СЪВРЕМЕННИ МЕТОДИ ЗА ИНТЕЛИГЕНТЕН АНАЛИЗ НА ДАННИ

Д-р инж.Мартин Пъшев Иванов,
главен асистент, НБУ –департамент „Информатика”,
e-mail: mivanov@nbu.bg

Анотация: В настоящата публикация се представят и анализират проблемите и задачите, възникващи в една относително нова област на обработка на данните – техният интелигентен анализ със средствата на математическата статистика, изкуствения интелект, машинното обучение и т.н. Отбелязано е значението на тази област за процеса на вземане на решение, както и основните направления, в които се развиват иследванията и използваните в практиката методи. Систематизирани са типовете задачи и средствата за решаването им. Разгледани са и някои специализрани приложения на Data Mining-подхода като Text Mining и Web Mining. Отбелязан е и подходът Big Data в обработването на големи съвкупности от дани.

Ключови думи: *Data Mining, Text Mining, Web Mining, Big Data, класификация, клъстеризация, асоциативни правила.*

1. Същност на проблема

1.1 *Общи бележки*

Всяка съзнателна и целенасочена човешка дейност е свързана със създаването, обработката и използването на значителни обеми информация. Тези естествени процеси обикновено остават незабелязани и недобре осмислени в ежедневната ни дейност, но с развитието на формалните инструменти за обработка на данните, техническата база на изчислителните системи и специализирания софтуер те придобиват все по-голямо значение. Във всички съвременни дейности набирането, съхраняването, обработката на информацията и извличането на полезни резултати от анализа ѝ е ключов процес, съществено влияещ върху желания краен резултат и тази трайна тенденция ще се задълбочава и ще има все по-голямо значение.

Съществена характеристика на съвременната информационна епоха е динамичното и изключително бързо нарастване на обемите данни и на обменяните информационни потоци. Източниците на данни са разнообразни:

- Бизнес - Интернет, електронна търговия, транзакции, фондова борса, публикувани официални счетоводни и финансови отчети и др.
- Наука - Remote sensing, bioinformatics, scientific simulation
- Общество и индивид - новини, развлечения, мултимедйна продукция, социални мрежи и др.

Натрупването на огромни обеми от организирани по различен начин първични данни създава предпоставки за по-задълбоченото им аналитично изследване и обобщаване, за прилагане на специфични аналитични техники с оглед разкриването на съществени за регистрираните явления и процеси вътрешни връзки и зависимости, които обикновено не могат непосредствено да се установят дори и от подготвен специалист-аналитик. Съвременните средства за интелигентен анализ на данни обикновено се обозначават чрез

повсеместно приетия термин ***Data Mining***, който много често се употребява без еквивалентен превод дори и в не-англоезичните източници (Aggarwal, 2015), (Sumathi,2006).

Терминът Data Mining е получил своето наименование от идеята за търсене и откриване на полезна и ценна информация, съдържаща се в големи бази или хранилища на данни, която идея от някои изследователи метафорично се уподобява на практиката на минната дейност за търсене и откриване на ценни руди и минерални изкопаеми. Аналогията идва най-вече от факта, че и в двата случая ценният добив идва в резултат на обработката на значително количество суров изходен материал.

Терминът Data Mining се оказва наистина трудно преводим, но близки до неговия смисъл са употребяваните термини като „добиване на данни", „разкриване на знания", „интелигентен анализ на данни". Намираме последния за най-сполучлив, защото той представя и съдържателната страна на тази дейност, основаваща се на съвременни методи и техники за аналитично изследване. Друго, също така често срещано наименование на предмената област е „разкриване на знания в бази от данни" (Knowledge Discovery in Databases, KDD), доколкото почти цялото информационно съдържание, които е обект на Data Mining, са съсредоточени и организирани в разнообразни бази и хранилща на данни.

Като термин Data Mining се появява през 1978 г. и придибива голяма популярност в съвременната си интерпретация през първата половина на 90-те години на миналия век. До този момент обработката и анализът на данни се извършват с конвенционални средства, основавщи се на класическата статистика, като се прилагат преди всичко към неголеми бази от данни.

На Data Mining се отделя нарастващо внимание в информационната индустрия и в обществото, особено през последните години поради широката достъпност на огромни обеми от данни, както и поради нарастващата необходимост тези данни да бъдат превърнати в полезна информация и знания. Така придобитите информция и знания могат с много голям успех да бъдат изплзвани в различни приложения, вариращи от маркетингови анализи, установяване на опити за измами, запазване на клиенти, до управление на продукцията и в мащабни научни изследвания.

Областта Data Mining може да бъде разглеждана като резултат на естествената евоюция на информационните технолгии (Aggarwal, 2015). Софтуерната индустрия на системи, базирани на бази от данни засвидетелства еволюционно развитие по отношение на следните функционалности: набиране на данни и създаване на бази от данни. Управление на данните (data management), включително съхраняване и възстановяване на данните, обработка на транзакциите в бази от данни, и разширен анализ на данните (включващ средствата и методите, използвани в хранилищата на данни и интелигентния им анализ). Така например ранните техники на разработване на информационни и бази от данни послужиха като предпоставка за последващото разработване на ефективни механизми за съхранение, извличане и възстановяване на данни, за изпълнение на заявки и обработка на транзакции. За многобройните системи за управление и поддържане на бази от данни интелигентния анализ на данните става по естествен път следващата ключова задача.

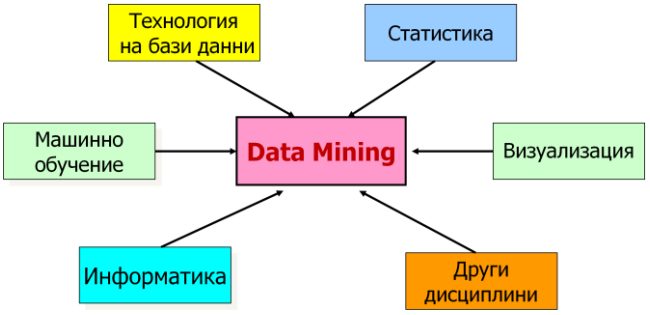
1.2. *Работно определение за Data Minig* (*Gregory Piatetsky-Shapiro*)

Едно общоприето определение на понятието Data Mining (Usama, Piatetsky-Shapiro, 1996) е следното: Data Mining е процесът на разкриване в „суровите" необработени данни на неизвестни преди нетривиални, неочевидни, обективни, практически полезни и достъпни интерпретации на знания, необходими за вземането на решения в различни области човешката дейност. Това определение акцентира върху следните особености на получените от този процес резултати:

- **Неочевидни** – закономерностите не могат да бъдат установени със стандартни методи за обработка на информацията или по експертен път.
- **Обективни** – закономерностите съответстват максимално на действителността (за разлика, напр. от експертното мнение, което винаги съдържа субективен елемент).
- **Практически полезни** – закономерностите (изводите) имат конкретно значение, на което може да се намери практическо приложение.

1.3 *Data Mining като интердисциплинарна област*

Data Mining представлява интердисциплинарна област, възникнала и развиваща се а на базата на такива съседни области като приложна статистика, машинно обучение, изкуствен интелект , теорията на базите от данни и др. Data Mining ползва теоретични резултати и приложни процедури, които се включват както в областите, показани на фиг.2, така и в много други направления на науката, съдържащи потенциал за постигане на целите на интелигентния анализ на данните (Graham, 2011), (Han, 2006).



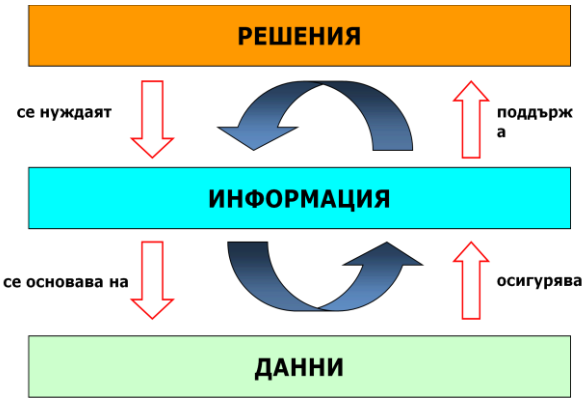
Фиг. 1: Интердисциплинарната област Data Mining и конактните ѝ научни области.

Понятието Data Mining може да бъде определено в няколко различни аспекта, в зависимост от целите на изследването и от контекста, в който то се употребява:

- **подход и методология** – определя насоките за изследване и развитие на ефективни модели и методи за анализ на данните и разкриване на съществените за практиката шаблони и връзки.
- **технология** – ефективни средства и техники на изпълнение на задачите по обработка на данните и извличане на даните, включително използването на специализран софтуер.
- **процес** - това е динамичен процес на поддържането на задачите за вземане на решения, чрез няколко свързани по между си фази.

1.4 *Data Mining и процесът на вземане на решение*

Data Mining като процес на анализ на данните и извличането на практически полезни връзки и зависимости е фаза, свързваща средствата и технологиите за набиране,съхраняване и достъп до данните от една срана, и процесите и моделите, и алгоритмите за вземане на решение от друга. Всеки акт на решение се нуждае от информация, изградена въз основа на данни и от установени знания, относно връзките и зависимостите между параметрите на управлявания процес. Схема на цикъла на информационното осигуряване на задачите на вземане на решение е показан на фиг. 1 .



Фиг. 2: Цикъл на информационното осигуряване на задачите за вземане на решение.

Фигурата илюстрира двата основни потока, които са включени в информационния цикъл на Data Mining, чиито резултати са придобиване на знание и вземане на решение:

- **Данни – информация – знания и решения**: вземането на решения изисква информация, която се основава на данни.
- **Задачи – действия и методи за решаване – приложения**: данните осигуряват информация, която поддържа решенията.

Големият обем информация, свързан с различните дейности в обществото от една страна позволява да се получат точни разчети, а от друга – превръща процеса на търсене на решение в сложна задача. В резултат на това се появява цял клас програмни системи, предназначени да облекчат работата на специалистите-аналитици – системи за поддържане на вземането на решения (*Decision Support System - DSS*). Могат да бъдат отбелязани три основни задчи, решавани в *DSS*:

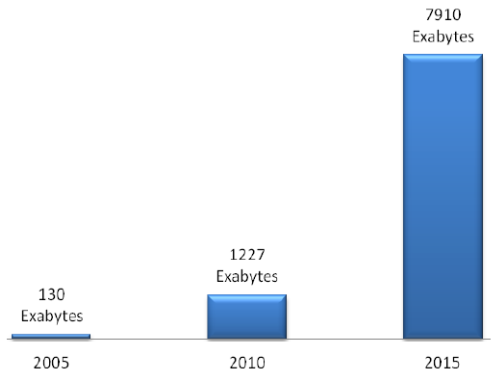
- въвеждане на данните;
- съхраняване на данните;
- анализ на данните.

2. Източници и тенденции в растежа на обема на данните

Източниците на данни в съвременния свят представляват необозримо множество. Те са повсеместни и са свързани с всички икономически, управленски, социални, научни, образователни страни на човешката дейност. Значителна част от данните са машинно генерирани, а обемът им нараства експоненциално във времето. Най-общо източниците на данните могат да бъдат систематизирани в следните групи:

- Релационни бази от данни;
- Хранилища на данни;
- Транзакционни бази от данни;
- Разширени информационни системи и приложения;
- Обектно-релационни бази данни;
- Текстови източници и приложения;
- Web.

Този списък е условен и подлежи на актуализация с разширението на областите, в които намират приложение информационните технологии. Експоненциалният ръст на данните, според изследване на на IDC Digital Universe Study за последните години е илюстриран на фиг. 3 .



Фиг. 3: Тенденция в ръста на обема данни.
Оценка на IDC Digital Universe Study за общия обем
генерирани и съхранени дани по години.

3. Данни, информация, знание

Следва да се вземе под внимание, че въпреки честото им смесване в популярния език, категориите „данни", „информация" и „знания" са свързани, но все пак съществено различни понятия и отразяват различни равнища на абстракция на представата ни за света и протичащите явления и процеси в него. За правилното разбиране и формулиране на проблемите на Data Mining е от значение тези различия да бъдат ясно дефинирани.

В най-общия случай **данните** представляват формална регистрация на някакви факти. Те могат да се представят чрез числови стойности, текст, графика, изображения, звук, аналогово или цифрово видео и др. Данните могат да бъдат получени в резултат на измерване, експерименти, аритметични и логически операции и др. Те трябва да бъдат представени във

форма, позволяваща съхранението им, предаването им и обработката им. Иначе казано – данните са необработения („суров") материал, предоставен от някакъв източник и използван от потребителя за създаване на информация въз основа на тези данни.

Информацията е по-висша форма на абстракция на представата ни за обектите и явленията и се представя чрез образци, схеми, асоциации, свързващи известните ни данни.

Знанието е най-висша форма на абстракция, която се получава въз основа на информацията като резултат от разсъдъчната ни дейност и от прилагането на някои, често интуитивни процедури. Знанията представляват съвкупност от сведения, които образуват описание, съответстващо на дадено равнище на осведоменост по определен въпрос, предмет, проблем и т.н. Знанията се използват за получаване на конкретни предимства и резултати.

4. Характеристики на данните, използвани в Data Mining

4.1 Измервания и скали за представяне на данните

Познаването на начините на получаване и представяне на различните типове данни е от съществено значение за избор на методи за тяхната обработка. Значителна част от данните се получават чрез измерване (респ.остойностяване) на параметри и атрубути на наблюдаваните явления и процеси. Под "остойностяване" се разбира процесът на присвояване на някакви символни (вкл.логически) или числови стойности на наблюдаваните характеристики в съответствие с някакво правило (т.е.изобразяването им върху някакво символно или числово множество). Множествата, носители на стойностите, в които се изобразяват характеристиките, заедно с правилата на изобразяването се наричат "скали". Според свойствата си скалите могат да бъдат различни категории. Преди всичко те могат да бъдат разделени на дискретни и непрекъснати:

- **Дискретни скали** – в случаите, когато остойностяването се извършва върху дискретно или изброимо множество от числови или символни стойности. В тази категория влизат и скалите, представящи логически типове данни (true/false, 0/1 и т.н.). Данните, представяни върху дискретни скали се наричат **дискретни данни**. Възможните стойности на тези данни могат да бъдат обозначени (номерирани) с естествените числа.
- **Непрекъснати скали** – множествата, носители на стойностите са непрекъснати и обикновено са числови, зададени в краен или безкраен интервал. Данните, представяни върху непрекъснати скали се наричат **непрекъснати данни**.

Освен това общо разделение е прието да се работи с четири основни типа скали – номинална, рангова (ординална), интервална и относителна.

- **Номинална скала** (nominal scale) – съдържа само категории – данните се представят символно, като не могат да бъдат подреждани в някакъв ред и върху тях не могат да бъдат прилагани аритметични операции. Номиналната скала може да се състои от наименования, категории, класификационни признаци на обекти. Примери за такива скали са имена на географски обекти, цветове, професии, символи в някаква азбука, пол, семейно положение и др. В тази скала са дефинирани единствено операциите „равно" (=) и „не равно" (≠). Данните, представяни в такава скала

понякога се наричат „категорийни”, но коректното им наименование носи името на скалата – „номинални”.

- **Рангова** (ординална) скала (ordinal scale) – в която представените обекти имат някаква относителна позиция един спрямо друг. Ако позициите се представят чрез числа, то тези стойности по никакъв начин не отразяват степента на различие на обектите. Ранговата скала дава възможност върху множеството на обектите да бъде дефинирана релация на наредба. Класически пример за такава скала са всички бални системи за оценяване на знания и умения на учащи се, системите за класиране в съревнования, някои метеорологични и сеизмични скали и др. За тази скала освен релациите „равно” (=) и „неравно” (≠) е дефинирана и релация на наредба (т.е.валидни са операциите за сравнение “по-голямо” > и „по-малко” <). При някои изследвания представените в тази скала данни също се определят като категорийни, доколкото върху тя не могат да бъдат прилагани основните математически операции.
- **Интервална скала** (interval scale) – скала, при която разликите във стойностите на представените величини могат да бъдат изчислени (като дължина на интервал), но определянето на количествено отношение между тях няма смисъл. Тази скала притежава свойствата на номиналната и ранговата скала, като позволява при това да се намери разликата между две величини и да се определи количествено стойността на признака. Типичен пример за такава скала са температурните скали на Целзий или на Фаренхайт, при които могат да бъдат измерени съответните стойности, те могат да бъдат съпоставени по величина („<”, „>”), могат да бъдат пресметнати и съпоставени интервалите на изменението им (като температурни разлики), но самите температурни стойности не могат да бъдат съпоставени в никакво количествено отношение (не можем да твърдим, че температурата 20^oC е три пъти по-ниска от температура 60^oC). За номиналната скала са приложими операциите: равно (=), не равно (≠), по-голямо (>), по-малко (<), операциите събиране (+) и изваждане (-). Номиналната скала представя по принцип непрекъснати числови данни.
- **Относителна скала** (ratio scale) – скала, в която има определена точка на отчитане (репер) и има смисъл определянето на количествено отношение за стойностите ѝ. Пример за такава скала са повечето от скалите за измерване на физически величини: маса, разстояние, скорост, енергия, остойносяванията в икономиката и финансите, физиологични характеристики в медицината и др. Това е числова скала, предоставяща най-голямо съвършенство при представяне на данните, защото освен операциите, дефинирани и допустими за горните три скали, тя позволява и прилагането на аритметичните операции „умножение” (.)и „деление” (/).

В някои случаи се употребява и т.нар. „дихотомична скала” (**dichotomous scale**), съдържаща само две възможни стойности. Пример за такава скала е представянето на логически стойности, на пол (мъж,жена), ден/нощ и т.н. По същество тази скала представлява разновидност на номиналната, но самостоятелната ѝ употреба се мотивира от множеството изследвания, в които тя е удобна форма за представяне на даните.

Възможни са и други класификации на данните според други трхни особености или целите на изследването, за което се използват. Така например според представянето им в базите от данни могат да бъдат различавани следните групи данни:

- **Релационни данни – това са** данните от релационните таблици.
- **Многомерни данни** – данни, представени в *OLAP*-кубове.
- **Измерение (*dimension*) или "ос"** – в могомерните данни – обединение на данни от един и същи тип, което позволява да се структурира многомерната БД.

По критерий "постоянство на стойностите в процеса на решаване на задачата" данните могат да бъдат:

- **Променливи данни** – които изменят своите стойности в процеса на решаванена задачите.
- **Постоянни данни** – данни, които съхраняват своите стойности в процеса на решаване на задачите и не зависято от външни фактори.
- **Условно-постоянни данни** – данни, които винаги могат да променят своите стойности, но тези изменения не зависят от процеса на решаване на задачата, а от външни фактори.

Според времевите си характеристики данните могат да бъдат:

- **Данни за период** – характеризират някакъв период от време (доход за месец, средна температура за месец, потребление на енергия, курс на валута за периода и т.н.).
- **Точкови данни** – представят стойността на някаква променлива в конкретен момент от време.

Според структурираното им представяне данните се разделят на следните категории:

- **Сруктурирани данни** – това са данни, които са групирани в релационни схеми (редове и колони в стандартна релационна база от данни). Конфигурацията на данните и съгласуваноста им позволяват да се получат отговори на прости или съставни заявки със средствата за манипулиране на релационни бази от данни (SQL-заявки).
- **Полу-структурирани данни** – това е форма на структурираните данни, която не следва изрично релационната схема. За тези данни е присъщо, че те са „самоописващи" се и съдържат етикети или други маркери (метаданни), които уточняват фактическото и, съдържание и позволяват то да бъде представено в организирана форма (напр.като йерархии от записи и полета).
- **Неструктурирани данни** – това са данни във формат, който не може да бъде представен чрез релационна схема или в самоописващи се структури. Такива са например свободния текст, файловете с графично и видео съдържание и др.

5. Основни задачи на Data Mining

В основата на съвременната технолгия Data Mining (discovery-driven data mining) е положена концепцията за разкриването на шаблони, отразяващи някакви многоаспектни отношения в данните (Han,2006). Шаблоните представляват закономерности, свойствени на наличните данни, които могат да бъдат компактно представени в разбираема за човек форма. Шаблони са центровете на групиране и величината на разсейване на наблюдаваните стойности, връзката между изменението на независими и зависими променливи, предсказаните технологични или икономически връзки въз основа на известни характеризиращи параметри и др. Търсенето на такива връзки-шаблони се извършва с методи, които не са ограничени от априорни предположения за структурата на извадката или за вида на разпределенията на стойностите на анализираните показатели.

Задачите в Data Mining се решават изключително за сметка на систематизирането и обработката на големи обеми данни. Решението на задачите се представя под формата на логически, конструктивни или формални модели, чиито ключови параметри се определят чрез подходящи техники и алгоритми. Определянето на стойностите (числови, логически и т.н.) на тези ключови параметри обикновено се отбелязва като „обучение на модела".

Задачите на Data Mining могат да се класифицират в две най-общи категории, в зависимост от поставената цел и от резултата, който се получава от анализа:

- **Дескриптивни задачи** – свързани са основно с установяването на някакви присъщи характеристики и връзки, описващи състоянието на наличните до момента данни и поведението на процесите, които те описват. Такива са задачите на първоначалния статистически анализ на данните, задачите за клъстеризация и др.
- **Предиктивни задачи** – отнасят се до съставянето на прогнози или предвиждане за бъдещото изменение на данните и параметрите на процесите въз основа на наблюдаваните им стойности и динамика до момента. Такива са задачите за класификация, регресионите статистически модели, анализът на динамични редове (time series) и т.н.

Сред изследователите и специалистите по Data Mining няма единно мнение какви точно типови задачи следва да бъдат предмет на областта. Обикновено основните тематични проблеми в Data Mining, посочвани в повечето авторитетни източници са следните: класификация, клъстеризация, прогнозиране и предвиждане, асоциация, визуализация, идентификация и анализ на отклоненията, оценяване, анализ на връзките, обобщаване. Следва да се отбележи, че проблемната област на Data Mining не е затворена, Data Mining представлява преди всичко ефективен прагматичен подход и той може да включва всякакви смислени и съдържателни задачи, имащи отношение към подпомагането на процеса на вземане на решения с нужните средства и модели за представяне на информацията и знанията.

6. Методи на Data Mining

Следва да се отбележи, че всъщност Data Mining не разполага със свои собствени методи, а използва методите на съседните на научно-приложни области (фиг. 1). Както

неколократно бе отбелязано, Data Mining е по-скоро конструктивна, прагматично ориентирана, а не теоретична област (Zaki,2014), (Olson,2008). Всеки метод, потвърдил на практика ефективността си, може да бъде изплзван за целите на Data Mining, дори и в случаите, когато неговата ефективност не е обоснована теоретично по формален път. Всички такива случаи, обаче, трябва да се вземат предвид от изследователите и практикуващите DM специалисти и да не се прилагат сляпо и немотивирано, като им се възлагат свърхголеми очаквания. Data Mining не е вълшебно решение на проблема за изследване на данните, а път и технология, които непрекъснато се усъвършенстват, разширяват и които откриват нови възможности и хоризонти за приложение.

Сред методите на Data Mining следва по-специално да бъде обърнато внимание на следните групи:

- **Методи на математическата и приложна статистика** – това са добре познати традиционни методи за обработка и анализ на масови съвкупности от данни, които отговарят на целите на подхода Data Mining. Такива методи са например:
    - Дескриптивен (първичен) анализ и описание на изходните данни.
    - Анализ на връзките (корелационен анализ и регресионен анализ, факторен анализ, дисперсионен анализ).
    - Многомерен статистически анализ (компонентен анализ, дискриминантен анализ, многомерен регресионен и корелационен анализ).
    - Анализ на динамични редове и прогнозиране.
- **Методи, основаващи се на теорията на изкуствения интелект, машинното обучение и евристични методи** – това са методи, чрез които може да бъде изследвана връзката между състоянието на съвкупност от входни данни или предпоставки и изходния резултат. Обикновено те позволяват да се моделира поведението на реални технически, технологични или икономически процеси, без да се анализира в детайли тяхната структура. Голяма част от методите от тази група нямат строго формална обосновка и се основават по-скоро на общи рационални принципи, потвърдени от многократния изследователски и практически опит. Тези методи привличат вниманието на разработчиците с разширените си възможности, със способността си да решават множество нетрадиционни задачи и с факта, че могат лесно и ефективно да бъдат реализирани в програмни инструменти и да бъдат вградени в специализирани софтуерни ситеми. Без да бъде изчерпателен, списъкът на тези методи включва:
    - **Изкуствени невронни мрежи** – подходящи за решаване на апроксимационни и оптимизационни задачи.
    - **Еволюционно програмиране**, включващо алгоритми за групово отчитане на аргументите.
    - **Генетични алгоритми** – за решаване на оптимизационни задачи, които се представят с по-сложни или неявни формални модели.
    - **Асоциативна памет** – за търсене на аналози и прототипи на поведение в съвкупност от данни.
    - **Размита логика и изчисления** – приложими когато моделът съдържа параметри, чиито стойности не могат да бъдат точно определени или тези стойности съдържат твърде високо ниво на паразитно влияние („шум”).

- **Дървета на решенията** – използвани в класификационните задачи за съставяне на класификациони правила.
- **Системи за обработка на експертни знания** – приложими в случаите, когато отсъства описание на модела във формален вид.

6.1 *Стратегии на обучение*

В случаите, когато е известен типът и структурата на задачата, за окончателната ѝ формулировка е необходимо да бъдат конкретизирани редица нейни определящи параметри. Така например в задачите за класификация се изисква да бъдат определени онези правила, атрибути и стойности, които ще доведат до коректното отнасяне на текущо разглеждания обект към съответния клас. В оптимизационните задачи, например, трябва да бъдат пресметнати онези стойности на управляващите параметри на модела, които ще доведат до желаното целево състояние и т.н. Определянето на целесъобразните стойности на параметрите на модела в много от задачите става чрез итеративен процес, наречен „обучение" на модела и може да извършен по пътя на две основни стратегии:

- **„Обучение с учител"** (контролирано обучение, supervised learning) – при които стойностите се определят въз основа на предварително зададени образци, описващи наблюдаваното явление или обект. Стойностите на параметрите на модела се определят чрез изчислителен алгоритъм (процес), като същите периодично се съпоставят с регистрираното в известните образци състояние на обекта и се коригират така, че грешката да бъде минимална. Процесът на обучение на модела обикновено включва два етапа. На първият етап се използва съвкупността от предварително зададените образи, наречена „обучаващо множество" (training set) или „обучаваща извадка". Вторият етап включва оценка на надеждноста и точността на модела. В него също се изплзва съвкупност от известни наблюдения на явлението, които не участват в процеса на обучение, наречена „тестово множество" (test set). Процедурата на втория етап проиграва съставения модел с даните от тестовото множество и сравнява моделираните и фактическите стойности. Пример за задача от този тип е класификацията, при която изследователят разполага с множество примерни образци за принадлежността на обектите към даден клас, заедно с текущите стойности на атрибутите им. Класически пример за тази стратегия са и методите за обучение на изкуствени невронни мрежи.
- **„Обучение без учител"** (unsupervised learning) – в този случай отсъства предварителна информация, представена в известни, налюдавани образци на връзка между параметрите на процеса. Обучаващият алгоритъм съдържа механизъм за адаптиране на тези параметри към най-подходящите им стойности чрез използването на съответни критерии за целта. Пример за такава задача е клъстеризацията, при която отсъства каквато и да е информация за принадлежността на наблюдаваните обекти към някаква група.

7. Задача за класификация

Класификацията е най-често решаваната задача от областта на Data Mining и в много от случаите – най-простата. По същество решението ѝ предполага отнасянето на даден обект към някаква предварително определена класификационна група (клас) въз основа на стойностите на атрибутите, които го описват – цвят, размер, предназначение, цена и т.н. Атрибутите на обектите, определящи съществените за класификацията им характеристики, се наричат техни **предиктори**. Класификацията на реални обекти обикновено е неоднозначна и зависи от целта и интерпретацията на изследователя. Класовете могат да бъдат определени чрез бизнес-правила, чрез своите граници или чрез някаква математическа зависимост. Класификацията принадлежи към групата модели, наречени "обучение с учител" (*supervised learning*). Целта на класификацията е да създаде кратък обобщен модел на изменението на зависимия атрибут (принадлежност към клас), свързано с поведението на атрибутите-предиктори.

Популярни определения на задачата за класификация са:

- системно разпределение на изучаваните предмети, явления, процеси по родове, видове, типове според някакви съществени признаци с цел удобство на изследването им, групировка на изходните данни и разполагането им в определен ред, отразяващ степента на сходството им.
- подредени по някакъв принцип множество обекти, които имат сходни класификационни признаци (едно или няколко свойства), избрани за определяне на сходството или различието между тези обекти.

Задачите за класификация могат да бъдат систематизирани в няколко основни категории:

Двоична класификация – с два възможни изхода, определени от принадлежността на обектът към дадена група („да" или „не").

Класификация с множество възможни изходи – обектът се класифицира в една от няколко възможни групи (класове) в зависимост от стойността на атрибутуте си.

Едномерна (по един признак) и **многомерна** класификация (по множество признаци) – в зависимост от това, колко атрибута (класификатора), описващи обекта се вземат под внимание в задачата за класификация.

7.1 Примерни *практически приложения на задачите за класификация*

- Като разполагаме със стари данни за кредитоискатели за обслужваните от тях кредити да определим надеждноста на новите въз основа на наблюдаваните тегни характеристики (професия, образование, пол, доход, наличие на постоянна работа, семейно положение, здравословно състояние и др.);
- Като разполагаме с данни за потребителите на битов ресурс (вода,електроенергия) да определим различна търговска политика (респ.тарифни планове, начини на отчитане и разплащане и др.).
- Диагностиката в медицината е типичен пример на сложна задача за класификация, в която според симптомите на заболяването и данните от клиничните изследвания се поставя диагноза (принадлежност към клас заболяване).

7.2 *Обща формулировка на задачата*

Всеки класификационен подход използва съвкупност от признаци, за да класифицира отделен обект. Нека тези признаци се представят от случайните променливи X_1, …, X_k (предикторни променливи), а Y е зависимата променлива, съдържаща стойностите (етикетите) на възможните класове. Всяка променлива X_i притежава домейн (област на изменение) $dom(X_i)$ $(i=1,...k)$. Зависимата променлива Y притежава домейн $dom(Y)$. P е съвместното вероятностно разпределение на $dom(X_1) x … x dom(X_k) x dom(Y)$. Обучаващата база от данни D_m е случайна извадка от P.

Предиктор d е функцията:

$d: dom(X_1) … dom(X_k) \rightarrow dom(Y)$.

- Ако Y е категорийна (номинална) променлива, задачата е класификационна и използваме етикета на класа C вместо Y.

  $|dom(C)| = J$.

- C – етикет на класа, d - класификатор.
- Нека r е запис, случайно подбран от P. Дефинира се мярка на погрешната класификация (*misclassification rate)* за d:

  $RT(d,P) = P(d(r.X_1, …, r.X_k) != r.C)$.

Дефиниране на задачата: Нека съвкупността от данни D е случайна извадка от вероятностното разпределение P, да се намери класификатор d,който да минимизира $RT(d,P)$.

Основни въпроси при съставянето на задачата за класификация

- Какво е естеството на съвкупността от данни, която трябва да се класифицира? Трябва да бъде определено предназначението на класификацията и то трябва да бъде свързано с интерпретацията на очакваните резултати.
- Каква трябва да бъде точността на класификацията? Постигането на висока точност в повечето случаи изисква продължителна работа на алгоритмите, което може да се окаже неподходящо за целите на бизнеса.
- Доколко разбираема трябва да бъде класификацията? Някои от моделите (дървета на решенията) дават резултати, които обясняват по-добре връзката между предикторната и зависимата променлива, докато при други (невронни мрежи) тази връзка не е така изяснена.

Моделите на задачата за класификация (според формата на представянето и метода за решаването им) са твърде разнообразни:

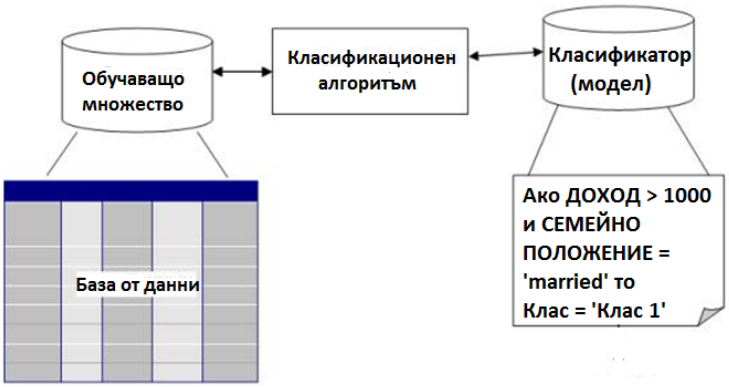
- Дървета на решенията (класификационни дървета);
- *CHAID* - Chi-square Automatic Interaction Detector;
- Случайни “гори” и усилени дървета;
- Логистична регресия;

- Невронни мрежи;
- Метод на “най-близкия съсед”;
- Наивна Бейсова класификация и др.

Методите за решаване на класификационните задачи като цяло се основават на стратегията „обучение с учител”. Както беше отбелязано, тази стратегия изисква наличието на две съвкупности от исторически данни-наблюдения на поведението на класифицираните обекти – с известни класове. Първата съвкупност образува обучаващото множество. С нейна помощ се настройват параметрите на класификационния модел. Втората съвкупност е тестовото множество, което се изплзва за оценка на точността на решението.

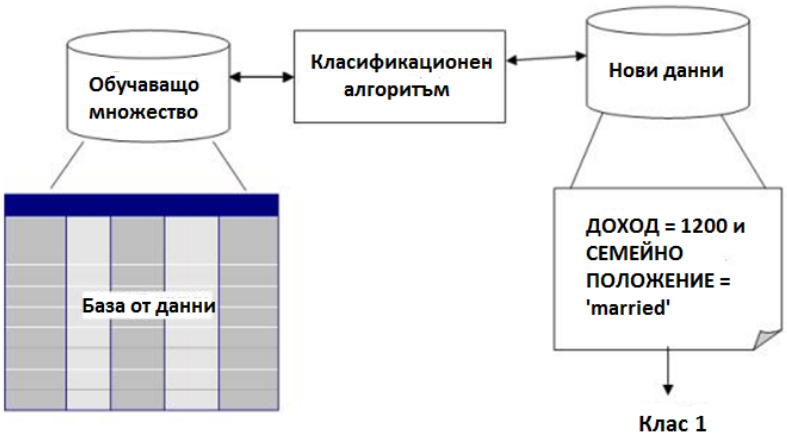
Процесът на класификация има за цел да се построи модел, който еднозначно съпоставя стойностите на атрибутите-предиктори с класа, към който принадлежи обекта. Основните фази на процеса “класификация” са две – конструиране на модела и неговото използване:

1. **Съставяне на модела** (фаза на обучение) – описание на множеството от предефинирани класове.
    1. Всеки обект от извадката се свързва с един от предефинираните класове чрез атрибута “етикет на клас” (ръководено обучение);
    2. Избира се множество от обекти, използвани за съставяне на модела – т.нар. “обучаващо множество”;
    3. Моделът се представя като система от класификационни правила, дървета на решенията или в някакъв формален вид.



Фиг. 3: Процес на класификация. Съставяне на модела.

2. **Прилагане на модела** – за класифициране на нови обекти
    1. Точността на модела се оценява чрез т.нар.”тестово множество”;
    2. Точността на модела се определя от процента правилно класифицирани обекти в тестовото множество;
    3. Тестовото множество не бива да съвпада с обучаващото множество.
    4. Ако точността е приемлива, съставеният модел се използва за практическото решаване на задачата.



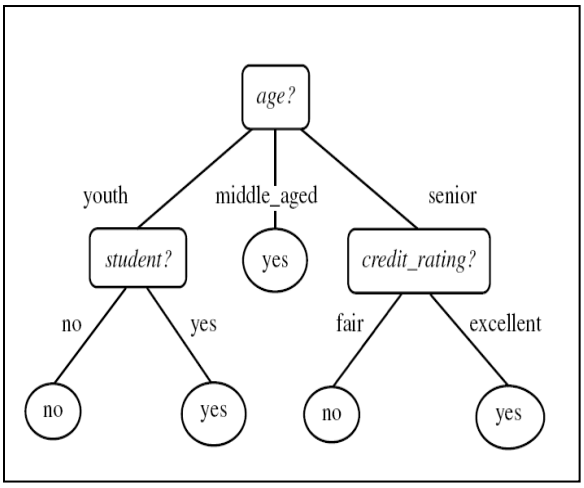
Фиг. 4: Процес на класификация – прилагане на модела.

Класификационните дървета (дървета на решенията) са най-често използваната форма за представяне и за решаване на класификационните задачи. **Класификационното дърво** (фиг. 5) е структура от възли и връзки между тях, като всеки възел има една входяща и две или повече изходящи връзки към съседни възли:

- **Възлите** биват: начален, междинни (разклонения) и крайни (листа). Началният възел съответства на началното състояние на процеса на класификация, като изразява първата логическа проверка, която трябва да бъде направена. Междинните възли съответстват на следващите логическите проверки, крайните възли – на класовете. Вътрешните възли съответстват на логическите проверки, крайните – на класовете.
- Всеки вътрешен възел има асоцииран разделящ предикат – т.е.логическа функция, сързана с изчисляването на условието за разклонение:

Класификационното дърво моделира процеса на вземане на решение относно класа, към който трябва да бъде отнесен класифицирания обект чрез поредица от проверки на логически условия. Всяка проверка се осъществява в конкретен възел, а поредицата от проверки протича по някой от клоните на дървото и води до класифицирането на обекта в определен клас. Дървовидният модел може лесно да бъде трансформиран в поредица от управляващи структури от вида **`if`** … **`then`** … **`else`** … .

Моделът на класификационото дърво в терминологията на теорията на графите се описва като свързан ориентиран граф без цикли (наречен също „дърво”).



Фиг. 5: Класификационно дърво.

На фиг. 5 е показан популярен пример на класификационно дърво, съставено за задача, в която кредитор решава дали да отпусне целеви кредит за закупуване на компютър на кандидат, в зависимост от възрастта му (“youth”, “middle_aged”, “senior”), кредитния му рейтинг (“fair”, “excellent”) и това дали е учащ или не (“student”?). Кандидат, който отговаря на условията (“youth”, “student”) бива одобрен за целеви кредит. Кандидат с атрибути за възраст “senior”и за кредитен рейтинг “fair” получава отказ.

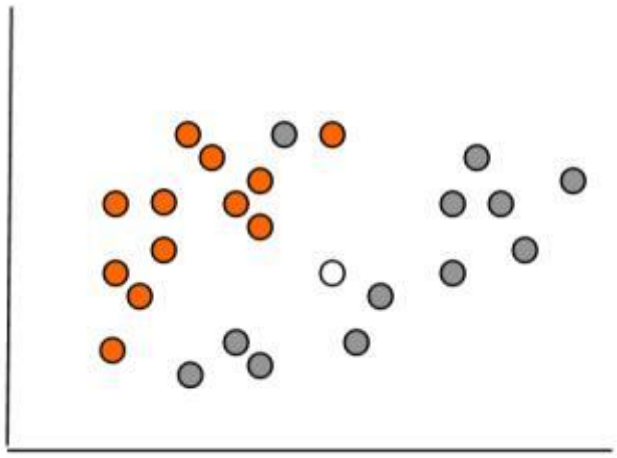
Други, често използвани методи за класификация, са невронните модели, наивна бейсова класификация и др. Всеки от тях използва специфична структура на представяне на връзките и зависимостите между атрибутите на класифицираните обекти. Поради ограниченото място тук тези методи няма да бъдат разглеждани. Следва да се има предвид, че всеки от методите има свои предимства, особености и е препоръчително да се използва в подходящи и обосновани случаи. Методите не гарантират съвпадащи решения на една и съща класификационна задача. Много често в практиката те се използват съвместно, като взаимно се допълват. Във всички случаи изборът на най-подходящия класификационен метод включва съпоставяне на резултатите от класификацията и оценка на точността на класификационното решение.

Оценката на ефективността и точността на съставения класификационнен модел е съществен етап от решаването на класификационната задача. Полученото решение (съставеният класификационен модел) се проверява върху тестови данни с известни целеви стойности на класификационния атрибут (известни класове). Проверката се състои в сравняването на предсказаните от модела класификационни групи и фактическите им стойности, взети от тестовите данни. Най-често използваните методи за оценка на точността на класификацията включват:

- **Матрица на грешките** (*Confusion matrix*, *contingency table*) – това е квадратна матрица, с брой редове и колони, съответстващи на класовете. Редовете на матрицата отговарят на действително наблюдаваните попадения в съответен клас, колоните – на класовете, получени при прилагането на класификационния модел върху тестовите данни. По такъв начин всяка клетка съдържа броя на случаите, принадлежащи реално към клас A, класифицирани от модела като клас B .

- **Брой (процент) на коректно класифицираните случаи** – проста метрика, обобщаваща данните от класификацията в най-груб вид.
- ***True Positive (TP) rate*** (дял на правилно класифицираните) – е частта от екземпляри, които са класифицирани от модела в клас *X*, разделена на всички налични екземпляри от клас *X*.
- ***False Positive (FP) rate*** (дял на неправилно класифицираните) – е частта от екземпляри, които са класифицирани от модела в клас *X*, но реално принадлежат на други класове, разделена на всички екземпляри, които не принадлежат към клас *X*.

Методите и показателите за оценка на качеството на класификацията са инвариантни по отношение използваните класификационни модели и методи.



Фиг.6: Множество от обекти на наблюдения, представени в двумерно пространство.

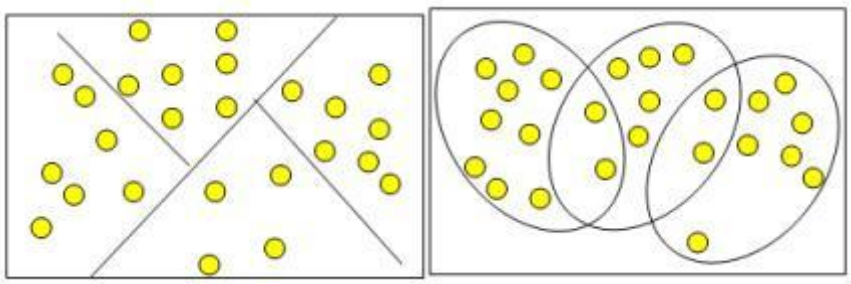
8. Задача за клъстеризация

Задачата за клъстеризация (или т.нар.”клъстерен анализ”) е подобна на задачата за класификация, на която е логично продължение. Тази задача също има за цел групирането на обекти в групи по сходство по даден признак. Разликата между двете задачи е, че класовете на изучаваната съвкупност от данни в задачата за клъстеризация не са предварително известни. Синоними на задачата за клъстеризация са “автоматична класификация”, “съставяне на таксономия”. Клъстеризацията е предназначена за разбиване на съвкупността от обекти на еднородни групи (клъстери). Ако данните от клъстеризацията бъдат изобразени като точки в пространството на признаците, то клъстеризацията може да се представи чрез определянето на концентрацията им около центрове на сгъстяване. Целта на клъстеризацията е търсене на съществуващи структури в съвкупността от данни, когато първите не са явно изразени. Често клъстеризацията се прилага като предварителна стъпка за групиране и пресяване на данните в *Data Mining*-процеса**.**

Клъстерът може да се характеризира като група обекти, имащи общи свойства. Два основни признака характеризират клъстерите:

- Вътрешна еднородност;
- Външна изолираност.

Клъстерите обикновено са „непресекаеми” или „ексклузивни” (т.е.нямат общи елементи) – (non-*overlapping*, exclusive), но могат да бъдат и пресекаеми, при което вторият от отбелязаните признаци се смекчава (фиг. 7). Клъстерите обикновено се изобразяват в n-мерно евклидиво пространство и се описват чрез съответни геометрични характеристики – координати на центъра (центроида), радиус и т.н. За оценка на сходството на обектите, попадащи в даден клъстер, се използват метрики, които обикновено се основават на евклидовото разстояние между точките или подобни на него характеристики.



Фиг. 7: Непресичащи се и пресичащи се клъстери.

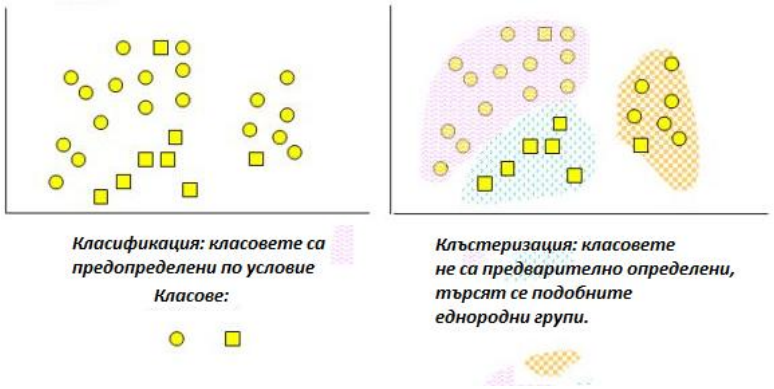
Клъстеризацията първоначално е прилагана в биологичните науки, антропология и психология, а впоследствие - и в решаването на различни икономически задачи, с оглед организирането и представянето на данните в нагледни структури и подпомагането на задачите за вземане на решение.

8.1 *Особености на задачата за клъстеризация*

- За разлика от задачата за класификация, клъстерния анализ не изисква априорни предположения за съвкупността от данните, не налага ограничения върху представянето на изследваните обекти, позволява да се анализират показатели от различни типове от данни (интервални данни, честоти, двоични данни и т.н.).
- Важно е да се вземе под внимание, че атрибутите, описващи обектите, трябва да се представят в сравними скали.
- Клъстерният анализ позволява да се съкрати размерността на данните и те да се представят в по-нагледен вид.
- Клъстерният анализ може да се прилага към съвкупности от динамични редове (*time series*). Могат да бъдат определени периоди на сходство на някои от показателите и да се определят групи от динамични редове със сходна динамика.

Таблица 1: Сравнителна таблица на основните характеристики на задачите за класификация и за клъстеризация.

Характеристика	Класификация	Клъстеризация
Контролируемост на обучението	Контролирано обучение	Неконтролирано обучение
Стратегия	Обучение с учител	Обучение без учител
Наличие на етикети на класовете	Обучаващата съвкупност се съпровожда от етикет, който указва класа, към който се отнася наблюдението.	Етикетите на класовете в обучаващата съвкупност са неизвестни.
Основание за класификация	Новите данни се класират въз основа на обучаващата съвкупност.	Съвкупността от данни се обработва с цел установяване на класове или клъстери.



Фиг. 8: Сравнение на задачите за класификация и за клъстеризация.

8.2 ***Методи за решаване на задачата за клзстеризация***

Методите за решаване на задачата за клъстеризация са две основни групи:

- **Йерархични –** чрез рекурсивно обединяване или разделяне на обектите в клъстери се създава дървовидна структура на клъстеризацията (изобразена като дендограма). Методите са подходящи за прилагане върху ограничено множество от неголям брой обекти.
- **Нейерархични -** При голям брой на наблюденията йерархичните методи на клъстерния анализ са неприложими. В такива случаи се използват нейерархични методи, основаващи се на итеративни техники за раздробяване и групиране на изходната съвкупност. В процеса на разделяне новите клъстери се формират до изпълнението на правилото за спиране на процеса. Най-популярният метод от тази група е този на k-средните (k-means), известен още като „бърз клъстерен анализ". Други разпространени методи за клъстеризация са EM (expectation maximization) , Fuzzy C-Means.

Всяка от групите методи включва множество специфични подходи и алгоритми. Важно е да се има предвид, че различните методи на клъстерния анализ могат да дадат различни

решения на задачата при едни и същи данни. Тази особеност се приема за естествена, но съществено усложнява задачата и следва да се отчита от практикуващите клъстеризацията специалисти. Целесъобразно е методите да бъдат комбинирани, а резултатите от прилагането им – критично и аналитично съпоставяни.

8.3 ***Качество на клъстеризацията***

Оценка на качеството на клъстеризацията може да бъде извършена чрез някоя от следните процедури:

- Ръчна проверка – приложима единствено при относително малък брой на обектите;
- Установяванена контролни точки и проверка на получените клъстери – когато има априорна информация за наличието на такива клъстери и за тяхното разпределение;
- Използване на количествени метрики, оценяващи вътрешното групиране на обектите в клъстерите и разстоянието между центровете им.

9. Задачи за прогнозиране

Задачите за прогнозиране се решават в най-разнообразни области на човешката дейност – наука, икономика, производство и множество други. Прогнозирането е съществен елемент от организацията на управлението както на отделния стопански субект, така и на икономиката като цяло.

Задачата за прогнозиране (forecasting) се счита за една от най-сложните задачи в областта на Data Mining. Тя изисква внимателен анализ на големи обеми от данни, както отчитане влиянието на множество свързани и динамични фактори върху параметрите на прогнозния модел. Прогнозирането е насочено към определяне на тенденциите на динамиката на конретен обект или събитие въз основа на ретроспективни данни. В някои случаи като задачи за прогнозиране се представят и проблеми, в които отсъства динамика на наблюденията във времето, а се търси адекатно описание на връзката между стойностите на зависими параметри и влияещите върху тях стойности на независими величини (фактори, предиктори).

Методите за решаване на задачата за прогнозиране се основават изключително на статистическия подход. Във всички случаи като решение на задачата се търси обоснована статистическа оценка на прогнозната стойност, съчетана с построяването на нейната доверителна област (тази област, в която вероятността за попадане на оценката е достатъчно голяма). Популярните методи за решаване на задачата са следните:

- **Регресионен анализ, обикновено линеен** – използването му се прилага за да се намери статистическа количествена връзка между стойностите на прогнозираната величина и предикторите (респ.времето). Въпреки разпространеното си приложение, използването на този подход е нецелесъобразно за прогнозиране, тъй като екстраполирането на прогнозните стойности извън наблюдавания времеви интервал води до бързо натрупване на грешка.

- **Динамични редове, времеви редици (time series)** – статистически методи, специално предназначени за изследване на тенденциите в редици от наблюдавани стойности. Широко прилагани методи с възможност за отчитане на тенденции, периодични и сезонни колебания и др.
- **Невронни мрежи** – притежават същите качества както методите на time series с възможност за адаптиране към измененията в особеностите на редиците от наблюдения.

Във всички случаи на прилагането на модели за прогнозиране следва да се съчетава с анализ на тяхната точност и адекватност, което се прави със статистически средства. Следва да се отбележи известно сходство между задачите за класификация и тези за прогнозиране, доколкото и в единия, и в другия случай целта е да се предскаже вероятното състояние (респ.принадлежност към клас) на изследвания обект въз основа на стойностите на описващите го променливи. Правилно проведеното изследване и адекватно съставения прогнозен модел могат да се окажат изключително полезни в много практически дейности на живота.

10. Разкриване на асоциативни връзки (*association rule mining*)

Това е една от най-разпространените задачи за анализ на даните в областта на Data Mining и се състои в определянето на често срещащи се съвкупности от данни, свързани в някаква неявна взаимна (асоциативна) връзка. Идеята се е появила от някои маркетнигови изследвания, в които се оценяват и анализират асоциативните зависимоси при покупката на свързани продукти в големите супермаркети – т.нар.”шаблони на покупките”. Подхдът е предложен от *Agrawal* през 1993 г. Задачата е известна още и под името „анализ на пазарната кошница” (market basket analysis). Това е съществен модел на *Data Mining*, който е предназначен за откриването на асоциативни връзки между множества от обекти в бази от данни. Първоначално задачата за разкриване на асоциативни правила (*association rule mining*) е била ориентирана към намирането на типични шаблони от покупки в супермаркети (купуване на свързани артикули), затова понякога тя се нарича и “задача за анализ на пазарната кошница”. Задачата работи с категорийни (номинални), а не с количествени данни. Тя се отличава от задачите за класификация и клъстеризация по това, че установяването на закономерностите става не въз основа на свойствата на анализирания обект, а между няколко събития, които настъпват едновременно.

Асоциативното правило има вида: “От събитието *A* следва събитие *B*".

„Ако клиентът купува **Клавиатура**, то той купува **Мишка** и купува **Подложка**”.

Асоциативно правило: (**Клавиатура**) $\rightarrow$ (**Мишка**) $\rightarrow$ (**Подложка**)

Основните термини, описващи задачата за разкриване на асоциативни правила са:

- ***Пазарна кошница*** (*market basket*) – това е съвкупността от артикули (обекти), които се придобиват от купувача в рамките на отделна транзакция. Транзакциите са достатъчно характерни операции, чрез които могат да се опишат резултатите от посещението на различни магазини.

- ***Транзакция*** (*transaction*) – множество от събития, които настъпват едновременно. Регистрирайки всички бизнес-поерации, по време на своята дейност, търговските дружества натрупват огромни масиви от транзакции. Всяка транзакция представлява съвкупност от стоки, купени от купувача за едно посещение. Получените в резултат на анализа шаблони включват списък на стоките и броя на транзакциите, които съдържат дадените стоки.
- ***Транзакционна*** *или* ***операционна база от данни*** (*transaction database*) е двумерна таблица, която включва номерата на транзакциите (*TID*) и списък на продуктите, набавени при тази транзакция.
- ***TID*** (*transaction ID*) – уникален идентификатор, определящ всяка сделка или транзакция (ред в транзакционната база).

TID	**Списък артикули**
100	Хляб, мляко, бисквити
200	Мляко, сметана
300	Мляко, хляб,сметана, бисквити
400	Колбаси, сметана
500	Хляб, мляко, бисквити, сметана
600	Бонбони

Фиг.9: Пример за транзакционна база от данни, съдържаща шест транзакции.

10.1 ***Формален модел на задачата***

Нека $I=\{i_1,i_2,\ldots,i_m\}$ е множество от литерали (артикули) i_j $i=1,\ldots,m$, D е база от данни за транзакции $D=\{T_k\}$:

- $T_k \in D$ – транзакция – $T_k \subseteq I$.
- *TID* – уникален идентификатор, свързан с T_k.
- X е подмножество на I ($X \subseteq I$).

Една транзакция T_k съдържа множеството артикули X, ако $X \subseteq T_k$.

Асоциативно правило е импликация във формата:

$X \rightarrow Y$, където X, $Y \subset I$, и $X \cap Y = \varnothing$.

- Всяко множество от артикули се нарича ***itemset***.
- Всяко множество от k артикула се нарича ***k-itemset***.

Асоциативното правило се характеризира от два вида метрики за своята значимост (сила):

- **Поддръжка** (*support*) – е броят или процентът транзакции от базата, съдържащи определен набор артикули:

  SUP(хляб,мляко,бисквити) = 3 или в проценти:

  SUP(хляб,мляко,бисквити) = (3/6)*100 = 50%.

- **Достоверност** (*confidence*) – показва каква е вероятността (условна) включването в транзакцията на артикул *A* да води и до включването на артикул *B*:

  conf = P(B|A)

10.2 *Алгоритми за намиране на асоцативни правила*

Алгоритмите за разкриването на асоциативни правила са в състояние да намерят всички правила от вида: "От *A* следва *B*" (*A -> B*) с различни стойности на поддръжка (SUP) и на достоверност (conf). В повечето случаи, обаче, е необходимо броят на тези правила да се ограничи с минимални и максимални стойности на поддръжката (SUP) и достоверността (conf). Ако стойността на поддръжката на правилото е твърде голяма, то в резултат на изпълнението на алгоритъма ще бъдат намерени очевидни и известни правила. Твърде ниската стойност на поддръжката ще доведе до намирането на голям брой правила, които няма да са известни и очевидни, но ще бъдат в значителна степен необосновани. Ако достоверноста е ниска, то ценността на такива правила е под съмнение.

Съществуват голям брой алгоритми за решаване на задачата за намиране на асоциативни правила, които използват различни стратегии и структури от данни. Най-известният от тях е алгоритъмът Apriori. Резултатите от изпълнението на тези алгоритми (генерираните множества от асоциативни правила) при зададени база от транзакции, и стойности на минималната поддръжка и достоверност съвпадат. Те са дефинирани еднозначно. Различията (и насоките за усъвършенстване) на алгоритмите са свързани с тяхната изчислителна ефективност и изискванията към необходимата памет.

11. Анализ на неструктуриран текст - Text Mining

С *Text Mining* се означава една специфична област на Data Mining, характерно за която е, че първичните данни не са представени в структуриран вид (т.е.не са организирани като таблици, записи, типизирани стойности и т.н.), а като свободен текст – т.е.в неструктуриран вид (Miner, 2012), (Feldman,2006). Изследователите обръщат внимание, че такива са данните в над 90 % от достъпните информационни източници, в които се съдържа изключително ценна информация и огромен потенциал за извличане на полезни знания. По същество задачите на Text Mining са много близки до формулираните по-горе задачи на Data Mining (класификация, клъстеризация, асоциативни правила и т.н.), но основният проблем при решаването им е свързан с неструктурираното представяне на данните в свободния текст и многообразието на формите на изразяване, както и с особеностите на вътрешната морфологична структура и богатството на естествения език. Този проблем е твърде тежък и до момента не е изцяло преодолян, но основните насоки са свързани с анализа на изходния текст и с отделянето на такива негови елементи, които могат да бъдат използвани за структурираното му представяне. Основните подходи са два:

- Използване на лингвистични методи за обработка на естествен език (Natural Language Processing - NLP);
- Методи, основаващи се на статистическия анализ на съдържанието на текста и откриването на онези негови определящи елементи и връзки между тях, които могат да послужат за иденифицирането и за категоризирането му.

Към настоящия момент Text Mining разчита предимно на втория подход, като за описанието на структурата на текста се използва специфичното му преобразуване в т.нар. модел ”торба от думи” (bag of words), при който морфологичната му структура се разрушава и текстът се представя чрез множество вербални единици (атрибути), заедно с честотата на появата им в източника на данните (документа). За да бъдат приложени класическите методи за класификация и клъстеризация, се въвеждат специфични метрики, оценяващи сходството на текста в отделните източници. Като цяло статистическите методи за обработка на текст и за решаването на задчите на Text Mining изискват много по-задълбочен подход, с анализ и запазване на семантичното текстово съдържание и все още са в експериментална фаза.

12. Анализ на Web-съдържание - Web Mining

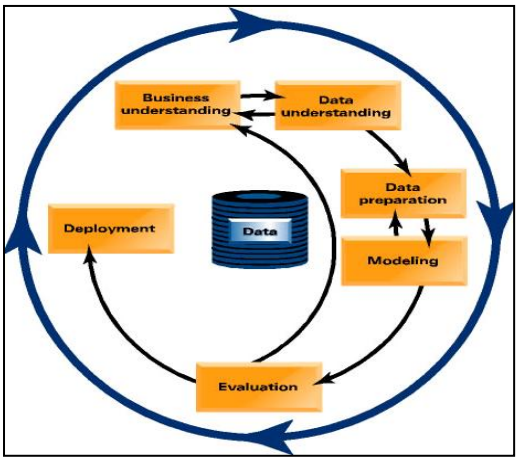
Световната мрежа (World Wide Web) представлява огромна разпределена глобална информационна служба за новини, реклама, информиране на потребителя, за финансова информация и управление, обучение, електронна търговия и много други информационни услуги. Web съдържа разнообразно и динамично множество от хипервръзки и достъп до Web-страници и предоставя огромно богатство от източници за Data Mining. Наред с това Web поражда значителни предизвикателства за ефективно използване на ресурсите и за разкриване на съдържащите се в тях знания. Задачите на Web Mining включват разкриване структурата от Web-връзки, анализ на Web-съдържанието и шаблоните за достъп във Web, разкриване на „авторитетно” Web-съдържание, автоматична класификация на Web-документи, анализ на използване на Web-ресурси и т.н.

13. **Процес на Data Mining**

За да бъде обосновано, методически завършено и полезно едно Data Mining изследване, то трябва да отговаря на определен стандарт. В специализираната литература се препоръчват различни виждания и концепции, които са предназначени да служат като подробен план за организиране на процеса на набиране на данни, анализиране на данните, анализиране и тиражиране на резултатите, внедряване и ползване на резултатите и контрол за качеството и за непрекъснатото усъвършенстване на процеса (Zaki,2014).

В средата на 90-те години на XX-ти век (септември 1996 г.) от европейски консорциум на компании (*SPSS/ISL, NCR, Daimler-Benz, OHRA*) е предложен стандартен общодостъпен модел на процеса Data Mining. Същият е наречен ***Cross Industry Standard Process for Data Mining – CRISP-DM.*** Стандартът CRISP-DM представлява изчерпателна Data Mining методология и модел на процеса, която предоставя както на новите потребители, така и на експертите пълен и подробен план за изпълнение на Data Mining проекта. *CRISP-DM* е приложим във всяка делова или научна област, която извършва *Data Mining* и не ограничава избора на използвания специализиран софтуер или инструменти. Дефинирането на стъпките на *Data Mining*-процеса е ориентирано към по-ефективно и с по-добро качество изпълнение на проектите, с по-малко разходи на време и средства. Водеща в разработването на CRISP-DM е идеята, че *Data Mining*-

процесът трябва да бъде надежден и да може да се повтори от специалисти със сравнително малък опит в областта.



Фиг. 9: Жизнен цикъл на Data Mining-проекта съгласно CRISP-DM.

CRISP-DM разделя жизнения цикъл на *Data Mining*-проекта на шест фази:

1. **Съставяне на представа за бизнеса –** тази фаза се фокусира върху изясняването на целите на проекта и изискванията от перспективата на бизнеса, след което преобразува това знание в дефиниция на задачата на *Data Mining* и в съставянето на предварителен план за постигане на целите.

2. **Изясняване на същността и качеството на данните –** тази фаза започва с първоначално набиране на данните и преминава в дейности, които целят по-добро запознаване с тези данни, проверка на тяхното качество, разкриване на най-общите им характеристики и особености (първоначално видими), както и установяването на интересни връзки и зависимости, които ще помогнат за издигане на хипотези за скритата в тях информация.

3. **Подготовка на данните –** тази фаза покрива всички дейности, свързани с конструирането на крайните набори от данни от изходните "сурови" данни. Задачите, свързани с подготовката на данните могат да се изпълняват многократно и в произволен ред. Задачите включват: съставяне на таблици, избор на записи и атрибути както и "прочистване" на данните, интеграция, трансформация, редукция и т.н.

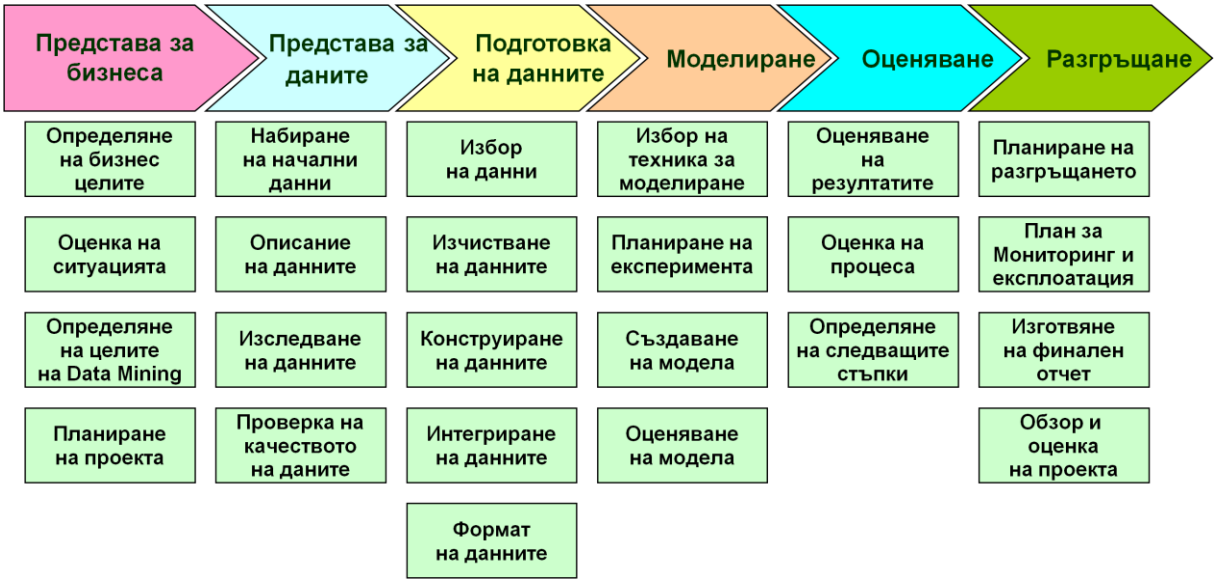
4. **Моделиране –** на тази фаза се избират и се прилагат различни техники на моделиране като техните параметри се калибрират към оптималните стойности. Типично е да съществуват няколко моделиращи техники за решаването на дадена *Data Mining*-задача. Някои от техниките имат специфични изисквания към формата на данните, поради което често се налага връщането към Фаза 3 "Подготовка на данните".

5. **Остойностяване –** фазата включва задълбочено оценяване на модела и преглед на стъпките, изпълнени за конструиране на модела за да се убедим, че бизнес целите се постигат правилно. Една от ключовите цели е да се определи дали има някакъв важен

въпрос на бизнеса, който не е разгледан достатъчно. В края на тази фаза се взема решение за използване на резултатите от *Data Mining*-проекта.

6. **Разгръщане (внедряване)** – тази фаза има два акцента: прилагане на резултатите от модела в практиката и създаване на условия за непрекъснатост на процеса *Data Mining.* Придобитите знания трябва да бъдат организирани и представени по начин, по който потребителят да може да ги използва ефективно. В зависимост от изискванията, фазата на разгръщане може да бъде съвсем проста като напр.генерирането на един отчет или сложна като осъществяването на повтарящ се *Data Mining*-процес в бизнес-сруктурата.

На фиг. 10 е изобразена в разгънат вид сруктурата на процеса CRISP-DM.



Фиг. 10: Структура на процеса CRISP-DM.

14. Обработка на масивни съвкупности от данни - Big Data

Поради динамичния ръст на обема на генерираните данни от разнообразните обществени, стопански и научни дейности все по-често възникват ситуации, свързани със съхраняването, поддържането и обработката на тези данни, с които съвременните компютърни технологии и програмни инструменти не са в състояние да се справят ефективно и в разумни срокове. Непрекъснатото усъвършенстване и нарастване на мощността на изчислителните средства съществено смекчават този проблем и разширяват диапазона на поддаващите се на решаване задачи, но като цяло проблемът остава тежък и трудно преодолим.

14.1 *Терминология*

Big Data („големи данни”) е термин, който е свързан с използването на големи набори от данни (Leskovec, 2016). Като Big Data се определят съвкупности (колекции) от набори данни, твърде големи и сложни, за да бъдат обработвани с наличните инструменти за управление на бази от данни. Като предизвикателство те включват набиране (придобивание, получаване), съхранение, търсене, споделяне, анализ и визуализация. Тенденцията към увеличаване на обема на съвкупностите от данни произтича и от допълнителната информация, която може да

се извлече от анализа на простите съвкупности от данни. Казано по друг начин Big Data е реализация на разширена технология за бизнес-интелигентност с цел съхранение, обработка и анализ на данни, която преди това е била игнорирана поради ограниченията на традиционните информационни технологии.

Основна особеност на използваните подходи в рамките на концепцията на „големите данни" е възможността за обработка на целия информационен масив за получаване на по-достоверни резултати от анализа. Традиционните подходи за анализ на данните (в т.ч. и методите на Data Mining) се основават на ограничена по обем представителна извадка, което закономерно води до увеличаване на грешката. Освен това този подход в много случаи на практика води до увеличени разходи за подготовка на данните, породени от особеностите на представянето им и изискването за яхното привеждане в определен формат.

Основните характеристики на Big Data се отбелязват в три направления (фиг. 11):

Фиг. 11: Трите основни характеристики на Big Data – обем, скорост и разнообразие – (*volume, velocity u variety*) – или V^3.

- **Обем** (Volume) – Big Data се характеризират с непрекъснато нарастване на обема на данните. Бизнес-звената са „потопени" в този нарастващ информационен поток, натрупващ с лекота терабайтове и дори петабайтове от информация.
- **Скорост** (Velocity) – обикновено в обработката на данните се свързва с някакви критични срокове за получаване на резултат. Времевият хоризонт на обработката на данните в реално време може да се окаже твърде кратък – от порядъка на минути, а за процеси, особено чувствителни към времето за реакция като установяване на агресивни опити за измама Big Data трябва да се прилага в поточен режим, за да се максимизира полезният резултат.
- **Разнообразие** (Variety) – Big Data обхваща разнообразни типове данни – структурирани и неструктурирани, различно мултимедийно съдържание в различни формати (нови източници – смартфони, таблети, сензори).

Акцентите върху тези три особености, с които се сблъскват проблемите на Big Data е маркират основните различия между тази сравнително нова област и традиционните технологии за обработка на данни, методите на т.нар."бизнес-интелигентност", а дори и тези на Data Mininig.

14.2 *Методики за анализ на „големи данни”*

Съществуващите множество разнообразни методики за анализ на огромни масиви от данни използват инструментариум, сходен с този, който се прилага и в Data Mining. Използване методи също се основават на резултати от статистиката, машинното обучение, кибернетични и евристични техники и др. Както и при интелигентния анализ на данни, списъкът на прилаганите методики не е ограничен и непрекъснато се допълва в резултат на множеството научни и експериментални изследвания, а също и вследствие на нарастващите потребности на практиката. Безусловно е, че колкото по-голям и диверсифициран е информационния масив, толкова по-точни и по-полезни ще са получените резултати.

14.3 Аналитичен инструментариум

Аналитичният инструментариум на Big Data към момента включва множество софтуерни продукти и фреймуърци, сред които има и такива, които се разпространяват свободно или на невисока цена. Най-типичните представители от тях са *MapReduce*, *Big Table* и *Apache Hadoop*.

MapREduce е рпроект, разработен от Google като средство за ефективно изпълнение на множество от функции с много голям обем от дани в пакетен режим. Компонентът „карта” (“map”) разпределя програмния проблем или задачи между голям брой системи и поддържа разположението на задачите по наичн, който балансира натоварването и управлява възстановяването от сривове. След като завършат разпределените изчисления, друга функция, наречена „редуциране” („reduce”) обобщава обратно заедно всички елементи, за да предостави резултата. Като пример за възможностите на MapReduce е определянето на броя на страниците на книга, написана на 50 различни езика.

Big Table е разработен от Google като разпределена система за съхранение на данни , предназначена за управлението и поддържането на мащабируеми структурни данни. Данните се организират в таблици с редове и колони. За разлика от традиционите релационни модели на бази от данни, Big Table представлява „разредена”, разпределена, трайна многомерна сортирана карта. Тя е предназначена за съхранението на огромни обеми данни върху общодостъпни сървъри.

Apache™ Hadoop® е проект, който заслужава особено внимание, тъй като се счита за един от осовополагащите за технологията BigData. Разработен е от фондацията Apache Software Foundation и представлява софтуерна рамка (библиотеки и набор от помощни програми) с отворен код за разработване и изпълнение на разпределени програми, работещи върху клъстери от стотици и хиляди възли. Използва се за реализация на търсещи и контекстни механизми на много сайтове с високо натоварване, в т.ч. за Yahoo! и Facebook. Разработен е на Java в рамките на изчислителния подход, съдържащ е в MapReduce, съгласно който приложението се разделя на голям брой еднакви елементарни задания, изпълнявани върху възлите на клъстера и по естествен начин трансформирани в краен резултат. Проектът включва четири модула:

- **Hadoop Common** – свързващо програмно осигуряване – набор инфраструктурни програмни библиотеки и утилити, използвани от другите програмни модуи и родствени проекти;
- **HDFS** – разпределена файлова система;

- **YARN** – система за планиране на заданията и за управление на клъстера;
- **Hadoop MapReduce** – платформа за програмиране и за изпълнение на разпределени MapReduce-изчисления.

Big Data се оформя като една от най-важните технологични тенденции която притежава потенциал за драматични промени на начините на организация и използване на информацията с цел разширяване на практиките на потребителите и за трансформиране на техните бизнес-модели.

Заключение

Интелигентният анализ на данни предлага обещаващи средства за ракриване на скрити връзки и шаблони, съдържащи се в големи обеми от данни. Тези скрити шаблони могат потенциално да бъдат използвани за предсказване на бъдещо поведение. Наличието на нови алгоритми за интелигентен анализ на данни, обаче, трябва да се посреща с повишено внимание и с известна педпазливост. Най-напред – тези техники могат да бъдат смо дотолкова добри и евефктивни, доколкото са добри данните, върху които се изпълняват. Наличието на качествени и надеждни данни е първото необходимо условие за тяхното изследване. След осигуряването на качествени данни, следващата стъпка е да се подбере най-подходящата техника за техния интелигентен анализ. Възможно е да се наложат известни компромиси при избора на целесъобразния метод, който може силно да зависи както от състоянието на данните, така и от целта на изследването. „Най-добрият" модел за обработка на данните и представяне на резултата често се намира чрез продължителен процес на проби и грешки – като се изпробват, а често и творчески съчетават различни технологии и алгоритми за да се постигнат възможно най-добрите резултати. Извън всякакво съмнение е, че обработката на данните със средствата на интелигентния анализ не се ограничва до няколко специфицирани предварително задачи, а също така, че този подход има значителен потенциал и съществено ще допринесе за технологичния, икономически и социален напредък на съвременното общество.

Библиография:

1. Aggarwal Charu C., Data Mining: The Textbook, Springer-Verlag, 2015.
2. David Hand, Heikki Mannila, Padhraic Smyth, Principles of Data Mining, The MIT Press, © 2001.
3. Feldman Ronen, James Sanger, The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data, Cambridge University Press, Dec 11, 2006.
4. Graham Williams, Data Mining with Rattle and R, The Art of Excavating Data for Knowledge Discovery, Springer Science+Business Media, LLC 2011.
5. Han Jiawei, Micheline Kamber, Data Mining: Concepts and Techniques, Second Edition, The Morgan Kaufmann Series in DataManagement Systems, 2006.
6. Leskovec Jure, Anand Rajaraman, Jeffrey D. Ullman, Mining of Massive Datasets, Stanford InfoLab, 2014 (като Интернет-ресурс: http://infolab.stanford.edu/~ullman/mmds/book.pdf, последно достъпен м.август 2016 г.).
7. Miner Gary, John Elder IV, Andrew Fast, Thomas Hill, Robert Nisbet, Dursun Delen Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications, Academic Press, Jan 25, 2012.

8. Olson David L., Dursun Delen, Advanced Data Mining Techniques, Springer-Verlag Berlin Heidelberg, 2008.
9. Sumathi S., S.N. Sivanandam, Introduction to Data Mining and its Applications, Studies in Computational Intelligence, Volume 29, Springer-Verlag Berlin Heidelberg 2006.
10. Usama M. Fayyad , Gregory Piatetsky-Shapiro , Padhraic Smyth Advances in Knowledge Discovery and Data Mining (American Association for Artificial Intelligence), Copirighted Material, 1996.
11. Zaki Mohammed J., Wagner Meira Jr., Data Mining and Analysis: Fundamental Concepts and Algorithms, Cambridge University Press, 2014.
12. Интернет-ресурс: Dean Jeffrey, Sanjay Ghemawat**,** MapReduce: simplified data processing on large clusters, Magazine Communications of the ACM, http://dl.acm.org/citation.cfm?id=1327492.
13. Интернет-ресурс: Cloud Big Table, A high performance NoSQL database service for large analytical and operational workloads, https://cloud.google.com/bigtable/ .
14. Интернет-ресурс: Официален Web-сайт - http://hadoop.apache.org/ .