прогнозувати значення невідомої змінної за відомими значеннями інших змінних без побудови регресійної моделі – алгоритм *k середніх*. Значення залежної змінної обчислюється ним за відомими значеннями *найближчих сусідів*, зазвичай як просте середнє арифметичне. Цей алгоритм доречно використовувати у випадках, коли не вдається виявити оптимальну форму лінії регресії.

2.4. Задача пошуку асоціативних правил і секвенційний аналіз

Суть. Задачу пошуку асоціативних правил, зазвичай, пояснюють на прикладі з так званого *аналізу корзин* (*basket analysis*). Припустимо, що власники супермаркету з метою поліпшення обслуговування покупців (або, що не одне й те ж, з метою збільшення доходів) цікавляться тим, які товари часто продаються разом. Уводиться поняття транзакції, яке тут означає набір одночасно придбаних товарів – вміст однієї корзини покупця. Зрозуміло, кожен покупець бере різну кількість товарів, тобто в транзакцію може входити від одного до великої кількості товарів. Тут криється певна методологічна проблема, якщо ми хочемо знову уявляти собі базу даних у вигляді плоскої таблиці. Один з варіантів полягає у тому, що в базу даних вноситься щоразу по одному товару із зазначенням в окремому стовпці номера транзакції.

Що стосується прикладу застосування цього виду аналізу, то зазначимо тут, що та ж сама задача передбачення успішності студента може розв'язуватися за допомогою пошуку наборів даних з певної множини характеристик, які фіксуються для кожного студента, скажімо, базою даних і лог-файлами системи управління навчанням. Для цього потрібно значення незалежної змінної включити доипереліку «товарів у корзині». Здійснюючи після цього пошук асоціативних правил («частих» наборів характеристик), ми відповідаємо на запитання на зразок такого: «З якими характеристиками діяльності студента в системі управління навчанням одночасно часто зустрічається характеристика *Високий рівень успішності*»? Звертаємо увагу на те, що одночасно з цим ми з'ясовуємо, які характеристики діяльності студента є важливими для досягнення ним високого рівня, що повинно допомогти нам у поліпшенні самої системи.

Секвенційний аналіз отримуємо як похідну простого пошуку асоціативних правил, увівши відношення порядку до множин об'єктів. Часто таким природним відношенням є час. Тепер ми цікавимося, які *послідовності подій* відбуваються частіше за інші. Ми можемо піти й далі, увівши додатково відстань – проміжок часу між двома послідовними подіями.

Форми представлення результатів і алгоритми. Як не дивно, результати роботи алгоритму з пошуку асоціативних правил представляються у вигляді все тих же конструкцій *якщо…то…*

На перший погляд здається, що пошук «частих наборів» можна здійснити простим перебором. Насправді читач, знайомий з комбінаторикою, погодиться, що навіть у базі даних невеликого розміру перебір усіх підмножин об'єктів є справою надто затратною за часом. Популярний алгоритм *Apriori* ефективно звужує кількість наборів під час перебору, спираючись на простий принцип: слід розпочати з одноелементних наборів, потім переходити до двоелементних і т. д.; при цьому якщо якийсь набір зустрічається рідко, то з розгляду вилучаються всі більші набори, які містять даний набір, оскільки вони теж гарантовано зустрічатимуться рідко. Існують різні модифікації цього алгоритму, зокрема й для секвенційного аналізу.

2.5. Задача кластеризації

Суть. У задачі класифікації потрібно віднести об'єкт до одного з наперед заданих класів за навчаючою вибіркою. Кластеризація ж полягає у пошуку *природних класів* у множині об'єктів, тобто множина класів невідома. Наприклад, потрібно розбити множину студентів на групи схожих між собою студентів за інтересами або особливостями сприйняття ними навчального матеріалу, з метою індивідуалізації навчальних підходів і методик викладання. Це може покладатися також в основу

побудови адаптивних комп'ютерних навчальних систем. Часто кластеризація передує розв'язанню задачі класифікації.

Подібно до регресійного, кластерний аналіз виник ще до формування DM. Разом з тим, існують (і продовжують з'являтися) суто машинні алгоритми кластеризації. Кластеризація не вимагає наявності навчаючої вибірки у тому сенсі, що сама вибірка з повними даними про значення всіх атрибутів розбивається на кластери.

Приступаючи до розв'язання задачі кластеризації, користувач має вирішити чи не найважливіше питання: що для даних об'єктів вважати *схожістю*? Це може бути як один із видів відстаней (звичайна Евклідова, «манхеттенських кварталів», «Карлсруе», за однією з координат тощо), так і, скажімо, коефіцієнти кореляції і т. п. Також потрібно визначити, що буде вважатися центром кластеру, відстанню між кластерами, чи дозволяється одному об'єкту належати до різних кластерів. Деякі задачі можуть розглядатися в рамках *теорії нечітких множин*. Іншими словами, від користувача вимагається певний рівень попередньої теоретичної підготовки, а також глибоке розуміння природи об'єктів, що вивчаються.

Форми представлення результатів і алгоритми. Найчастіше використовуються агломеративні ієрархічні алгоритми, робота яких зводиться до послідовного укрупнення кластерів: спочатку кожен об'єкт множини вважається кластером, потім найближчі кластери об'єднуються в один і т. д. Алгоритм зупиняє роботу, коли відстань між кластерами є значимо більшою порівняно з розмірами утворених кластерів, або перевищує наперед заданий поріг. інформативною формою представлення результатів для ієрархічного алгоритму є *дендрограма* – графічна візуалізація послідовності утворення кластерів, на якій також видно відстані між кластерами. Утім, дендрограма часто замінюється спеціальною таблицею, у якій відображається та ж сама інформація.

Серед інших відзначимо елегантний алгоритм *k* середніх: (1) кількість кластерів задається наперед; (2) спочатку довільно призначаються центри кластерів; (3) потім з'ясовується, до якого з кластерів належить кожен об'єкт, тобто до якого з центрів він розташований найближче; (4) для утворених так кластерів наново обчислюються їхні центри; (5) усе повторюється з кроку 3. Алгоритм зупиняє роботу, коли координати центів кластерів перестають змінюватися. Основні недоліки алгоритму: кількість кластерів задається наперед; результати можуть відрізнятися залежно від початкових координат центрів.

До задач кластеризації умовно можна віднести задачі колаборативної фільтрації – пошуку (зокрема. в соціальних мережах) груп людей, близьких за інтересами чи будь-якими іншими характеристиками.

2.6. Деякі застосування

Деякі з можливих застосувань DM в освіті ми вже розглянули як ілюстрації до основних класів задач. В огляді [9] зміст 306 публікацій згрупований за 11 напрямами (називатимемо далі студентами будь-яких осіб, що навчаються).

1. Аналіз і візуалізація даних.
2. Забезпечення зворотного зв'язку між студентами іироблення рекомендацій для студентів.
3. Прогнозування успішності студентів.
4. Створення моделей студентів.
5. Виявлення несподіваних проявів у поведінці студентів.
6. Групування студентів.
7. Аналіз соціальних мереж.

8. Конструювання ментальних карт (concept maps).
9. Конструювання систем управління курсами.
10. Планування навчального процесу.

Перші 4 напрями досліджень за кількістю проаналізованих авторами огляду станом на 2009 рік публікацій виявилися найбільш популярними. Зауважимо, що перший напрям може лише умовно належати до EDM, оскільки передбачає лише певну попередню обробку даних у вигляді різних статистичних характеристик, таблиць і діаграм частот тощо. Цікаво відзначити, що раніше саме цей напрям переважав за кількістю публікацій, тоді як на час написання огляду помітно почала переважати тема *прогнозування*.

Оскільки огляд є у вільному доступі в Інтернеті, читачеві рекомендується самостійно ознайомитися з тематикою конкретних публікацій. Ми ж обмежимося відзначенням лише деяких цікавих, на наш погляд, тем досліджень, користуючись базовою класифікацією задач DM, викладеною в підрозділі 2.3.

На перший погляд, кожен з напрямів досліджень 1–11 припускає розв'язання однієї з базових задач DM. Але це далеко не так. У більшості досліджень розв'язується певна комбінація базових задач з використанням різних алгоритмів. Розглянемо, для прикладу, напрям *групування студентів*. Тут у межах одного дослідження можуть розв'язуватися подані нижче проблеми.

1. Як розбити всіх студентів на групи згідно з рішенням, яке приймається (наприклад, за індивідуальними навчальними стилями); хто із студентів найкраще підходить як партнер для конкретного студента з метою виконання певного колективного проекту (*кластеризація*, *колаборативна фільтрація*).

2. До якої з утворених груп помістити нового студента; яким буде його успішність (*класифікація*, *регресія*).

3. Якими є особливості даної групи студентів; як правильно спланувати навчальне навантаження студентів (*пошук асоціативних правил*, *секвенційний аналіз*).

З наведених раніше прикладів можна дійти висновку, що *класифікація* і *регресія* ефективно використовуються для *прогнозування* різного роду освітніх (у тому числі навчальних) показників. Також, крім очевидної за своєю значимістю задачі прогнозування рівня навчальних досягнень студентів, слід відзначити такі цікаві застосування цього виду EDM: аналіз лог-файлів і баз даних систем електронного навчання з метою поліпшення процесу навчання і контенту навчальних матеріалів; відбір студентів до спеціалізованих груп (за інтересами, за здібностями, за особливостями сприйняття інформації, стилями навчання); поліпшення якості комп'ютерних тестів (наприклад: які завдання тесту найкраще визначають рівень успішності?); відбір тих контекстуальних змінних, які найбільше впливають на якість навчання, побудова на основі цієї інформації лаконічної але ефективної системи забезпечення якості; побудова ментальної моделі студента в середовищі електронного навчання; виявлення нетипових рис у поведінці студента або й викладача (низька вмотивованість, схильність до обману, зловживання тощо); класифікація веб-документів.

Практично всі з перерахованих вище тем досліджень можуть також передбачати використання алгоритмів *пошуку асоціативних правил* (та й *кластеризації*). Відмінності, як ми переконалися раніше, полягають у деталях мети дослідника. Більш показовими у сенсі використання пошуку частих наборів об'єктів є автоматизація створення ментальних карт; планування оптимального індивідуального плану студента, розкладу занять, розподілу ресурсів; секвенційний аналіз допоможе правильно спланувати все це у часі.

Кластерний аналіз може передувати процесу класифікації у випадках, коли класи об'єктів (чи то учасників навчального процесу, чи то навчальних матеріалів, методів, подій тощо) не задані наперед жорстко. Зауважимо, що умоглядна класифікація, наприклад, розподіл студентів за рівнями «просунутості» згідно з якимись наперед заданими критеріями може виявитися далекою від реального стану речей. Кластерний аналіз допоможе це виявити і запропонувати нове бачення критеріїв. Також звертаємо увагу читачів на важливість з точки зору застосувань *колаборативної фільтрації*, яку можна вважати «полегшеним варіантом» кластеризації, оскільки тут нам важливо не так розбити всю множину об'єктів на класи, як підшукати для якогось конкретного об'єкта певну множину найбільш схожих на нього об'єктів.

Програмне забезпечення. Як уже зазначалося, окремі модулі DM входять нині до пакетів практично всіх популярних статистичних програм, таких як SPSS, для розв'язання окремих задач достатньо табличних процесорів. Існують і спеціалізовані програми, з яких хотілося б виділити дві: це Deductor Studio (http://basegroup.ru/deductor/description) з доброзичливим російськомовним інтерфейсом користувача, якісною документацією, супроводжуваною прикладами, і безкоштовною, щоправда, обмеженою у кількості досліджуваних об'єктів академічною версією, яка часто використовується у викладанні для студентів IT спеціальностей; WEKA (http://www.cs.waikato.ac.nz/ml/weka/) – безкоштовний продукт авторів книги [10], який розповсюджується за ліцензією GNU, містить реалізацію великої кількості алгоритмів DM.

3. ВИСНОВКИ ТА ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

У наш час найбільш цікаві і перспективні дослідження проводяться на перехресті різних наук, і DM є чудовим свідченням того, як суто комп'ютерні технології можуть слугувати задачі аналізу даних у різних галузях, у тому числі й освіті. Через мультидисциплінарність досліджень зростає потреба в ефективній організації роботи дослідницьких груп, членами яких є представники різних окремих галузей. При цьому є очевидною необхідність певного рівня обізнаності кожного члена колективу із специфікою роботи колег. Освітянину, який хоче досягти прогресу у поліпшенні якогось з аспектів якості освіти, і розуміє необхідність застосування методів і технологій аналізу даних, але який разом з тим не є фахівцем у цій галузі, доцільно звернутися за допомогою до фахового аналітика. Але для цього він повинен принаймні грамотно сформулювати суть задачі. Завданням даної статті є популяризація напряму EDM якраз на такому рівні. З цією метою розглянуто як базові задачі DM, так і принципи роботи основних алгоритмів, на прикладах їх застосування для досліджень в галузі освіти, й окреслено сучасний стан EDM.

Що стосується перспектив, то в першу чергу слід зазначити зростання застосування Байєсівських мереж, зокрема, в галузі освітніх вимірювань, у рамках нової парадигми ECD (Evidence-centered Design) [3], яка допускає значно більш вільне трактування тесту як інструменту вимірювання, послаблюючи вимоги до його стандартизації і застосування жорстких моделей класичної психометрії, і є безумовно орієнтованою на комп'ютерні технології.

Багатообіцяючою перспективою є повна інтеграція DM до систем електронного навчання, подібно до того. як це відбувається у пошукових системах (Google), різноманітних веб-системах вироблення рекомендацій, усього того, що часто позначають як Web 2.0.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Барсегян А. А. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP: уч. пос. / А. А. Барсегян, М. С. Куприянов, В. В. Степаненко, И. И. Холод. – СПб. : БХВ-Петербург, 2007. – 384 с.
2. Ковальчук Ю. О. Відбір до ВНЗ як прикладна задача теорії освітніх вимірювань / Ковальчук Ю. О., Лісова Т. В. // Вища освіта України. Тематичний випуск «Європейська інтеграція вищої освіти України в контексті Болонського процесу». – 2014. – №3 (додаток 1). – С. 69–73.
3. Almond R. G. Bayesian Networks in Educational Assessment. / Almond R. G., Mislevy R. J. at all. – Springer, 2015. – 666 p.
4. Baker R. The State of Educational Data Mining in 2009: A Review and Future visions. / Baker R., Yacef K. // Journal of Educational Data Mining. – 2009. – Vol 1. – No 1.— Pp. 3–17.
5. Korb K. B. Bayesian artificial intelligence / Kevin B. Korb, Ann E. Nicholson. p. cm. – Chapman & Hall/CRC computer science and data analysis, 2003. – 365 p.
6. Romero C. Handbook of Educational Data Mining. / by Cristobal Romero (Editor), Sebastian Ventura (Editor), Mykola Pechenizkiy (Editor), Ryan S.J.d. Baker (Editor). – CRC Press, 2010. – 536 p.
7. Romero C. Data mining in e-learning (Advances in Management Information). / Romero C., Ventura S. – Wit Press, 2006. – 328 p.
8. Romero C. Educational Data Mining: a Survey from 1995 to 2005. / Romero, C., Ventura, S. // Expert Systems with Applications. – 2007. – No 1. – Vol 33. – P. 135–146.
9. Romero C. Educational data mining: a review of the state of the art. / Romero, C., Ventura, S. // IEEE Transactions on Systems, Man, And Cybernetics – Part C: Applications And Reviews. – 2010. – Vol. 40. – No. 6. —Pp. 601–618.
10. Witten I. H. Data Mining: practical machine learning tools and techniques. – 3rd ed. / Ian H. Witten, Frank Eibe, Mark A. Hall. p. cm. – (The Morgan Kaufmann series in data management systems), 2011. – 665 p.

Матеріал надійшов до редакції 29.09.2015 р.

ПОИСК, ИЗВЛЕЧЕНИЕ И АНАЛИЗ ДАННЫХ В ОБРАЗОВАНИИ: СОВРЕМЕННОЕ СОСТОЯНИЕ И ПЕРСПЕКТИВЫ РАЗВИТИЯ

Ковальчук Юрий Алексеевич
доцент, кандидат физико-математических наук, декан факультета
Нежинский государственный университет имени Николая Гоголя, г. Нежин, Украина
yu.kovalchuk@i.ua

Аннотация. Рассмотрены основные задачи (классификация и регрессия, поиск ассоциативных правил, кластеризация) и принципы работы базовых алгоритмов Data Mining в контексте их использования для различных исследований в области образования, являющихся предметом относительно нового самостоятельного направления Educational Data Mining. Приведены данные о наиболее популярных темах исследований в рамках данного направления, обозначены перспективы его развития. Статья рассчитана на читателей, которые занимаются исследованиями в области образования на различных уровнях, особенно тех, кто причастен к использованию систем электронного обучения, но мало знаком с этим направлением анализа данных.

Ключевые слова: Educational Data Mining; классификация; регрессия; поиск ассоциативных правил; кластеризация; система электронного обучения.

DATA MINING IN EDUCATION: CURRENT STATE AND PERSPECTIVES OF DEVELOPMENT

Yurii O. Kovalchuk
Docent, PhD (Mathematics and Physics), Dean of Faculty
Nizhyn Mykola Gogol State University, Nizhyn, Ukraine
yu.kovalchuk@i.ua

Abstract. The main tasks (classification and regression, association rules, clustering) and the basic principles of the Data Mining algorithms in the context of their use for a variety of research in the field of education which are the subject of a relatively new independent direction Educational Data Mining are considered. The findings about the most popular topics of research within this area as well as the perspectives of its development are presented. Presentation of the material is illustrated by simple examples. This article is intended for readers who are engaged in research in the field of education at various levels, especially those involved in the use of e-learning systems, but little familiar with this area of data analysis.

Keywords: Educational Data Mining; classification; regression; association rules; clustering; e-learning system.

REFERENCES (TRANSLATED AND TRANSLITERATED)

1. Barsegian A. A. Data Analysis Technology: Data Mining, Visual Mining, Text Mining, OLAP: tutorial. / A. A. Barsegian, M. S. Kupriyanov, V. V. Stepanenko, I. I. Holod. – SPb. : BHV-Petersburg, 2007. – 384 p. (in Russian).
2. Kovalchuk Y. O. University admission as an applied problem of the theory of educational measurement / Kovalchuk Y. O., Lisova T. V. // Higher Education of Ukraine. Special issue “European Integration of Higher Education of Ukraine in the context of the Bologna process.” – 2014. – №3 (annexe 1). ––P. 69–73 (in Ukrainian).
3. Almond R. G. Bayesian Networks in Educational Assessment. / Almond R. G., Mislevy R. J. at all. –– Springer, 2015. – 666 p. (in English).
4. Baker R. The State of Educational Data Mining in 2009: A Review and Future visions. / Baker R., Yacef, K. // Journal of Educational Data Mining. – 2009. – Vol 1. – No 1.— Pp. 3–17 (in English).
5. Korb K. B. Bayesian artificial intelligence / Kevin B. Korb, Ann E. Nicholson. p. cm. – Chapman & Hall/CRC computer science and data analysis, 2003. – 365 p. (in English).
6. Romero C. Handbook of Educational Data Mining. / by Cristobal Romero (Editor), Sebastian Ventura (Editor), Mykola Pechenizkiy (Editor), Ryan S.J.d. Baker (Editor). – CRC Press, 2010. – 536 p. (in English).
7. Romero C. Data mining in e-learning (Advances in Management Information). / Romero C., Ventura S. – Wit Press, 2006. – 328 p. (in English).
8. Romero C. Educational Data Mining: a Survey from 1995 to 2005. / Romero, C., Ventura, S. // Expert Systems with Applications. – 2007. – No 1. – Vol 33. – Pp. 135–146 (in English).
9. Romero C. Educational data mining: a review of the state of the art. / Romero, C., Ventura, S. // IEEE Transactions on Systems, Man, And Cybernetics – Part C: Applications And Reviews. – 2010. – Vol. 40. – No. 6. – Pp. 601–618 (in English).
10. Witten I. H. Data mining : practical machine learning tools and techniques. – 3rd ed. / Ian H. Witten, Frank Eibe, Mrk A. Hall. p. cm. – (The Morgan Kaufmann series in data management systems), 2011— 665 p. (in English).