ИЗВЛИЧАНЕ НА ЗНАНИЯ ОТ ДАННИ В ОБРАЗОВАТЕЛНО ПРОСТРАНСТВО

Даниела Орозова
Бургаски свободен университет

DATA MINING IN EDUCATIONAL SPACE

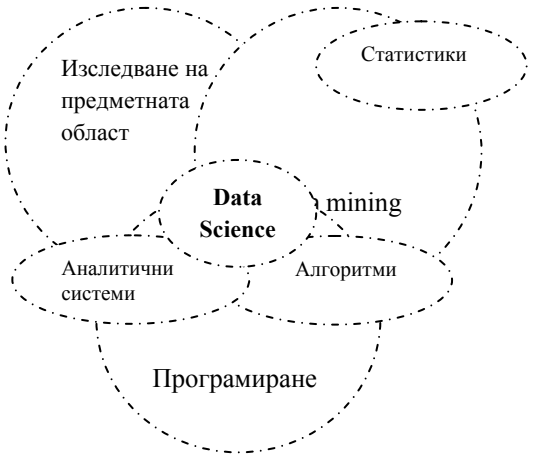
Daniela Orozova
Burgas Free University

Abstract: *Системата за електронно обучение на БСУ, базирана на Мудъл, позволява надграждане и развитие с цел провеждане на качествено персонализирано и адаптивно е-обучение. Разгледани са особености на базата от данни на тази система, примерни справки и съответните им SQL заявки. Поради нарастващия обем на натрупваните данни се предлага използване на агенти за извличане на знания от данните.*

Key words: *Big Data, Education Space, Agent oriented architectures, E-learning, Data Science.*

1. Увод

Съвременните тенденции в компютърните системи са свързани с развитието на високопроизводителни изчисления (HPC) и работа с феноменна големи данни (BigData) [16]. В тази статия интересът е насочени към оптимизиране на достъпа до големи обеми от данни в съвременно оразователно продтранство, чрез прилагане на агенти за извличане на знания от данните. Обработката и съхранението на големите данни изисква нов поглед и съвместно прилагане на редица утвърдени технологии, както е показано на фигура 1 [18]. Появява се понятието *Data Science*, което се свързва с извличането на предварително неизвестни знания директно от данни, чрез процес на откриване и аналитичен анализ на хипотези.

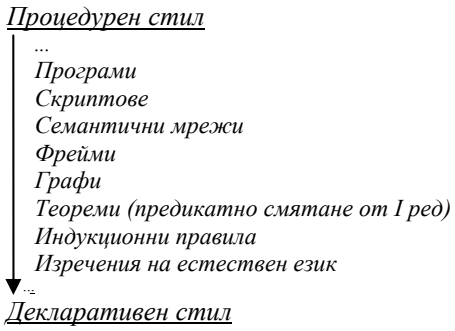


Фиг. 1. Възникване на понятието Data Science

Когато говорим за извличане на знания, данните и знанията трябва да бъдат формализирани – представени чрез определен подход. На фигура 2 се показват различни формализми за представяне на знания, започвайки от процедурен формат, който е добре структуриран и преминавайки към декларативен формат, по-свободен в описанието.

Процедурният подход позовава простотата и лекотата на разбиране. Представянето е чрез алгоритми, симулиращи реално поведение. Този тип позволява проследяване на проблем от алгоритмичен стил и е лесен за анализиране.

Декларативният подход е по-свободен и гъвкав. Тук структурата на управление е отделена от представените знания, обикновено като система от правила. Основната характеристика при това представяне е модулност.



Фиг. 2. Формализми за представяне на знания

Начинът на представяне на знанията е един от най-важните проблеми, които трябва да бъдат решени при проектирането на обучаваща среда. Знанията за предметната област са основата, върху която се гради експертността на системата. Те фактически представят познанията на самата система и в този смисъл трябва да бъдат пълни и удобни за използване. Това условие отговаря на изискването към преподавателя да бъде експерт в областта на познанието, което преподава. За попълването на знанията от предметната област в системата се грижат специалисти от областта и педагози.

2. Архитектура на образователното пространство

Съвременното образователно пространство трябва да осигурява контекстно-зависима, адаптивна и персонализирана доставка на образователни услуги и учебно съдържание. Пространството е множество от различни активни компоненти, опериращи в информационната среда.

Една *информационна среда* включва (пасивни) информационни ресурси, необходими за организиране, провеждане и документиране на електронен обучителен процес. Тя може да се разглежда като околната среда, в която оперират и до която имат достъп интелигентни агенти от различни типове. Изграждането образователно пространство включва три типа компоненти [14]:

- *Дигитални библиотеки* – хранилища на електронно учебно съдържание;
- *Онтологии* – представят основните концепции и релациите между тях за предлаганите учебни дисциплини;
- *Бази от данни* – релационни хранилища с данни и извеждана от тях информация, необходима за функциониране на образователното пространство.

Развитието на технологиите дава възможност да се прилагат модели на иновативно преподаване, обучение и оценяване, чрез създаване на обучителни среди, кои-

то могат да интегрират в обучението използването на различни ресурси, дейности и подходи. Облачно-базираните технологии, социалните мрежи и мобилните устройства все повече стават част от нашето ежедневие. Все по-често се използват електронни форми за подпомагане на обучението, като тенденцията е налагане по-силно на мобилното обучение (м-обучение). М-обучението представлява електронно обучение (е-обучение), чрез използване на различните видове мобилни устройства и свързаните с тях технологии за подпомагане процеса на обучение [7]. В сферата на образованието, ако няколко университета изградят общ облак, те ще могат да предложат сравними по качество и обем информационни услуги, което ще се отрази положително на качеството на самото обучение.

През 2012 г. по Оперативна програма „Разлитие на човешките ресурси“, Бургаският свободен университет спечели проект „Университетски център за електронни форми на дистанционно обучение и услуги при БСУ (УЦДО) – възможност за учене през целия живот”. В рамките на проекта бяха проведени проучвания и анализи и бе взето решение в качеството на система за създаване, ползване и поддържане на електронно учебно съдържание в БСУ да се използва системата за управление на е-обучение с отворен код Мудъл [8]. Тази система позволява надграждане и развитие с цел провеждане на качествено персонализирано и адаптивно е-обучение и интеграция с други средства за добавяне на нови функционалности. По време на работата на системата през годините се напрупват данни, свързани с обученеито на различните студенти по изучавани от тях дисциплини. Последващите анализи на натрупваните през годините данни са интересно поле за научни изследвания. Нашите усилия са насочени в посока на интегриране на *Data Mining* техники със системата за електронно обучение.

3. Задачата за моделиране на потребителя в образователното пространство

Чрез подходящи средства за следене и отчитане на дейността на потребителя се изгражда негов модел. Моделът на потребителя служи за персонализиране на системата към знанията и уменията на потребителя и прилагане на адекватни подпомагащи стратегии в работата му.

Моделът на потребителя по своята същност е структура от данни, представяща индивидуални характеристики за потребителя, на дадена програмна система. Той представя познавателните процеси и схващания на потребителя за приложната област. Сложният процес на създаване, поддържане и обновяване на модела, въз основа на наблюдаваното поведение, се нарича *диагностициране* [10].

Моделирането на потребителя дава възможност за осъществяване на следните дейности от страна на системата:

- адаптиране на поведението на системата към индивидуалния потребител за подпомагане на научно-приложната дейност;
- определяне на грешни схващания на потребителя и вземане на мерки от страна на системата за отстраняването им;
- възможност за активност на системата и осъществяване на смесена инициатива на диалога: напредване в учебната програма; преход към подходяща или следваща тема, когато текущата е усвоена; предложение за помощ в най-подходящите моменти; генериране на нови задачи с трудност малко над оценките за текущите знания и умения на обучаемия; адаптиране на обясненията и др.

В процеса на проектиране на модела на потребителя трябва да се отговори на следните въпроси:

Защо се моделира потребителя? – каква е общата цел, кои модули на системата се нуждаят от данните, съдържащи се в модела на потребителя.

Кой се моделира? – какво е отношението на моделираните хора към системата, колко индивидуални са моделите – какви и колко са потребителите, които се моделират. Често се прави модел на типове потребители или базов модел на потребител.

Какво се моделира? – какви аспекти на потребителя се моделират и как се интерпретира моделът. Някои от най-често моделираните аспекти на потребителя са: ниво на експертност, убеждения, грешни предположения, цели, планове или намерения, квалификация и др.

Как се моделира потребителя? – това са методологията и източниците за моделиране. Използват се общи методи като: абстрахиране, класификация или развитие на прагматичен модел. Част от информацията е заложена предварително в модела, друга информация се получава от отговори на потребителя или се извежда индиректно от отговорите му, трета информация се получава от предишни сесии на системата.

„Моделът на потребителя е източник на знания, съдържащ в явен вид информация за някои специфични характеристики на потребителя, засягащи диалога със системата" [10]. В интелигентните програмни системи се създава отделен модул, който строи модела и снабдява другите модули на системата с информация и предположения за потребителя. Информацията съдържаща се в модела не е статична. Тя се изменя и конкретизира при действията на потребителя. Този модел съществува и има смисъл само, ако информацията от него се използва активно от системата за генериране на различни действия, относно различните потребители.

Източниците за получаване на информация, необходима за изграждане на модела на потребителя могат да бъдат:

- неявни – от поведението на потребителя в конкретната ситуация;
- явни – от диалога на системата с потребителя;
- структурни – от вътрешните връзки между обектите и познанията в предметната област;
- фонови – базиращи се на оценки за средното ниво на подготвеност на потребителя.

Проучват се различни техники и средства за следене и отчитане на дейността на потребителя и подходи за моделирането му. То се разглежда в два аспекта:

а) Моделиране на знанията и уменията на цял клас от потребители.

Създават се групи (класове) от потребители с общи характерисики, на които се предлагат общи функционалности и ресурси. Цели се, системата да е така проектирана, че да задоволява техните колективни нужди. Тази идея стои в основата на стереотипният подход, при който потребителя се класифицира към някоя предварително дефинирана група и по този начин му се приписват общите за групата характеристики. Стереотипите могат да се представят във вид на йерархия и могат да разширяват тези, които са над тях. Когато обаче се изисква индивидуална адаптация или предоставяне на специфична помощ, стереотипните модели не са достатъчни.

б) Моделиране на знанията и уменията на отделен потребител

Индивидуалният модел на потребителя служи за описване на историята на работата, за фиксиране на текущото състояние на знанията, за прогнозиране развитието на уменията на потребителя и разкриване на погрешните му заключения.

Основни области на приложение на моделиране на колективен или индивидуален потребител са интелигентните системи за обучение, експертните системи, програмните потребителски среди и др.

Според Barr и Feigenbaum [11] съществуват два основни подхода за моделиране на отделен обучаем (потребител):

- Моделиране с припокриване

При моделите с припокриване (overlay approach) знанието на потребителя се представя като подмножество на общото знание, поддържано от системата. Знанието на системата може да бъде определено от предметното знание за областта, която се изучава, знанието на експерта в областта или очакваното знание за обучаемия. Моделът с припокриване е сред най-често използваните модели. Знанието на системата е декомпозирано на независими компоненти и е покрито със система от означения, показващи нивото на усвояване на всяка отделна компонента. Обикновено тези компоненти са представени като йерархия или семантична мрежа от възли, свързани пряко с понятията от областта. За оценяване на знанията се използват логически или числови стойности. При една от разновидностите на този вид моделиране – диференциалните модели – системата поддържа множество от знания, които трябва да бъдат усвоени и друго множество от знания, които не е задължително да бъдат усвоени. Моделите с припокриване могат да оценяват знанието на потребителя в различни предметни области, което прави тези системите гъвкави и адаптивни.

- Моделиране с библиотека от грешки

При този подход (bugs theory) формално се диагностицират грешки, чрез списък от предварително дефинирани грешно усвоени и липсващи елементи от знания. Тук моделът на потребителя се състои от модел на експерта, допълнен със списък от допускани грешки в областта.

Моделиране на някои психологически качества като упоритост, работоспособност, съсредоточеност, време за което потребителят забравя определено знание или умение, ако не го упражнява е представено в [12]. Една от новите насоки е способността за моделиране на индуктивни или дедуктивни разсъждения, изучаване и идентифициране на креативното мислене и действие на обучаемите. Такъв модел се реализира при средата Creativity Assistant, разработена в Де Монтфорт университета, Лестър [ELDE]. В [13] е представена програмна среда за събиране и съхраняване информация за настроенията и действията на обучаемите в процеса на усвояване и прилагане на знания. Събраните данни се представят в структури, наречени „карти на креативност" (Creativity Maps). Разработва се концепция за откриване и използване на подходящи анализатори на картите на креативността.

В предметно-съдържателен план индивидуализацията на програмната система позволява да се осъществява вариативност на постъпващите към потребителя съобщения във функции на обяснения по степен на тяхната детайлизираност, насоченост, честота на включване, ориентация на обучаващия и управляващ ефект, което се определя от конкретните характеристики на когнитивния стил на потребителя. В резултат на това може да се извършва предварително планиране на стратегиите за търсене, диференциране на подмножество от оценъчни съобщения и управляващи въздействия.

4. Базата от данни на Мудъл

Използването на софтуер с отворен код си има своите предимства и недостатъци. Предимствата в общия случай са ниската цена за инсталиране, внедряване и експлоатиране на продукта и наличието на богата гама от функционалности. Недостатъците (от програмистка гледна точка) са свързани с редица трудности като необходимостта от проучване на голям обем сорс код при надграждане на системата и усложнена бизнес логика, породена от стремежа за постигане на универсалност на

функционалностите и др. Базата от данни на Мудъл съдъжа над 200 таблици [5]. Основните таблици са 50, а останалите са групирани на база функционалности, като името на всяка таблица съдържа името на модула, който обслужва.

Използвайки натрупаните данни, могат да се генерират редица полезни справки за участниците и работата им, например: справка за участниците в дадена дейност, справка за продължителност на участие в дадено обучение, дата на започване и дата на завършване на участието; данни за обучаемите и курсовете, в които участват; общ брой участници; брой курсове по специалности; брой участници, преминали обучение в отделните курсове и др. В Мудъл не съществуват стандартни средства за създаване на посочените справки. За тяхното изготвяне е необходимо познаване не само на структурата, но и на съдържанието на таблиците с данни за потребителите. Примерните SQL кодове са написани на MS SQLServer, но могат да се ползват и други системи за управление на бази от данни. Основни таблици от базата данни на Мудъл са:

- Таблица **role** съдържа информация за потребителските роли в системата. Ролите се създават от администратор на системата и съответно, конкретните им имена могат да се различават във всяка отделна инсталация на Мудъл.

- Таблица **user** описва потребителите на системата, с техните основни характеристики: име (*firstname*), фамилия (*lastname*), адрес (*address*), e-mail и др.

Ако са необходими допълнителни характеристики за потребителите, може да се създаде допълнителна таблица, която да се свързва с таблицата *user* по номера на потребителя. Ако са необходими данни за допълнителни характеристики, които не се съхраняват по подразбиране в таблиците на Мудъл, можем да създадем нова таблица, наптимер *user_info*, която да представя конкретни стойности за допълнителните данни, за потребител [2]. Всеки отделен запис от ще таблицата съдържа информация за това, кое допълнително поле (*filedid*), за кой потребител (*userid*), каква стойност има (*data*). Тогава, за да се изведат данните за един потребител, за всяка от допълнителните характеристики трябва да се направи вложена заявка.

Таблица **course** съдържа основни данни за курсовете като кратко и пълно име на курса (*shortname* и *fullname*) и неговите характеристики. В допълнение всеки курс има зададена категория – полето *category,* което изпълнява ролята на външен ключ и служи за връзка с първичния ключ на допълнителна таблицата **course_categories**. В тази таблица са зададени като категории основните специалности (направления).

В таблица **context** се задават различни контекстно зависими характеристики. В Мудъл има дефинирани като константи няколко нива на контекст (context level, описани във файла lib/accesslib.php). В зависимост от стойността на *contextlevel*, полето *instanceid* се свързва с различни таблици. Например, ако *contextlevel* е 50, то *instanceid* съдържа *id* на курс от таблицата course.

Таблица **role_assignments** описва кой потребител (*userid*) каква роля (*roleid*) в какъв контекст (*contextid*) притежава. Например, един потребител може да е с роля студент за даден курс и с роля автор за друг.

Изплзвайки тези таблици, например заявка за извеждане на списък с имената на курсовете и специалността, в която се чете курса има вида:

```
SELECT course_categories.name, course.fullname, course.shortname
FROM course_categories LEFT JOIN course
ON course_categories.id=course.category
ORDER BY course_categories.sortorder
```

Пример на заявка за групиране на курсовете по специалности и извеждане на броя на създадените курсове за всяка специаллност има вида:

```
SELECT course_categories.name, COUNT(*) AS num
FROM course, course_categories
WHERE course_categories.id=course.category
GROUP BY course_categories.name
ORDER BY num
```

Алтернативна SQL-заявка със същото условие има вида:

```
SELECT course_categories.name,
        (SELECT COUNT(*)
         FROM course
         where course_categories.id=course.category) AS num
FROM course_categories
ORDER BY num
```

Следващо развитие е свързано с автоматизирано изследване и оценяване на качеството на провежданите електронни и дистанционни форми на обучение на базата на справки за индикатори, извличани от БД на използваната система за електронно обучение.

5. Извличане на знания от данните

Процесът Data Mining (в превод на български – известен като *Извличане на закономерности от данни)* има за цел извличането на имплицитна, непозната отнапред и потенциално полезна информация от файлове, бази от данни и други източници на данни. Изведените от данни зависимости и обобщения се наричат *модели, образци* или *шаблони*, които се представят чрез линейни уравнения, правила, клъстери, графи, дървета и т.н [8].

Това е процесът на извличане на валидна и предварително неизвестна информация от различни хранилища с данни и използването ѝ при взимане на решения. Data Mining също се разглежда като процес на откриване на шаблони, корелации и тенденции в големи обеми от данни с помощта на статистически и математически средства и методи на изкуствения интелект [8]. Използва методи от различни области като: статистика, невронни мрежи, машинно обучение, генетични алгоритми и др. Този процес е ефективен при наличие на обобщени данни, каквито се съхраняват в складовете от данни (Data Warehouse). При работата с малки множества от данни могат да бъдат използвани методи на класическия изследователски анализ, прилаган от статистиците. В случая на големи бази от данни възникват нови проблеми. Някои от тях са свързани с начина за ефективно съхраняване и намиране на данни, а други се отнасят към такива фундаментални въпроси, като: как да бъдат избрани характерни представители на данни, как тези данни могат да бъдат анализирани за приемливо време, как да се разбере, дали някоя намерена зависимост отразява действителната реалност, а не е резултат от случайно съвпадение в определена част от данните.

По принцип *data mining* техники биха могли да се прилагат върху различни информационни хранилища като: релационни бази от данни, складове с данни, транзакционни бази от данни (OLTP), обектно-ориентирани и обектно-релационни БД, БД за специфични приложения (пространствени БД, БД с временни редове, текстови БД и мултимедийни БД), файлове с данни в текстов или бинарен формат и World-Wide Web. Успешно могат да бъдат прилагани и при хранилища с данни, свързани с електронни системи и среди за обучение.

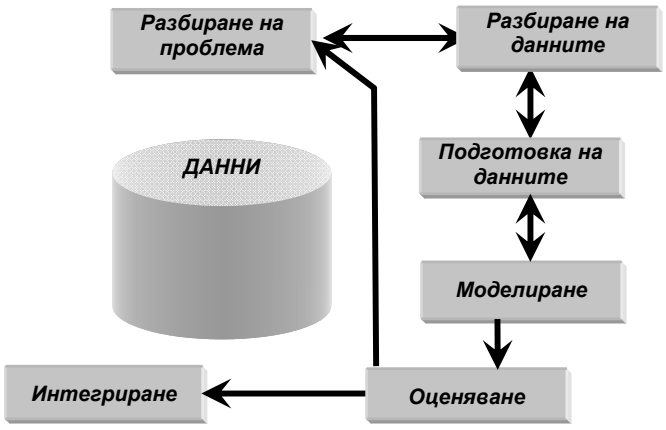
Процесът на извличане на закономерности от данни се състои от шест основни етапа [9]:

(1). Разбиране на проблемната област (Business understanding) – фокусира върху дефинирането на целите на изследванията и съответните изисквания от гледната точка на потребителя. Тук изследователят трябва да разбере, какво точно иска от него потребителя. Често потребителят има множество конкуриращи се цели и ограничения, които трябва да бъдат балансирани по подходящ начин. След завършването на етапа, тези знания трябва да бъдат превърнати в дефиниции на задачи за сондиране на данни и да се състави предварителен план как тези цели могат да бъдат постигнати.

(2). Разбиране на данните (Data understanding) – започва с първоначално събиране на данни и продължава с дейности, целящи задълбочаване на знанията на изследователя за естеството на данните. На този етап е необходимо да бъдат идентифицирани проблеми, свързани с качеството на данните, да бъде получено първоначално мнение за характера на данните, да бъдат намерени интересните подмножества на данните, за да се формират първоначални хипотези за скритата в данните информация. На този етап трябва да бъдат намерени отговори на такива въпроси като: дали данните са пълни? Дали те са коректни или съдържат грешки? Има ли липсващите данни и ако е така, как те са представени?

(3). Подготовката на данните (Data preparation) – покрива всички дейности по създаване от първоначални „сурови“ данни на крайното множество от данни (т.е. данни, които ще бъдат използвани от моделиращите средства). Етапът на подготовката на данни често се налага да бъде прилаган многократно и по различно време. Задачите по подготовката на данни включват в себе си избор на таблиците с данни, техни атрибути и отделни записи, както и трансформация и изчистване на данни.

Различието между непрекъснати и символни променливи е важно, тъй като някои от техниките за анализ на данни, подходящи за един тип променливи, не са подходящи за други. Непрекъснатите променливи се измерват по числова скала и по принцип могат да приемат произволни числови стойности. Символните променливи могат да приемат само определени, дискретни стойности. Символните стойности могат да бъдат *подредени* (т.е. да имат естествен начин на подреждане – например, *степен на образование*), или *номинални* (т.е. представят само имена на определени категории, като, например, *семейно положение*).



Фигура 3. Схема на процеса Data Mining според CRISP-DM [17] (Cross-Industry Standard Process Model for Data Mining)

(4). Моделиране (Modeling) – този етап се състои от избор и прилагане на различни техники за моделиране, целящи извличане на закономерности от данните. Параметрите на моделите се калибрират до свои оптимални стойности. Тъй като някои модели имат свои специфични изисквания към формата на данните, на този етап често се налага връщане към етапа за подготовката на данни.

(5). Оценка на модела (Model evaluation) – на тази стъпка се оценява степента, до която моделът отговаря на целите, поставени от потребителя и се прави опит да се определи, дали съществуват някакви причини от проблемната област, правещи модела неприложим. Резултатите трябва да бъдат сравнени с оценъчният критерий, дефиниран в началото на изследванията. Етапът се състои във внимателно преглеждане на всички стъпки, изпълнени при създаването на конкретния модел, за да се потвърди, че те постигат поставените цели. В края на този етап се приема решение за използване или не на получените резултати.

(6). Експлоатация на модела (Deployment) – свързана е с необходимост от наблюдение и изготвяне на стратегия за експлоатация. На този етап следва да се определи дали и кога да се поднови процедурата по извличане на знания от данни и при какви условия.

Огромното количество събрани данни значително превишава възможността на човека те да бъдат ефективно използвани без помощта на специализирани мощни средства за анализ на данни. Стига се до въпроса как да се избегне вече известната ситуация „богата на данни, но бедна на информацията".

Основен акцент в процеса е прилагането на съответен Data Mining алгоритъм или алгоритми, позволяващи получаването на знания, описващи: връзка между свойства на данните, модели на данните, резултатите от класификацията и клъстеризацията на данните и др. Интересно е какви типове зависимостите могат да бъдат открити в процеса на извличане на закономерности от данни. За целта разглеждаме различни типове задачи, решаването, на които води до получаване на интересуващите ни зависимости:

- **Задачата за *описание и обобщение на данни*** (Data Description and Summarisation) цели намирането на описание на основните характеристики на данни, обикновено в обобщена форма. Това позволява на потребителят да получи представа за структурата на данните. В повечето от случаите описанието и обобщението на данни е само една подзадача, изпълнявана на ранните стадии на едно изследване. Началният изследователски анализ на данните може да помогне да бъде разбрана природата на данните, да бъдат построени първоначалните хипотези за скрита в данните информация. Резултатите от описанието или обобщението на данните могат да бъдат представени във вид на графики, диаграми, таблици, правила и др.
- **Задачата за *сегментация (или клъстеризация)*** има за цел разбиването на данни на интересни и смислени подгрупи или клъстери. Всички членове на една подгрупа поделят общите характеристики. Клъстеризацията може да се използва и за създаване на таксономии, т.е. организацията на обекти в йерархия от клъстери, които групират заедно сходните обекти. Сегментацията може да бъде напълно самостоятелната цел на изследване, т.е. определяне на сегментите да бъде основната цел на DataMining процеса. Но често тя е само една стъпка от решаването на други задачи. В тези случай целта на сегментацията е се да намерят еднородни подмножества от данни, които след това могат по-лесно да бъдат анализирани. Обикновено в големи бази от данни различни зависимости между данни влияят едни върху други и затрудняват

намирането на интересни закономерности. В този случай подходящото сегментиране на данни значително улеснява основната задача.

- **Анализ на крайностите** – данните могат да съдържат определени обекти, които не отговарят на общия модел на данни. Такива обекти се наричат крайни или екстремни (outliers). Повечето от методите за анализ разглеждат екстремните обекти като шум, който се игнорира. Обаче, в някои приложения, такива като диагностиката или разкриването на измами, рядко срещаните обекти са значително по-интересни от данните подчиняващи се на общи закономерности. Задачата за намиране на подобни обекти се нарича анализ на крайностите (outlier detection). Екстремните стойности могат да бъдат намерени чрез използване на статистическите тестове, базиращи се на определен вероятностен модел на данни или чрез използване на мерки за разстояние, смятайки че екстремните са тези обекти, които се намират на едно значително разстояние от всички клъстери. Методите, базирани на отклонение, опитват да идентифицират крайностите чрез анализ на разликите в основните характеристики на обекти в групата. Анализът на крайностите позволява да бъдат разкрити измами с използване на кредитните карти чрез откриване на покупки, чиято стойност е много по-голяма за дадената сметка в сравнение с редовни плащания, осъществявани по сметката. Крайностите могат също така да бъдат открити при анализа на местоназначение и типа на покупката или честота на покупките.
- **Задачата за описание на понятия** (concept description) цели създаване на разбираеми описания на понятия или класове. Тук целта не е да бъдат създадени пълни модели с висока предсказваща точност, а да бъдат създадени описания на отделни класове, позволяващи да бъде разбрана тяхната структура. Например, една компания би била заинтересована да научи повече за свои надеждни и ненадеждни клиенти. От създадени описания на тези два класа, компанията може да определи подходи, позволяващи й да запази лоялните си клиенти и да се освободи от нелоялните.

Описанията на понятия могат да бъдат използвани и за целите на класификацията. От друга страна, някои класификационни техники създават класификационни модели, които могат да се разглеждат като описание на понятия. Важното различие се състои в това, че класификацията цели създаването на пълни модели, т.е. модели, които могат да бъдат приложени към всички данни. Описанието на понятия не е задължително да бъде пълно – достатъчно е то да описва важни части от понятия или класове. Описанията на понятия най-често се представят във вид на правила или прототипи.

- **Класификация и предвиждане.** Класификацията предполага наличието на множеството от обекти, описвани чрез определени атрибути и принадлежащи към различни класове. Целевият атрибут приема дискретни (символни) стойности и е известен за всеки обект. Целта е да бъдат построени класификационни модели (класификатори), които назначават коректен клас на по-рано неизвестни обекти – обекти с предварително неизвестна стойност на целевия атрибут. Класификационните модели предимно се използват за предсказващото моделиране. Стойността на целевия атрибут може да бъде зададена предварително (например от потребителя) или изведена от сегментацията. Преди построяване на класификационни модели желателно е да бъде прилаган анализ на изключения, който позволява те да бъдат премахнати и по този начин да бъде подобрен класификационният модел. От своя страна, един класификационен модел може да бъде използван за откриване на отклонения от общата норма.

Предсказването е много сходно с класификацията, като разликата е в това, че при предсказването целевият атрибут не приема дискретни, а непрекъснати стойнос-

ти. Това означава, че целта на предсказването е да се намери числовата стойност на целевия атрибут за по-рано неизвестни обекти. Този тип задачи често се нарича регресия. Когато предсказването работи с данни за времеви серии, тази задача носи название предвиждане (forecasting). Регресионните модели могат да бъдат представени във вид на регресионни дървета, невронни мрежи, генетични алгоритми и т.н.

o **Анализът на зависимостите** се състои в намирането на модел, който описва важни зависимости (или асоциации) между отделни елементи от данни или събития. Зависимостите могат да бъдат точни или вероятности.

- Асоциациите са един специален случай на зависимостите, те описват сходството между елементите на данни, т.е. тези елементи: двойки атрибут-стойност, които често се срещат заедно. Типичното приложение на асоциациите е анализ на потребителските кошници. Така следното асоциативно правило „В 30% от всички покупки бира и фъстъци са били закупени заедно" е типичен пример на една асоциация. Алгоритмите за намиране на асоциации са много бързи и извеждат много асоциации. Избор на най-интересните от тях е предмет на интензивна изследователска работа.

- Времеви серии (редове) са специален вид зависимости, извличани от данни, представляващи последователности от събития във времето. Така например в областта на анализа на потребителските кошници последователните шаблони описват покупките на определен потребител или група от потребители във времето.

За извличане и представяне на моделите данни се използват различни статистически и математически средства и методи на изкуствения интелект като: невронни мрежи, машинно обучение, генетични алгоритми, дървета на решенията и др. Най-често използваните статистически средства са: групиране на данни, статистика, пресмятане на средни величини и отклонение, честотни разпределения и др. Моделите създавани от съответните средства представят компресиран образ за входните данни и съдържат основните познания, извлечени от данните на базата на определящия алгоритъм. Основни средства, използвани от техниките за извличане на закономерности от данни са [9]:

- *Дървета на решенията* – служат за класифициране на данни, като използват теглови коефициенти, за да се разпределят данните на все по-малки групи. С всеки възел в дървото има свързано правило, което предсказва целевата стойност с определени „сигурност" (confidence) и „подкрепа" (support). „Сигурност" е мярка на вероятността, с която възела на дървото предсказва целевата стойност. Това е съотношението на броя на случаите във възела, в които условието е предсказано вярно към общия брой случаи във възела. „Подкрепа" е съотношението на броя на случаите във възела, в които условието е предсказано вярно към общия брой случаи в базата от данни, т.е. мярка – колко случаи от базата са свързани с условието. Така дървото на решението е последователност от правила, които осигуряват итеративно разделяне на данните в дискретни групи, посредством проверка на условие във всеки възел.

- *Асоциативни правила* – чрез тях се класифицират данните, на базата на набор от правила, подобни на правилата в експертните системи. Тези правила могат да бъдат генерирани чрез процеса на търсене и изследване комбинации от правила или се извличат от дървета на решенията.

- *Невронни мрежи* – знанията са представени под формата на връзки, свързващи набор от възли. Силата на връзките дефинира свързаността между данните. За определяне значението на целевия показател се използва набор от входни параметри, математически функции и тегла. Изпълнява се повтарящ се цикъл на обучение, като

невронните мрежи променят теглото, докато се определи действителната стойност на входния параметър. След това невронните мрежи се превръщат в модел, които може да бъде приложен към нови данни с цел прогнозиране.

- *Генетични алгоритми* – моделът използва повтарящи се еволюционни модели, включващи операциите селекция, мутация и кръстосване (смесване). За избор на дадена особеност или отклонение се използва „функция приспособление" (fitness function). Генетичните алгоритми се прилагат често при оптимизиране на топологията на невронните мрежи и теглата им. Въпреки това те могат да се използват като средство за моделиране и самостоятелно.

- *Машинно обучение* (machine learning) – са компютърни програми, които анализират данни, с цел да получат образци от информация, които се използват за решаване на специфични проблеми.

- *Иновационен анализ* (OLAP) – като сложен статистически анализ в реално време, следвайки дедуктивна (query-oriented) стратегия за анализ на данните. Потребителите формулират хипотези и изпълняват заявки, за да разберат значението на данните. Обикновено този анализ се използва като допълнение към Data Mining средствата. Самият Data Mining анализ следва индуктивна стратегия за анализиране на данните.

- *Изводи на базата на сравнение или изводи базирани на прецеденти* – тези алгоритми се основават на откриването на аналози от миналото, най-близки до текущата ситуация, с цел уточняване на стойност или предсказване на резултати.

Съществуват и редица други средства за анализ на данните. Някой средства за извличане на закономерности върху данните автоматично избират най-подходящия техника и настройки, а при други те се указват от потребителя.

5. Агент за извличане на знания

Когнитивният агент [4] е автономен софтуер, който има способността да развива своите знания, посредством откриване на нови знания чрез математически методи. Като се има предвид различната степен на интелигентност, когнитивните агенти могат да бъдат класифицирани в три категории:

- Когнитивни агенти на базата на правила, които развиват знанията си с помощта на вградени в системата механизми за извод.
- Когнитивни агенти, които извличат нови знания с помощта на техники на машинно обучение.
- Когнитивни агенти на базата на анализ на данни чрез по-мощни Data Mining техники за откриване на знания [3].

Общата рамка на архитектурата на такава система включва следните компоненти:

- Базата от знания – съдържа всички знания необходими на агента. В изследваната от нас обучаваща среда това са индукционни правила от вида „IF условия THEN действия". Индукционното правило по своята същност е Булева формулла

R: X → Y, където X и Y са множества от клаузи.

X наричаме условие на правилото, а Y е неговото следствие. Клаузата се композира от два елемента (a,b) ∈ (A,V), където A е множество от променливи, а V е множество от стойности. Елемент (a,b) принадлежи на Декартовото произведение A x V. Също така в базата се съхраняват факти, изведени от околната среда или изведени от агент чрез механизмите за извод.

- Базата от мета-знания. Тя е структурирана като множество от мета-правила, задаващи връзката между правилата.
- Модул за извличане на нови правила. Основната роля на този инструмент е непрекъснато да се извичат нови знания, намирайки връзки между правилата, с цел ускоряване на процеса на извод.

Базирайки се на данните, натрупвани през годините в средата за електронно обучение Мудъл, изследваме възможността за извличане на знания и връзки между данните за студентите в процеса на обучението. Правилата, които се извличат и анализират (в случая индукционни правила) стават голям брой, който непрекъснато нараства. Идеята е, да групираме правилата в отделни групи. Когато се налага агент да вземе решение за обучавания обект на база на правилата от базата знания – той да не обработва последователно всички правила, а само правилата от групата, която е с най-голяма степен на сходство с търсената цел. Сходството между две правила измерва степента на прилика между тях, различието определя степента на несъответствие. Интуитивно, правилата са подобни, когато те споделят много идентични компоненти като клаузи.

На базата на клъстерен анализ правилата могат да се групират в класове. Агентът ще има способността да взима решения, като намалява мащаба на правилата, които се обработват по време на работа на системата. Така преди да започнете процеса на извод, агентът трябва да локализира съответствие с клъстер от правила на текущия факт от базата от знания, с който ще се работи.

4. Заключение

Базирайки се на натрупани данни от работата на система за електронно обучение с различни потребители, прилагайки средства от областта на извличане на знания от данните могат да се взимат различни решения като:

- да се оптимизират техниките за избор на тестови елементи и подходящ вид намеса в дейността на обучаемия;
- да се идентифицират типове обучавани, на които да се предлага подходящо продължение на обучението;
- да се предвидят обучаваните, за които има опасност да не се справят с обучението;
- да се правят анализи относно степента на придобиване и забравяне на знанията за различни интервали от време и за различни типове задачи, както и сравнение на показателите през годините.

При дистанционната форма, прилагайки електронното обучение се разкрива нов облик на процеса на преподаване, използвайки най-новите информационни технологии. Интегрирането на системите за електронно обучение с Data mining средства е необходимо за процеса на персонализиране на курсове за дистанционно обучение. На базата на получените резултати могат да се въведат допълнителни мерки за анализ и промяна на обучаващите курсове. Това от своя страна е път към повишаване на качеството на обучението във висшето училище.

Благодарности

Изследването е свързано с работата по проект НИД-4/2015 г. към фонд Научни изследвания на Бургаския свободен университет „Персонализиране на средата за електронно обучение“.

Литература

1. S. Stoyanov, I. Ganchev, M. O'Droma, H. Zedan, D. Meere, V. Valkanova, Semantic Multi-Agent mLearning System, A. Elci, M. T. Kone, M. A. Orgun (Eds.): „Semantic Agent Systems: Foundations and Applications", Book Series: Studies in Computational Intelligence, Vol. 344, Springer Verlag, 2011, ISBN: 978-3-642-18307-2
2. Хаджиколев, Е., С. Хаджиколева, А. Вангелова, Едно решение за автоматизирано генериране на потребителски справки в Мудъл, Пета национална конференция по Електронно Обучение във висшите училища, 16-17 Май 2014 г., гр. Русе, 181-187, 2014.
3. Chemchem, A., Drias, H. From data mining to knowledge mining: Application to intelligent agents. International Joint Conference on Artificial Intelligence: Expert Systems with Applications, (2014).
4. Shen, W., Hao, Q., Yoon, H. J., Norrie, D. H. Applications of agent-based systems in intelligent manufacturing: An updated review, Advanced Engineering INFORMATICS, 20(4), (2006), pp.415-431.
5. Moodle, Database schema introduction, http://docs.moodle.org/dev/Database_ schema introduction.
6. Официален сайт на Мудъл, https://moodle.org/.
7. Орозова Д., Е. Николова, Възможности на мобилното обучение (m-learning), международна конференция БСУ, 13-14 юни 2008, 345-350.
8. Попчев И., Бизнес интелигентност: практически аспекти, Международна конференция 2006, сборник доклади том III, БСУ, Бургас, 336-340.
9. Han, Jiawei and Micheline Kamber. 2007. Data Mining: Concepts and Techniques. San Fransisco, CA: Morgan Kaufmann publishers.
10. Anderson L. W., D. R. Krathwohl (Eds). A taxonomy for learning, teaching and assessing: A revision of Bloom's Taxonomy of educational objectives: Complete edition, New York, Longman, 2001.
11. Barr A., Feigenbaum E., The Handbook of Artificial Intelligence, vol.2, 1984.
12. Орозова Д., К. Паскалев, Когнитивно моделиране в интелигентни инструментални среди, в: Съвременни тенденции в развитието на фундаменталните и приложни науки – VI национална конференция на Съюз на учените в България, Стара Загора, 1995, стр. 279-284.
13. Stoyanov, S., Cholakov, G., Valkanova, V. Sandalski, M., Personalized, Reactive and Proactive Providing of e-Learning Services, E-Learn 2011, Honolulu, USA.
14. D. Orozova, S. Stoyanov, I. Popchev, „Virtual education space", International Conference „Knowledge – tradition, innovation, perspectives", Burgas, Bulgaria, ISBN: 978-954-9370-99-7, vol. 4, pp.153–159, 2013.
15. С.Стоянов, Д.Орозова, И.Попчев, Виртуално образователно пространство – настояще и бъдеще, Юбилейна конференция на БСУ, том. II, стр. 410-418.
16. Орозова Д, М. Георгиева, Приложения на „Големите данни", Списание „Компютърни науки и комуникации", Том 3, № 4 (2014), БСУ, Бургас, стр.118-121.
17. Mark Hornick, Erik Marcade, Sunil Venkayala, Java data mining: strategy, standard, and practice, A practical Guide for Architecture, Design, and Implementation, 2006.
18. https://www.datascience.com/resources/topic/education