**ТЕМА 4. ИЗВЛИЧАНЕ НА ЗНАНИЯ ОТ ДАННИ (DATA MINING)**

**Съдържание на темата:**

4.1.Характеристика на процеса. Основни понятия.

4.2.Етапи в процеса Data Mining.

**4.1.Характеристика на процеса. Основни понятия**

Терминът Data Mining в превод означава “добиване” или “извличане на данни”. С него се свързват и термина ”интелигентен анализ на данните”.

Откриването на знания (Data Mining), според един от основателите на направлението, е „процес на идентифициране в данните на по-рано неизвестни, нетривиални, практически полезни и достъпни интерпретации на знания, необходими за вземане на решения в различни сфери на човешка дейност“(Gr. Piatetsky-Shapiro). Закономерностите са предварително неочевидни, което означава че не могат да се получат със стандартните методи на обработка или по експертен път.

В търсене на знания се претърсват фирмените бази данни, архивите на наличните компютърни системи, текстове на електронни документи, таблици и други набори от данни. За разлика от извличането на данни посредством заявки към база данни, извличането на знания от данни генерира предварително неизвестна и потенциално полезна информация.

DM е стандартизиран през 1996 г. процес (CRISP- DM: CRoss Industry Standard Process for DM). Той е **неутрален по отношение на сферата на приложение**, което е негово съществено предимство. Почти всеки бизнес има потенциал за използване на технологията. Основните причини за прилагането й са:

* ограниченият достъп до специалисти в различните предметни области, които могат да дадат практически решения на съществуващите проблеми. Te могат да бъдат заместени от системи, извличащи закономерности от данните и тези модели да се използват за прогнозиране, класификация и др.
* наличието на големи обеми от реални данни, организирани в бази (складове) от данни. В тях се съдържат ценни за бизнеса връзки и зависимости, които могат да се получат с прилагане на методите на DM.
* достъп до информационни технологии за прилагане на методите на DM, които са съвместими с използваните във фирмите. Т.е. методите за извличане на знания не влизат в противоречие с използваните системи във фирмите, а само обработват натрупаните данни.

Извличането на знания се използва предимно от компании със силен акцент върху потребителите - търговия на дребно, финансови, комуника ционни и маркетингови организации. Дава се възможност на тези компании да определят връзката между "вътрешни" фактори като цена, позициониране на продукта или уменията на персонала, както и "външни" фактори като икономически показатели, конкуренция и др.

В смисълa на DM се използват и понятията:

-интелигентен анализ на данни;

-изследване на знания в бази от данни (Knowledge Discovery in Databases- KDD). Но DМ е само стъпка от този процес и не бива да се отъждествява с него.

Data mining включва интеграция от техники от различни дисциплини като БД, статистика, машинно обучение, разпознаване на шаблони, НМ, визуализация, извличане на информация, обработка на образи и сигнали и анализ на пространствени и времеви данни, т.е. технологията е мутидисциплинарна. Откритите знания се прилагат за вземане на решения, контрол на процеси, управление на информацията, обработка на заявки и др.

Резултатите от ДМ в голяма степен зависят от нивото на подготовка на данните. Около 75-80% от работата по ДМ се състои в набиране и подготовка на данните. Извличането на полезни знания е невъзможно без разбиране същността на данните. Както и неграмотното прилагане на някои инструменти за анализ може да доведе до големи загуби.

Преди да се използва технологията е необходимо да се обмислят нейните ограничения и проблеми и да се разбере какво тя може да постигне. Например, ДМ не може да отговори на незададени въпроси или да замени човека- аналитик.

**Характеристики на данните, използвани в Data Mining**

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 20⁰C е три пъти по-ниска от температура 60⁰C). 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 упоребява 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 данните.

Възможни са и други класификации на данните според други техни особености или целите на изследването, за което се използват.

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.

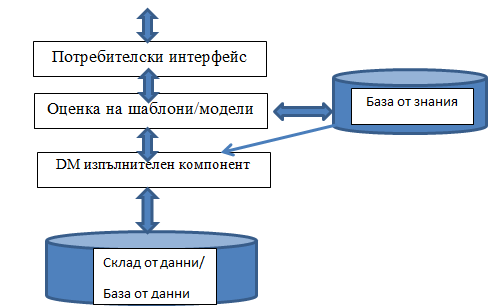
Според структурираното им представяне данните се разделят на следните категории:

- Структурирани 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 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.

На фиг. 4.1. е представена схема на типична DМ система. Основните й модули присъстват в БИС.



**Фиг. 4.1. Схема на типична DМ система**

Знанията в базата от знания (БЗ) са за предметната област, в която работи системата. Използват се да ръководят търсенето или оценката на значимостта на моделите. Знания, като очакванията на потребителя, които може да се използват за оценка на значимостта на модела, също може да бъдат включени. Други примери за знания от предметната област са допълнителни ограничения или прагове и метаданни (описвaщи данни от много хетерогенни източници).

**Data mining** изпълнителен компонент е най-важната част на DМ системата. Състои се от набор функционални модули за задачи като асоцииране и корелация, анализи, класификация, прогнозиране и др. Инструментите намират закономерностите самостоятелно и построяват хипотези за взаимовръзките им, т.е. **не се работи с предварително зададени хипотези**, както напр. при OLAP или статистическите методи.

Основна особеност на ДМ е съчетанието на математически инструментариум с последните достижения в сферата на информационните технологии, т.е. на строго формализираните методи и тези на неформалния анализ – количествен и качествен анализ на данните.

**Модулът за оценка на моделите** обикновено съдържа метриките на значимост (интерес) и фокусира търсенето към интересни модели. Той може да използва праговете на интерес за филтриране на откритите модели.

**Потребителският интерфейс** е модул за комуникация между потребителите и DМ системата, позволявайки интерактивност за потребителя чрез специфициране на DМ заявката, осигуряване на информация за фокусиране на търсенето. В допълнение този модул позволява на потребителя да преглежда схемите на БД и СД или структурите от данни, да оцeнява получените модели и ги визуализира в различни форми.

При анализа на данни по принцип не е възможно да се използва цялата интересуваща ни съвкупност от обекти. Достатъчно е да се разглежда част от нея, т.е. извадка, за да се получи интересуващата ни информация. За разлика от статистическите подходи, тук теоретично **не се изисква строго определено количество данни**. Това може да доведе до получаване на недостоверни модели. За това получените модели трябва да се проверят за адекватност с тестови данни.

Друга особеност на ДМ е, че работи върху **реалните значения** на данните, а не върху приближени/осреднени величини.

Извличането на знания винаги се предшества от извличане на данни в избран обем.

**Решаваните задачи** чрез DM са главно от следните групи:

- Класификация. Представлява основна и най-разпространена задача за анализ на данни. По същността си е отнасяне на обектите към предварителни зададени класове. Съставят се модели, описващи даден клас. Моделът е основан на анализ на набор от обучаващи данни. Полученият модел може да се представи в различни форми като асоциативни правила, дървета от решения, математатически формули или невронни мрежи. Типични алгоритми, извършващи класификация, са дървета на решенията, методи на опорните вектори, бейсовска класификация, метод на най-близкия съсед, невронни мрежи и др.

- Клъстеризация (групиране). Идентифицират се групи от обекти подобни един на друг и отличаващи се от други групи обекти с определени характеристики. Като цяло, етикетите на класовете не се представени в данните на обучение, просто защото те не са известни отначало. Използва се принципа за максимизиране на подобието в класа и минимизиране на подобието между класовете.

- Прогнозиране. Представлява подобна на класификацията задача, но целта е да се предскаже стойност на числова променлива, а не да се разпознае клас. Откриват се зависимости в «исторически» данни, върху които може да се направи изпреварващо отразяване на бъдещето.

- Асоцииране – откриване на асоциации от свързани събития или стоки в набора от данни. Например, закупуването на една стока се асоциира с покупката на друга- покупка на очила се асоциира с купуване на калъф за тях.

- Oткриване на последователности от свързани събития. Нарича се още «намиране последователност от шаблони». Позволява намиране на времеви закономерности между транзакциите. Тази задача е подобна на асоциирането, но нейната цел е установяване на закономерности не между едновременно настъпващи събития, а между свързани във времето събития (с определен интервал от време). Последователността определя висока вероятност от верига, свързани във времето събития. Например, закупуването на компютър обикновено е последвано от покупка на принтер или аксесоари.

- Редукция на данни. Представлява процес на консолидиране на голям брой променливи (случаи) в по-малко множество.

- Визуализация. Подпомага откриването на зависимости в данните чрез графичен анализ. Това включва разглеждане на всяка променлива поотделно, както и разглеждане на връзките между променливите. За числови променливи, се използват предимно хистограми и плотови диаграми. Те позволяват да се види разпределението на техните стойности, да се открият отклоненията или да се намери друга информация, която е от значение за задачата. За категорийни променливи се използват главно лентови диаграми.

- Анализ на извънредни стойности и намиране на редки случаи, които се различават съществено от нормите. БД може да съдържа обекти които не отговарят на общото поведение или модел на данните. Тези данни за обекти се наричат извънредни (екстремни) стойности. Повечето DМ методи изхвърлят тези данни като шум или изключения. Обаче в някои приложения, като засичане на измами, редките събития са по-интересни от обичайните. Този вид анализ намира приложение най- често в борбата с престъпления с кредитни карти, анализ на производството и неоторизиран достъп до мрежи.

Решаването на тези задачи позволява изпълнение на действия като: прогностично моделиране, анализ на връзките, сегментация на данните и т.н.

Методите, прилагани за решаване на горните задачи, се разграничават в 2 групи- контролирани и неконтролирани методи. Контролираните са тези, използвани в класификацията и прогнозите. За контролираните методи трябва да имаме налични данни, в които е известен входа и съответния му изход. Например данните може да се разглеждат като двойки (x,y), като х е входът, а у е коректният изход. Тези данни служат за обучение «с учител», от които при класификация или прогноза, алгоритъмът "научава" връзката между входа и изхода и може да я приложи в друга извадка от данни или за валидиране. Този модел може да се тества (провери) с друга извадка от данни с известни резултати. Обобщено, при тези методи се използват 3 извадки от данни- за обучение, за валидация и за тестване. Данните за обучение са най-големи по обем. От тях се извеждат (създават) различните модели. Данните за валидация се използват за оценка на моделите и така те могат да бъдат сравнявани. Тестовите данни се прилагат за оценка на работата на избран модел с нови данни.

Неконтролираните алгоритми за обучение са тези, при които няма изходна променлива, която да се класифицира или предсказва. Следователно, не съществува "обучение" от случаи, в които да е известна изходна/резултатна променлива. Примери за неконтролирани методи на обучение («без учител») са асоциативните правила, техниките за клъстеризация и др.

Трябва да се отбележи, че нито един метод не осигурява решение на целия спектър от задачи на ДМ.

Някои от основните конвенционални методи за DM са разгледани в следващи теми.

Само малка част от генерираните шаблони/модели са интересни за даден потребител. Шаблоните са интересни, ако са лесно разбираеми от човека, ако са валидни за нови или тествани данни с някаква степен на сигурност или са потенциално полезни и нови, също ако валидират хипотези, които потребителят е търсел за потвърждение. Интересните шаблони представят знания. Често не е реалистично и ефективно да се генерират всички възможни шаблони. Вместо това зададените от потребителя ограничения и интереси трябва да се използва за фокусиране на търсенето.

Получените модели се представят чрез дърво на решенията, асоциативни правила, математически формули и др.

**4.2.Етапи в процеса Data Mining**

Процесът по откриване на знания е итеративен и интерактивен. Основните му етапи включват (фиг. 4.2):

1.Разбиране на приложната област и дефиниране на целите на изследванията, както и съответните изисквания на потребителя. След завършването на етапа тези знания трябва да бъдат превърнати в дефиниции на задачи за сондиране на данни и да се състави предварителен план за това, как тези цели могат да бъдат постигнати.

2. Набиране/получаване на данните. Може да се използват данните от различни бази от данни, които се разполагат в общо хранилище (слад от данни). Базите от данни може да бъдат вътрешни за фирмата (напр. покупки на стоки, доставки и т.н.) или външни (напр. кредитни рейтинги).

На този етап е необходимо да бъдат идентифицирани проблеми, свързани с качеството на данните, да бъдат намерени интересни подмножества на данните, за да бъдат формирани първоначални хипотези за скритата в данните информация.

3. Почистване и предварителна обработка на данните.

Включва се проверка дали данните са в добро състояние, тъй като „замърсените“ данни може да доведат до грешни резултати или провал на процедурата по DM. Решава се как да се борави с липсващи данни и как да се „изглади“ шума в данните. Oпределя се стойностите на величините в разумни граници ли са? Има ли очевидно големи различия в стойностите на една променлива? Ако е избран контролиран метод за DM, от данните се подготвят 3-те множества: за обучение, валидация и тестване.

По много и различни причини някои стойности в таблиците от БД/СД остават незапълнени или са изтрити. При липсващи данни може да се приложи някой от следните подходи:

- Заменяне на липсващата стойност с най-вероятната стойност, на базата на цялата налична в таблицата информация. Възможно е да бъдат използвани множество методи и модели за определяне на най-вероятната стойност, като е желателно методът да може да се прилага автоматично и в същото време резултатът да е лесно интерпретируем.

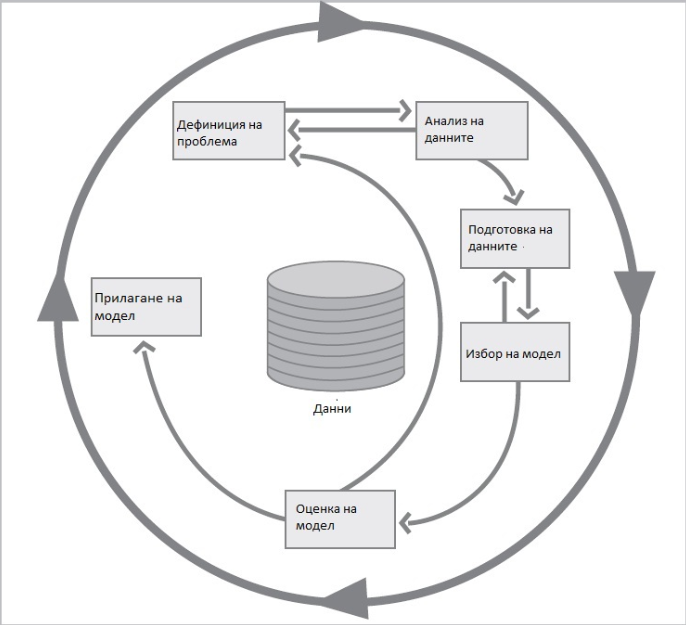
- Заменяне на липсващата стойност със средната за колоната;

- Използване на етикет за маркиране на липсващата стойност;

- Запълване на липсващата стойност ръчно, на база експертна преценка за размера й;

- Игнориране на колоната или реда в таблицата, което е свързано със загуба на информация и пропускане на интересни случаи.

4. Определя се задачата за DM- класификация, прогноза или др.

**

**Фиг.4.2. Етапи на процеса DM**[[1]](#footnote-1)

5. Избира се и се прилага метод за DM (невронни мрежи, дървета от решения или др.). Методът представлява прилагане на алгоритъм за търсене на интересуващи ни модели, връзки и зависимости в БД и други информационни хранилища. Те се представят в специална форма, като асоциативни правила, дървета, клъстери и др. Тъй като някои методи имат специфични изисквания към формата на данните, често се налага връщане към етапа за предварителна обработка на данните.

Прилагането на метода за DM е типично итеративен процес- опитват се множество варианти и често дори с един и същ алгоритъм като се избират различни променливи или настройки на алгоритъма.

6. Интерпретиране и оценка на модела (шаблона). Прави се избор на най-добрия алгоритъм и където е възможно се прилага тестовата извадка от данните, за да се получи представа колко добре се изпълнява.

Дефинират се количествени мерки за оценка на моделите (шаблоните). Моделите се разглеждат като знания, ако превишават зададен праг на полезност. Моделите се оценяват доколко отговарят на поставените бизнес цели, колко е времето за изпълнение, съответствие на поставените критерии и др. Генерираният модел води до разбиране нa същността на изследваните процеси, тъй като намира обяснение за връзката между получаването на определени резултати и налични факти. Може да се включи визуализация на извлечените зависимости. Те трябва да са валидни за нови данни с известна степен на сигурност.

6. Прилагане на модела. Изготвя се план за използване на резултатите от анализа на данните, а след това постигнатото се отчита.

Разгледаният процес се определя като **структурно** извличане на знания.

**Литература:**

1. Атанасова Т., Интелигентни компютърни системи, изд. ”Наука и икономика”, Варна, 2011 г.
2. Щерев Й., Анализ на данни, изд. „Фабер“, В.Търново, 2010
3. Peter C. Bruce; Galit Shmueli; Nitin R. Patel [Data Mining For Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel® with XLMiner®, Second Edition](https://www.safaribooksonline.com/library/view/data-mining-for/9780470526828/), Published by John Wiley & Sons, 2010

**Интернет адреси:**

1. Kurt, T. An Introduction to Data Mining, <http://www.thearling.com/dmintro/dmintro_2.htmS>.

The CRISP-DM User Guide <http://lyle.smu.edu/~mhd/8331f03/crisp.pdf>

Data Mining- A Short Introduction http://www.scribd.com/doc/47842005/week-13-Datamining-Association-rules

1. The CRISP-DM User Guide <http://lyle.smu.edu/~mhd/8331f03/crisp.pdf> [↑](#footnote-ref-1)