Лекция 1. Извличане на закономерности от данни – основни понятия

1.1. Какво е "извличане на закономерности от данни"?

Бързият прогрес в развитието на средства за придобиване и съхраняване на данни, наблюдаван в последните години, доведе до бързото нарастване на големи бази от данни в практически всички области на човешката дейност - от данни за транзакции в един супермаркет, детайли за телефонни разговори и разнообразни данни за населението, до снимки на астрономическите обекти, бази от медицински записи и бази от ДНК данни. Това огромно количество събрани данни значително превишава възможността на човека те да бъдат ефективно използвани без помощта на специализирани мощни средства за анализ на данни. Като резултат, данните, събрани в огромните бази от данни, са превърнали в "гробници от данни" – архиви, които почти не са посещавани от никого. Така се стигна ситуацията, описвана като "богата на данни, но бедна на информацията". Тази все по-разширяваща се пропаст между данните и информацията доведе до създаването на научна дисциплина *"Извличане на закономерности от данни (ИЗД)" (Data Mining)*, която цели да превърне гробниците от данни в "златни мини", от които се добиват знания.

Извличането на закономерности от данни се състои в анализ на (често много големи) множества от наблюдавани данни с цел да бъдат открити в тях неочаквани зависимости, или те да бъдат обобщени по нови начини, които са разбираеми и полезни за притежателите на данни.

Изведените от данни зависимости и обобщения често се наричат *модели* или *шаблони*, които се представят чрез линейни уравнения, правила, клъстери, графи, дървета и т.н. Думата "*наблюдавани данни*" е използвана в по-горната дефиниция като противоположност на "експериментални данни". Обикновено ИЗД работи с данни, които са били събрани за цели, отлични от ИЗД-анализа, което означава, че поставените от ИЗД цели не оказват никакво влияние на стратегията за събиране на данни. Това е една от характеристиките, по които ИЗД значително се отличава от статистиката, където данни често се събират чрез използване на ефективни стратегии за получаване на отговори на специфични въпроси. По тази причина ИЗД често се нарича "*вторичен*" анализ на данни.

В дефиницията на ИЗД се набляга на *големия размер* на бази данни. При работата с малки множества от данни могат да бъдат използвани методи на класическия изследователски анализ, прилаган от статистиците. В случая на големи бази данни възникват нови проблеми. Някои от тях са свързани с начина за ефективно съхраняване и намиране на данни, но другите се отнасят за такива фундаментални въпроси, като как да бъдат избрани характерни представители на данни, как тези данни могат да бъдат проанализирани за приемливо време, как да се разбере, дали някоя намерена зависимост отразява действителната реалност, а не е само резултат от някакво случайно съвпадение само в определена част от данните.

Важната характеристика на извлечените зависимости и структури са тяхната *новост*. Ясно е, че степента на новото трябва да се измерва относително априорните, базови знания на потребителя. За съжаление, много малко ИЗД алгоритми взимат под внимание подобни знания. Съществуващите формални методи за определяне на степента на новост на намерените закономерности ще бъдат разгледани по подробно в една от следващите лекции.

Макар че степента на новост е едно важно свойство на зависимости, които ние търсим в данните, само тя не е достатъчна да квалифицира една зависимост като "знание". За целта тя трябва да бъде и *разбираема*. Например, прости зависимости често са по-разбираеми от колкото сложни и по тази причина могат да бъдат предпочетени от потребителя при всички други равни условия.

1.2. Естеството на данни

Данните е едно множество от измервания, направени в определена среда или върху определен процес. В най-простия случай имаме колекцията от обекти, като всеки от тях се характеризира с множеството от *p* измервания, еднакви за всички обекти. В този случай такава колекция може да се разглежда като *n x p матрица от данни*. *n* редове представят *n* обекта, върху които бяха направени измервания (например, медицински пациенти, притежатели на кредитни карти или астрономическите обекти от типа на звезди или галактики). В различни научни дисциплини (контексти) тези обекти се наричат *индивиди, случаи, единици, обекти, примери* или *записи*.

Другата размерност на тази матрица съдържа множеството от *p* измервания, направени върху всеки от обектите. Най-често се предполага, че върху всеки обект е направено едно и също множество от измервания, макар че това не е задължително (например, върху различни пациенти могат да бъдат извършени различни медицински тестове). Тези *p* колонки на матрицата на данни се наричат (пак в различни контексти) *променливи, признаци, атрибути* или *полета*. Във всички ситуации идеята е една и съща: тези имена обясняват измерването, представено чрез всяка от колонките.

Пример 1.1. Бюрото по преброяване на населението на САЩ събира данни за населението на страната на всеки 10 години. Част от тази информация става публично достъпна, като всички данни, позволяващи индивидът да бъде идентифициран, се изтриват. Данните са достъпни във вид на 5% и 1% извадки (обърнете внимание, че дори 1% извадка съдържа около 2.7 милиона записи). Данните съдържат десетки променливи, от типа на възраст, доход, нивото на образование и т.н. На фиг. 1.1 е представен пример на подобни данни. Обърнете внимание, че таблицата съдържа различни типове променливи – непрекъснати (например *Age* (възраст)) и символни (например *Marital Status* (семейно положение)). Някои стойности липсват – това е често явление в бази от реални данни. По-сложен въпрос е с наличието на шума – например, дали доходът (income) на човека с ID 248 е действително 100 000$ или това е грешка при въвеждане?

ID	Age	Sex	Marital Status	Education	Income
248	54	Male	Married	High school graduate	100000
249	??	Female	Married	High school graduate	12000
250	29	Male	Married	Some college	23000
251	9	Male	Not married	Child	0
252	85	Female	Not married	High school graduate	19798
253	40	Male	Married	High school graduate	40100
254	38	Female	Not married	Less than 1st grade	2691
255	7	Male	??	Child	0
256	49	Male	Married	11th grade	30000
257	76	Male	Married	Doctorate degree	30686

Фиг.1.1. Пример на данни, събирани от Бюрото по преброяване на населението на САЩ

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

Една типична задача за подобен тип данни е намирането на зависимости между различни променливи. Например, би било интересно да се види доколко добре доходът на човека може да бъде предсказан от стойностите на други променливи. Интересна е и задача да бъдат намерени естествени групи от хора с близки показатели.

Разгледаната по-горе $n \times p$ матрица на данни е често само идеализацията или опростяване на ситуацията, наблюдавана на практиката. Например, в едно множество от медицински записи една и съща променлива (например *кръвно налягане*) може да има множество от стойности, като всяка стойност отразява измерването, направено в различен ден. Някои пациенти могат да имат данните представени във вид на *изображение* (рентгенова снимка, например). Възможни са и данните във вида на *текст* (диагнози и коментарии на специалиста). Освен това, са възможни и йерархични релации между пациенти в термините на доктори, болници и географското разположение.

Очевидно е, че колко по-сложни са структури данни, толкова по-сложни ще бъдат ИЗД модели и алгоритми, които ние трябва да използваме. Обаче, макар че много от реалните множества от данни не пасват точно на описания "плосък" прост формат във вид на матрицата на данни, повечето от съдържащата се в тях информация може по принцип да бъде запазена и в "плоския" формат (чрез подходящо дефиниране на p променливи), макар че това често води до определени загуби на структури, вградени в тези данни. Не зависимо от този факт, от тук нататък ще подразбираме (ако не е казано явно обратното), че наблюдаваните

данни са съществуват във вид на *n x p* матрица на данни, като ще смятаме, че и *n* и *p* могат да бъдат много големи. В различни контексти матриците на данни се наричат с различни имена, сред които *множество от данни, обучаващите данни, извадка, база данни* и т.н.

Пример 1.2. Текстовите документи са много важен източник на информацията и ИЗД методи могат да помогнат за извличане на полезни текстове от големи колекции от документи (от типа на Web). Всеки документ може да се разглежда като поредицата от думи и знаци за пунктуацията. Типичните задачи за анализиране на текстови бази данни са класификацията на документи по предварително определени категории, клъстеризация на подобни документи и намирането на документи, които най-добре отговарят на спецификациите, зададени в заявката.

Една колекция от текстови документи може да се представи като матрицата на данни, в която редове представляват документи, а колонки – думи. Елементът *(d, w)*, съответстващ на документа *d* и дума *w*, може да бъде представен чрез число, отразяващо колко пъти *w* се среща в *d*, или само чрез *1*, ако *w* присъства в *d*, или *0* – в противен случай. Макар че този подход води до загубата на информацията за подреждането на думи в документа (и, следователно, до голяма степен семантичния контекст), той представлява достатъчно добре съдържанието на документа. За една колекция от документи броят на редове е броят на документи, а броят на колонки – броят на различни думи. По този начин, една голяма многоезична колекция от документи може да има милиони редове и стотици хиляди колонки.

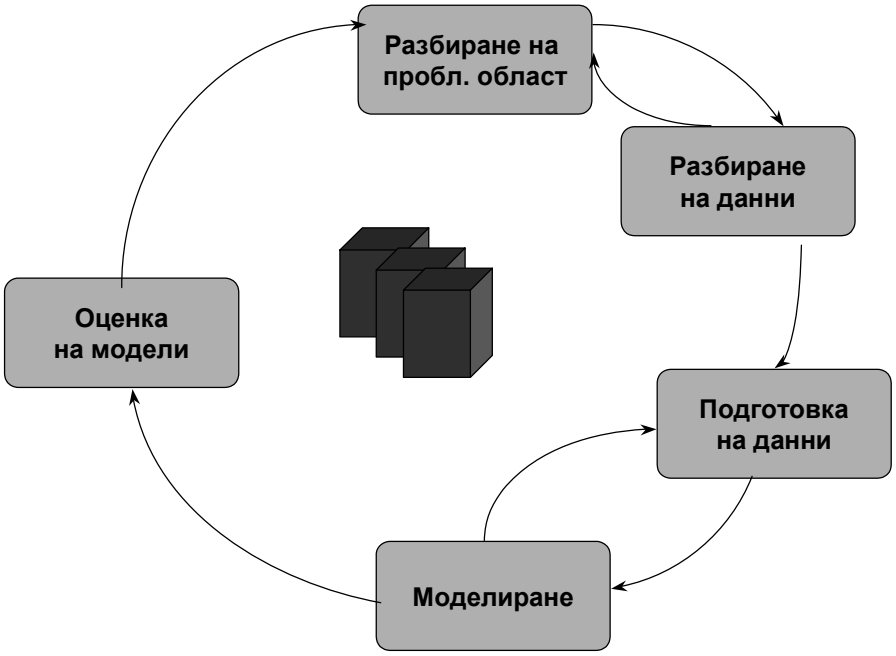
Пример 1.3. Един друг много разпространен тип на данни е *данни за транзакции (сделки)*, от типа на списъка на покупки в магазина, като всяка транзакция се описва с дата, идентификатор на покупателя и списъка на закупени неща заедно с тяхната цена. Подобният пример е файл-архив с записи на използвания на Web (Web transaction log), който съдържа последователността от тройки (user id, web page, time), означаващи че определен потребител имал достъп до указаната страница в определено време. Създателите и притежатели на Web сайтове често са заинтересувани да разберат типични начини, по които потребители се предвижват по техните сайтове.

Както и в случая с текстови документи, можем да трансформираме множеството от транзакции в матрицата на данни. За целта представете една разпръсната матрица, в която един ред отговаря на отделния потребител, а една колонка – на отделната Web страница или покупка. Елементите на подобната матрица могат да бъдат двоични (например, указващи дали конкретният потребител някога посещавал конкретната страница) или представени чрез цели числа (например, показващи колко пъти потребителят посещавал страницата). Ясно е, че подобно "плоско представяне" води до загуба на значителната порция от информация, свързана с развитие процеса във времето (например, в какъв порядък и кога бяха закупени нещата), както и до информация за структурни връзки между отделните елементи (например, йерархии на категориите на закупените неща, връзки между Web страниците и т.н.). Обаче, представянето чрез плоската матрица от данни позволява да дефинираме разстояния между потребители чрез сравнение на техни *p*-мерни вектори на използването на Web страници, което на свой ред позволява потребителите да бъдат клъстеризирани (групирани) на базата на техните Web шаблони.

1.3. Модел на ИЗД процеса

И така, ИЗД е една интердисциплинарна област на научни изследвания, представляваща пресичането на няколко различни научни дисциплини, такива като математическата статистика, машинно самообучение, експертни системи, бази данни, информационно-търсещи системи, разпознаване на образи и др. В научната литература освен термина Data Mining често се среща и терминът Knowledge Discovery in Databases (KDD) – извличане на знания от бази данни. Макар че някои от автори разглеждат Data Mining само като една част от по-общ процес на KDD, в последно време се налага тенденцията и двата термина да се разглеждат като синоними – в настоящия курс ние ще следваме тази тенденция.

Извличането на закономерности от данни е един интерактивен процес, който започва с разбирането и дефиницията на задача, подлежаща на решаване, и завършва с анализа на получените резултати и стратегията за тяхното използване на практиката. Адаптирайки методологията за ИЗД изследвания - CRISP-DM (Cross-Industry Standard Process Model of Data Mining) [Chapman et al. 99], един цикъл на изследвания в областта на ИЗД може да се представи като процес, състоящ от пет основни етапа (виж фиг.1.2). Последователността на етапите не е строга: често се налага преминаването напред – назад между отделни етапи в зависимост от резултата на всеки от етапите.



Фиг. 1.2. CRISP-DM модел на процеса на ИЗД

- *Разбирането на проблемната област* е началния етап, който се фокусира върху разбирането на целите на изследванията и съответните изисквания от гледната точка на потребителя. След завършването на етапа тези знания

трябва да бъдат превърнати в дефиниции на ИЗД задачи и да бъде съставен предварителния план, как тези цели могат да бъдат постигнати.

- *Разбирането на данни* започва с първоначално събиране на данни и се продължава с дейностите, целящи задълбочаване на знания на изследователя за естеството на данните. На този етап е необходимо да бъдат идентифицирани проблеми, свързани с качеството на данни, да бъде изградено първоначално мнение за характера на данни, да бъдат намерени интересните подмножества на данни, за да могат да бъдат формирани първоначални хипотези за скритата в данни информация.
- *Подготовката на данни* покрива всички дейности по създаване от първоначални "сурови" данни на крайното множество от данни (т.е. данни, които ще бъдат използвани от моделиращите средства). Етапът на подготовката на данни често се налага да бъде изпълняван многократно и по различно време. Задачите по подготовката на данни включват в себе си избор на таблиците от данни, нейните атрибути и отделни записи, както и трансформация и изчистване на данни.
- *Етапът на моделиране* се състои от избора и прилагане на различни техники за моделиране, целящи извличане на закономерностите от данни. Параметрите на моделите се калибрират до свои оптимални стойности. Тъй като някои модели имат свои специфични изисквания към формата на данни, на този етап често се налага връщането към етапа за подготовката на данни.
- *Оценката на модели* е необходима за по-задълбочено разбиране на създадените модели от гледната точка не само на изследователя, но и на потребителя. Важно е да бъдат внимателно прегледани всички стъпки, изпълнени при създаването на конкретния модел с цел да се увери, че те постигат поставените пред тях цели. В края на този етап се приеме решение за използване на получените в процеса на ИЗД резултати.

В следващите раздели ще се спрем по-подробно на описаните по-горе основни етапи на процеса за извличането на закономерности от данни.

1.3.1. Разбиране на проблемната област

Разбиването на основната задача за разбиране на проблемната област на подзадачи, решавани на този етап, както и съответните резултати, са представени на фиг. 1.3.

Цели на изследването

Първата задача на изследователя е да разбере, какво точно иска от него потребителят. Често потребителят има множество конкуриращи цели и ограничения, които трябва да бъдат по подходящ начин балансирани. Целта на изследователя още в началото да бъдат изяснени всички важни фактори, които могат да повлияят на резултата от изследването. *Недостатъчното внимание, отделено на тази стъпка, може да доведе до загуба на значителни усилия за получаване на правилни отговори на*

неправилно поставени въпроси.

Оценка на ситуацията Тази задача включва внимателното проучване, намиране и оценяване на наличните ресурси, ограничения, предположения и други фактори, които трябва да бъдат взети под внимание при определяне на ИЗД цели и съставяне на плана на изследването.

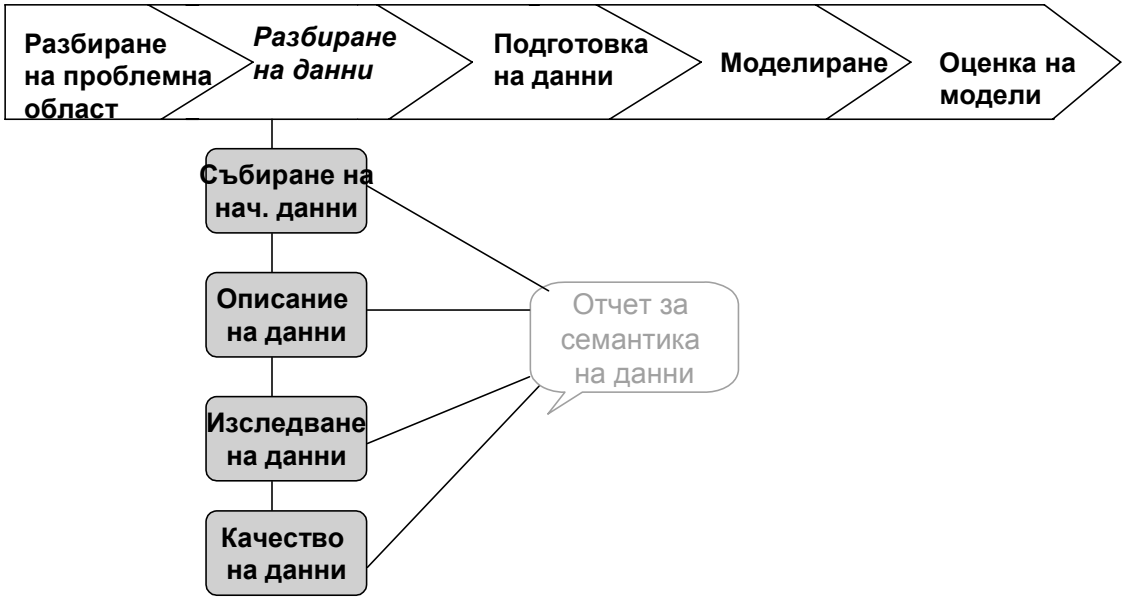
Определяне на ИЗД цели Целите на изследването се определят и се поставят в термините на проблемната област. ИЗД-цели определят целите на изследването в *техническите термини*. Например, целта на изследването може да бъде: "Да бъдат увеличени продажби по каталог за вече съществуващи покупатели", докато ИЗД цел може да бъде "Да бъде предсказано колко кухненски уреда ще бъдат закупени от един покупател на базата на данни за неговите покупки за последните три години, демографската информация (възраст, заплата, град и т.н.) и цената на стоките".

План на изследването Целта е да бъде създаден план за постигането на формулираните по-рано ИЗД цели. Планът описва множество от стъпки, които трябва да бъдат изпълнени в останалата част от изследването, както и начален избор на подходящи алгоритми.

Фиг. 1.3. Етап на разбиране на проблемната област

1.3.2. Разбиране на данни

Разбиването на основната задача за разбиране на данни на подзадачи, решавани на този етап, както и съответните резултати, са представени на фиг. 1.4.



Фиг. 1.4. Етап на разбиране на данни

Събиране на начални данни Тази задача включва достъп (или събиране) на данни от всички налични ресурси, идентифицирани на етап 1.3.1. Началната колекция включва всички видове данни, необходими за разбиране на характера на данни, използвани при изследването. Ако данните се извличат от различни източници, процедурите по тяхното интегриране трябва да бъдат извършени или на тази стъпка, или на стъпката за подготовката на данни.

Описание на данни Описанието на данни включва техния формат, броя на записи във всяка от таблиците, имената и типове на атрибути и т.н.

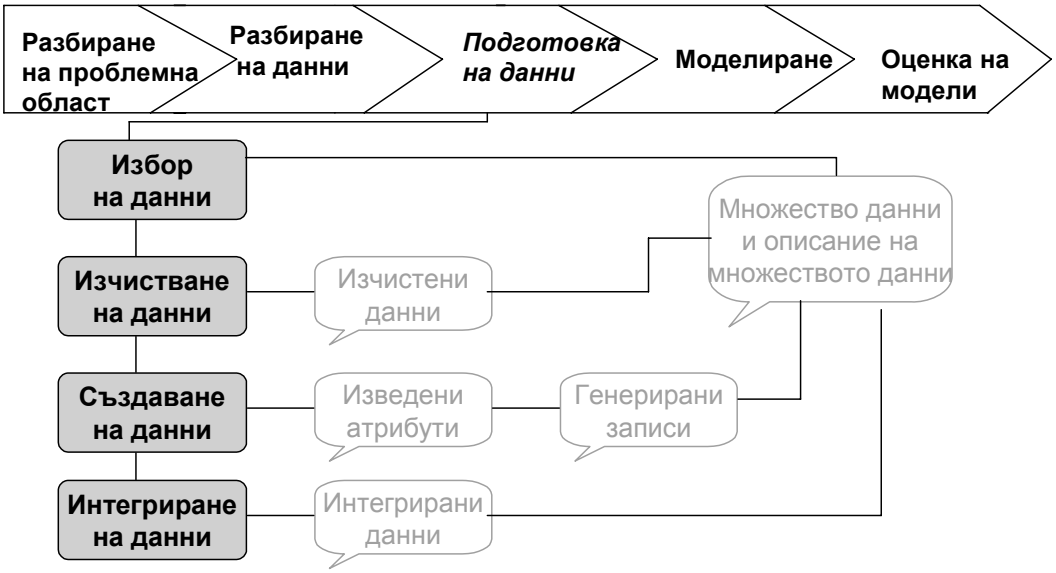
Изследване на данни На този етап изследователят трябва да направи първоначални хипотези за характер на данни чрез анализ на разпределение на ключови атрибути, изследване на релации между двойки на някои от атрибути и друг прост статистически анализ на някои извадки от първоначални данни.

Определяне на качеството на данни На този етап трябва да бъдат намерени отговори на такива въпроси като: дали данните са пълни? Дали те са коректни или съдържат грешки? Има ли липсващите данни? Ако да, как те са

представени? Отговорите на всички тези въпроси се документират в отчета, описващ семантиката на данни.

1.3.3. Подготовка на данни

Разбиването на основната задача за подготовка на данни на подзадачи, решавани на този етап, както и съответните резултати, са представени на фиг. 1.5.



Фиг. 1.5. Етап на подготовката на данни

Избор на данни На тази стъпка трябва да се вземе решение, кои данни ще се използват за анализа. Критерият включва релевантността на данни за ИЗД цели, качеството на данни и техническите ограничения на обем и типове данни, които могат да бъдат обработени от наличните ИЗД алгоритми.

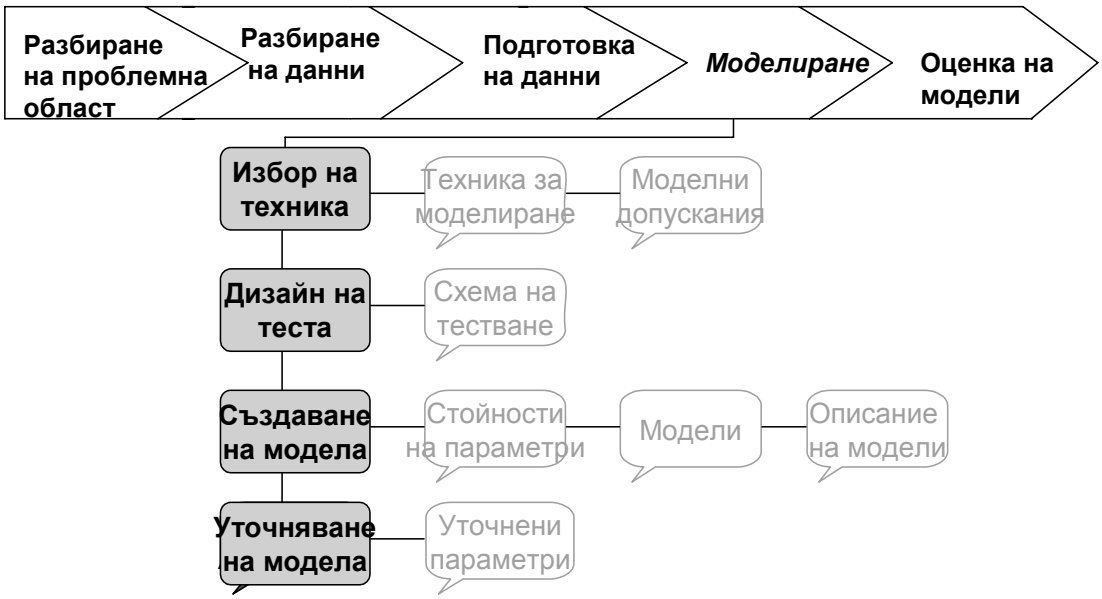
Изчистване на данни Целта е да бъде повишено качеството на използваните данни. Това може да бъде постигнато чрез, например, избор на "чисти" подмножества от данни или чрез прилагане на съответните техники за премахването на шума и обработка на пропуснатите стойности.

Създаване на данни Тази задача включва конструиране на нови данни чрез създаване на нови (изведени от първоначални) атрибути и/или различни трансформации на данни от типа на тяхното нормализиране, дискретизиране и т.н.

Интегриране на данни На този етап информацията се комбинира от различни източници (таблици) с цел да бъдат създадени нови записи или стойности.

1.3.4. Моделиране

Разбиването на основната задача за намиране на модели на данни на подзадачи, решавани на този етап, както и съответните резултати, са представени на фиг. 1.6.



Фиг. 1.6. Етап на намирането на модели на данни

Избор на техника за моделиране На тази стъпка трябва да бъде избрана една конкретна техника за моделиране, която ще бъде приложена първа. Ако се прилагат множеството от техники, тази стъпка трябва да бъде изпълнена за всяка от техники. Не всички алгоритми са приложими за всяка от ИЗД задачи (тяхното описание е приведено в следващия раздел), освен това не всички те са достъпни. Така че избраната техника не винаги е най-добрата от всички възможни. Освен това, трябва да се помни, че повечето от техники за моделиране се базират на неявни предположения за характера на данни, тяхното качество и формат, така че всички тези неща трябва да бъдат предварително проверени преди избора и прилагането на съответната техника.

Дизайн на схемата за тестване Преди построяване на модела е необходимо да бъде създадена схема за проверка на неговото качество. Например, при класификационни задачи в качеството на критерия за качество на модела често се използва процентът на грешки. По тази причина схемата на тестване трябва да определя как

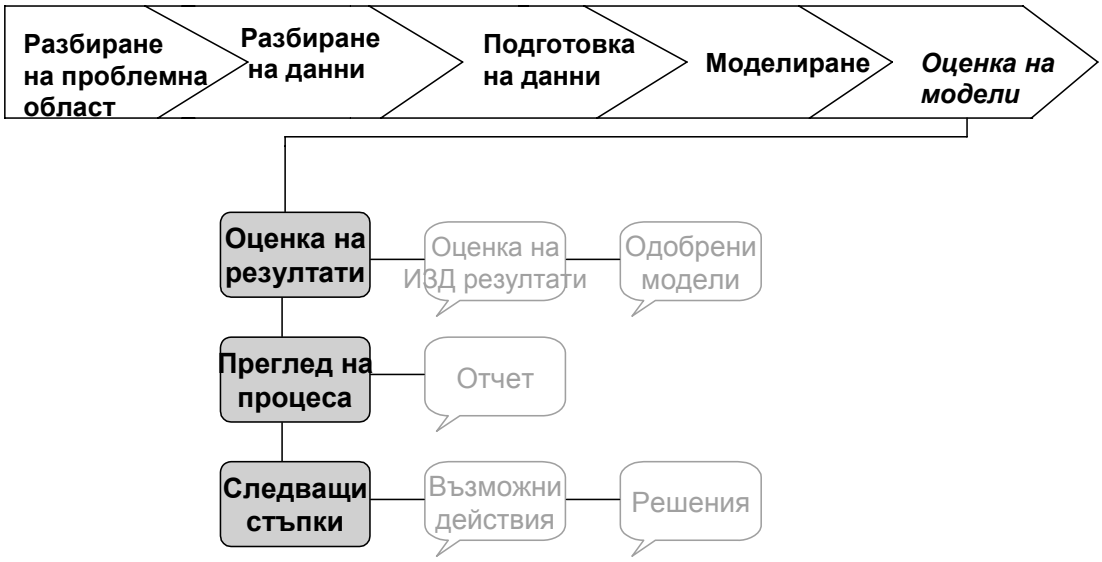
данните трябва да бъдат разбити на обучаващо и тестово множества, като моделът ще се създава върху обучаващото множество и се тества върху тестовото.

Създаване на модела На тази стъпка избраната техника за моделиране се прилага към подготвеното множество от данни с цел създаване на един или няколко модела. Този процес включва настройка на множество от различни параметри на избраната техника.

Уточняване на модела На тази стъпка моделите се оценяват от ИЗД гледна точка и се избират оптимални параметри на модела.

1.3.5. Оценяване на модели

Разбиването на основната задача за оценяване на модели на подзадачи, решавани на този етап, както и съответните резултати, са представени на фиг. 1.7.



Фиг. 1.6. Етап на оценяване на модели

Оценка на резултати На предишните стъпки се оценяваха такива фактори на модела като неговата точност и общност. На тази стъпка се оценява степента, до която моделът отговаря на целите, поставени от потребителя, и се прави опит да се определи, дали съществуват някакви причини от проблемната област, правещи модела неприложим. Резултатите трябва да бъдат сравнени с оценъчният критерий, дефиниран в началото на изследвания. Ако е възможно, добре е моделът да бъде оценен върху някакво реално приложение.

Преглед на ИЗД процеса На този етап изглежда, че крайният модел удовлетворява критерия за успех. Сега е желателно да бъдат по-задълбочено прегледани всички стъпки на ИЗД процеса, водещи до създаването на този модел, за да се определи, дали не бяха пропуснати някои важни от потребителската гледна точка фактори, които евентуално биха могли да доведат до по-добри модели.

Определяне на следващите стъпки В зависимост от резултатите на оценяване и прегледа на ИЗД процеса се определят следващите стъпки на изследването. На този етап трябва да бъде взето решение, дали изследването трябва да бъде прекратено или да бъдат направени следващите итерации на ИЗД процес.

1.4. Естеството на закономерности: типове на ИЗД задачи

В този раздел ще се опитаме да отговорим на въпроса какви типове зависимостите могат да бъдат открити в процеса на извличане на закономерности от данни. За целта ще разгледаме различни типове ИЗД задачи, решаване на които води до получаване на интересуващите ни зависимости (модели или шаблони).

1.4.1. Описание и обобщение на данни

Задачата за *описание и обобщение на данни* (Data Description and Summarization) цели намирането на стегнато описание на основните характеристики на данни, обикновено в елементарна, обобщена форма. Това позволява на потребителя да получи представа за структурата на данни. Понякога описанието и обобщението на данни може да бъде поставено като самостоятелната задача на ИЗД изследване. Например, един търговец може да бъде заинтересован в намиране на описанието на промени в пазара, разбити по съответните категории. Измененията и разликите в сравнение с предишния период трябва да бъдат обобщени и описани. Този тип на задачи се отнася към най-леките в областта на ИЗД.

В повечето от случаи описанието и обобщението на данни е само една от подзадачи, изпълнявана на ранните стадии на едно ИЗД изследване. В началото на изследването потребителят често не знае нито точната цел на анализа, нито естеството на данни. Началният изследователски анализ на данни може да помогне да бъде разбрана природата на данни, да бъдат построени първоначалните хипотези за скритата в данни информация. Например, разпределението на възрастта на покупатели и тяхно място на живеене може да даде подсказка към това, кои части от групата на покупателите трябва да бъдат обект на нови маркетингови стратегии.

Желателно е описанието и обобщението на данни да се извършва преди решаване на други ИЗД задачи. По тази причина в разгледаният по-горе модел на ИЗД процес тази задача е част от етапа на разбиране на данни.

Резултатите от описание или обобщение на данни могат да бъдат представени във вид на графики, диаграми, таблици или правила.

1.4.2. Сегментация (клъстеризация)

Задачата за *сегментация* (или *клъстеризация*) има за цел разбиването на данни на интересни и смислени подгрупи или клъстери. Всички членове на една подгрупа споделят общите характеристики, например, при анализа на потребителската кошница цял асортимент на покупки може да бъде разбит на отделни сегменти в зависимост от съдържащи се в него предмети. Алгоритмите за сегментация клъстеризират обекти на подгрупи (клъстери), базирайки се на принципа за максимизиране на вътрешно-клъстерното сходство и минимизиране на външно-клъстерното сходство. Клъстеризацията може да се използва и за създаване на таксономии, т.е. организацията на обекти в йерархия от клъстери, които групират заедно сходните обекти.

Сегментацията може да се извършва ръчно или (полу-)автоматично. Изследователят може да направи предположение за броя на сегменти, използвайки предварителните основни знания за проблемната област, или базирайки се на резултати от описанието и обобщение на данни, а ИЗД алгоритъм да извършва сегментирането на данни на указания брой сегменти. Съществуват и техники за автоматична клъстеризация, които сами могат да намерят по-рано неизвестни и скрити структури в данни без намеса на човека.

Сегментацията може да бъде напълно самостоятелната цел на ИЗД изследване, т.е. определяне на сегментите да бъде основната цел на ИЗД процеса. Обаче, доста често тя е само една стъпка при решаване на други ИЗД задачи. В тези случаи целта на сегментацията е да бъдат поддържани данни в размери, удобни за други ИЗД цели, или да намери еднородни подмножества от данни, които след това могат много по-лесно да бъдат анализирани. Обикновено в големи бази различни зависимости между данни влияят едни върху други и затрудняват намирането на интересни закономерности. В този случай подходящото сегментиране на данни значително улеснява основната задача. Например, анализ на зависимостите между покупките в милиони потребителски кошници е една много тежка задача. Тя може да бъде значително улеснена, ако бъде сведена до намирането на интересни зависимости в отделни интересни сегменти на кошниците, например такива с много голяма обща стойност, кошниците, съдържащи стоки за поддържане на къщи, или кошниците от отделните дни или часове.

Пример на една типична задача за сегментация: една автомобилна компания регулярно събира информация за свои клиенти, която включва такива техните социално-икономически характеристики, като доход, възраст, пол, професия и т.н.

Използвайки клъстерният анализ компанията може да раздели клиентите си на по-разбираеми подгрупи и да анализира структурата на всяка от тях. След това към всяка от подгрупите да бъде приложена отделна специализирана маркетингова стратегия, насочена към повишаването на продажбите.

1.4.3. Анализ на крайностите

Данните могат да съдържат определени обекти, които не отговарят на общия модел на данни. Такива обекти се наричат *крайни* или *екстремни* (outliers). Повечето от ИЗД методи разглеждат екстремните обекти като шум, който се игнорира. Обаче, в някои от приложения, такива като диагностиката или разкриването на измами, рядко срещаните обекти са значително по-интересни от подчиняващи се на общи закономерности. Задачата за намиране на подобни обекти се нарича *анализ на крайностите (outlier detection).*

Екстремните стойности могат да бъдат намерени чрез използване на статистическите тестове, базиращи се на определен вероятностен модел на данни или чрез използване на мерки за разстояние, смятайки че екстремните са тези обекти, които се намират на едно значително разстояние от всички клъстери. Методите, базирани на отклонение, опитват да идентифицират крайностите чрез анализ на разликите в основните характеристики на обекти в групата.

Анализът на крайностите позволява да бъдат разкрити измами с използване на кредитните карти чрез откриване на покупки, чиято стойност е много по-голяма за дадената сметка в сравнение с редовни плащания, осъществявани по сметката. Крайностите могат също така да бъдат открити при анализа на местоназначение и типа на покупката или честота на покупките.

1.4.4. Описание на понятия

Задачата за *описание на понятия* (concept description) цели създаване на *разбираеми* описания на понятия или класове. Целта е не да бъдат създадени пълни модели с висока предсказваща точност, а да бъдат създадени описания на отделни класове, позволяващи да бъде разбрана тяхната структура. Например, една компания би била заинтересована да научи повече за свои надеждни и ненадеждни клиенти. От създадени описания на тези два класа компанията може да изведе начини, позволяващи й да запази лоялните си клиенти и да се освободи от нелоялните.

Описанието на понятия е тясно свързано като със сегментация, така и с класификация. Сегментацията води до преброяването на обекти, принадлежащи към едно понятие или клас без да дава никакво разбираемо описание на този клас. Обикновено, сегментацията се прави преди създаването на описание на понятия. Някои техники позволяват и двете задачи да бъдат изпълнени едновременно.

Описанията на понятие могат да бъдат използвани и за цели на класификацията. От друга страна, някои класификационни техники създават класификационни модели,

които могат да се разглеждат като описание на понятия. Важното различие се състои в това, че класификацията цели създаването на *пълни* модели, т.е. модели, които могат да бъдат приложени към *всички* данни. Описанието на понятия не е задължително да бъде пълно – достатъчно е, ако то описва само важни части от понятия или класове. Ако се върнем към по-горния пример, достатъчно е да имаме описание на тези клиенти, които са напълно лоялни към компанията.

Описанията на понятия най-често се представят във вид на правила или прототипи.

1.4.5. Класификация и предсказване

Класификацията предполага наличието на множеството от обекти, описвани чрез определени атрибути и принадлежащи към различни класове. Целевият атрибут (клас) приема дискретни (символни) стойности и е известен за всеки обект. Целта е да бъдат построени класификационни модели (класификатори), които назначават коректен клас на по-рано неизвестни обекти - обекти с неизвестна стойност на целевия атрибут. Класификационните модели предимно се използват за предсказващото моделиране. Стойността на целевия атрибут може да бъде зададена предварително (например от потребителя) или изведена от сегментацията.

Преди построяване на класификационни модели желателно е да бъде прилаган анализ на изключения, който позволява те да бъдат премахнати и по този начин да бъде подобрен класификационният модел. От своя страна, един класификационен модел може да бъде използван за откриване на отклонения от общата норма.

Класификационните модели могат да се представят във вид на класификационни дървета, правила, невронни мрежи и т.н. Начини на построявана (научаване) и използване на такива класификационни методи бяха подробно разгледани в курса по машинно самообучение.

Предсказването е много сходно с класификацията, като единствената разлика е в това, че при предсказването целевият атрибут (клас) приема не дискретни, а *непрекъснати* стойности. Това означава, че целта на предсказването е да се намери числовата стойност на целевия атрибут за по-рано неизвестни обекти. Този тип задачи често се нарича *регресия*. Когато предсказването работи с данни за времеви серии, тази задача носи название *предвиждане (*forecasting).

Регресионните модели могат да бъдат представени във вид на регресионни дървета, невронни мрежи, генетични алгоритми и т.н.

1.4.6. Анализ на зависимостите

Анализът на зависимостите се състои в намирането на модела, който описва важни зависимости (или асоциации) между отделни елементи от данни или събития. Макар че зависимостите могат да бъдат използвани за предсказването, те

основно се използват за разбирането. Зависимостите могат да бъдат точни или вероятностни.

Асоциациите са един специален случай на зависимостите, които са станали силно популярни в последно време. Асоциациите описват сходството между елементите на данни (т.е. тези елементи (двойки атрибут-стойност), които често се срещат заедно). Типичното приложение на асоциациите е анализ на потребителските кошници. Така следното асоциативно правило "В 30% от всички покупки бира и фъстъци са били закупени заедно" е типичен пример на една асоциация.

Алгоритмите за намиране на асоциации са много бързи и извеждат много асоциации. Избор на най-интересните от тях е предмет на интензивна изследователска работа.

Последователните шаблони са един специален вид на зависимостите, извличани от данни, представляващи последователности от събития във времето (времеви серии). Така например в областта на анализа на потребителските кошници последователните шаблони описват шаблон на определен покупател или група от покупатели във времето.

Зависимостите се представят чрез асоциативни правила, Бейсови мрежи, различни визуализационни техники и др.