**ТЕМА 7. OСНОВНИ МЕТОДИ ЗА РЕШАВАНЕ НА ЗАДАЧИТЕ ПРИ ИЗВЛИЧАНЕ НА ЗНАНИЯ**

**Съдържание на темата:**

7.1. Класификация.

7.2. Клъстеризация.

7.3. Прогнозиране.

**7.1. Класификация**

Задачата за класификация (сегментиране) представлява разпределение на наблюдаваните същности в определени групи по някакъв признак/признаци. Може да се каже също, че класификацията е процес на намиране на модел (или функция), който описва и разграничава класовете данни или понятия.

Изискванията при прилагане на класификация са:

* Всяко разделяне в групи се изпълнява само по един признак;
* Елементите от различните групи трябва взаимно да се изключват.
* Разделянето по следващ признак трябва да бъде последователно.

При класификация данните се разделят на обучаващо и тестово множество, приблизително в пропорция 2:1. Обучаващото множество включва данни за създаване на модела, а тестовото за проверка на работоспособността му с нови данни. Тестовото множество не трябва да зависи от обучаващото.

Процесът на класификация се състои от 2 етапа: създаване на модел и използването му.

При конструирането на модела:

- всеки пример се отнася към един от предварително определените класове;

- на този етап се използва обучващото множество, на чиято база се създава модела;

- моделът се представя чрез класификационни правила, дърво на решенията, математическа формула или др.

Използването на модела е за класификация на нови или неизвестни значения:

* Оценява се точността на модела:

= известните значения от тестовия пример се сравняват с резултатите от използването на получения модел.

= нивото на точност е процентът на правилно класифицирани примери в тестовото множество.

* Ако точността на модела е допустима, той може да се използва за класификация на нови примери.

Освен нивото на точност за оценка на методите за класификация се използват следните величини:

* Скорост- определя времето за създаване на модела;
* Устойчивост- коректно изпълнение дори при наличие на непълни данни;
* Интерпретируемост- моделът да бъде разбираем и обработваем.

При класификацията се спазва принципът за максимално подобие на обектите в класа и минимално подобие между класовете.

Класификацията се използва от фирмите за обслужване сигурността на данните, по-конкретно установяване на нивото на конфиденциалност на дадена информация- конфиденциална, частна, чувствителна, публична. Често се включва и времеви елемент, чрез който след изтичане на период информацията преминава в друг вид. Например от конфиденциална към публична.

Класификацията се използва и за редица бизнес дейности във фирмата, като сегментиране на потребителите, доставчиците и други контрагенти на фирмата, за избор на канал за дистрибуция, контрол върху качеството на стоките, изразходването на ресурсите, контрол на финансови аномалии и др.

Класификацията представлява вид обучение с учител, защото за всеки пример от обучаващото множество е известен класа, към който той принадлежи.

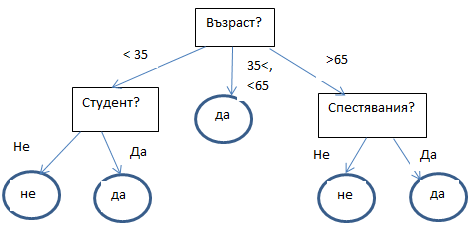
Методите за класификация се класифицират на традиционни (конвенционални) и интелигентни. Към интелигентните спадат невронните мрежи, размити методи, агенти за извличане на знания. Към традиционните: метод на опорните вектори, дървета на решенията и др.

**Дървото на решенията** е често използвана дървовидна структура за представяне на моделите, както е и метод за класификация. Идеите за създаването му са предложени от Ховленд (Hoveland) и Хант (Hunt) в края на 50-те години на 20-ти век.

В дървото на решенията всеки възел обозначава тест на стойност на зависима[[1]](#footnote-1) променлива. Всеки клон представя изход от теста, а листата на дървото– класове (фиг.7.1). С други думи: един клон представлява класификационна група и в зависимост от отговора на въпроса към тази група клона се разделя на други клони или листа (крайни възли). За избора на променлива за тестване, методът използва информативността й (информационната печалба[[2]](#footnote-2)) или статистичеки критерий. Последният използва индекс Gini (по името на италианския икономист Corrado Gini), който оценява „разстоянието“ между разпределните класове.

Gini(с)= 1- ∑ pj2 , където с е текущият възел, а pj- вероятността във възел с да има клас j.

Създаването на дърво на решенията не изисква знания за предметната област. Представянето на знания като дърво е интуитивно и лесно разбираемо от хората.



**Фиг.7.1. Примерно дърво на решенията**

Основен недостатък на дървото на решенията е, че при усложняване на моделираната ситуация, сложността на дървото нараства [експоненциално](http://basaga.org/wiki/index.php?title=%D0%95%D0%BA%D1%81%D0%BF%D0%BE%D0%BD%D0%B5%D0%BD%D1%86%D0%B8%D0%B0%D0%BB%D0%BD%D0%BE&action=edit&redlink=1).

Aлгоритми, които построяват дърво на решенията са ID3, C4.5. и CART.

**Асоциативните (класификационни) правила** са както форма на представяне на модела- резултат от DM, така и метод за откриване на закономерности от свързани събития и класификация. Te са във вида **ако X, то Y.** Значимостта на едно асоциативно правило се измерва със следните две мерки:

* подкрепа – процента на транзакциите, в които съответната асоциация е наблюдавана;
* доверие – процента на транзакциите, в които ако транзакцията съдържа **X**, то тя ще съдържа и **Y** или колко често, когато е изпълнено условието **X** се изпълнява **Y**.

Подкрепата и доверието са обективни мерки за измерване значимостта на шаблони, изразени като асоциативни правила.

Така всяко асоциативно правило може да се запише във вида:

**ако X то Y (подкрепа, доверие)**

Например, на фиг.7.2. са показани 2 таблици за продажби на стоки с регистрация на транзакциите в лявата и честота на срещане на стоката- в дясната.

****

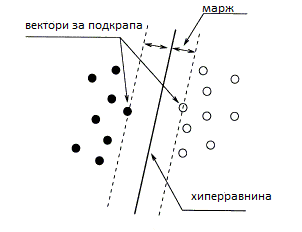
**Фиг. 7.2. Честота на срещане на стоките в транзакциите**

За правилото „ако А, то С“ подкрепата е 50% (от дясната таблица), а доверието се получава по формулата **подкрепата(А,С)/подкрепата(А)** или **50%/75% = 66.6%.** Тогава правилото може да се запише:

**ако А, то С (50%,66.6%).**

С мерките се асоциира праг, който може да се контролира от потребителя. Например правила, които не удовлетворяват праг на доверие 50% не може да се разглеждат като интересни.

**Методът на опорните вектори** (Support Vector Machines-SVM) за първи път е представен през 1995 като надзираван метод за класификация на входни вектори (Cortes & Vapnik, 1995). В основата си методът се базира на откриването на такава хиперравнина, която да раздели с най-висока точност обектите в два класа (фиг.7.3).



**Фиг. 7.3. Хиперравнина, разделяща обектите в два класа**

Алгоритъмът за обучение на SVM изгражда модел, чрез който примерите се присвояват към един от класовете. Обучаващото множество се състои от вектори, които имат най-малко разстояние от хиперравнината. Такива вектори се наричат вектори за подкрепа, откъдето идва и наименованието на метода.

Класификационната хиперравнина се избира по време на обучение като уникалната равнина, която разделя известните положителни примери от негативните с максимален марж. Маржът представлява разстоянието от хиперравнината до най-близката точка от положителното и отрицателното множество. Колкото маржът е по-голям, толкова разделянето на наблюденията е по-голямо. По време на обучението SVM „търси“ хиперравнина с най-голям марж. По този начин се осигурява максимално разделение между класовете.

Когато данните са линейно неделими[[3]](#footnote-3), се налага въвеждането на ядро на функцията. При него изходните данни се трансформират във функции в многомерното пространство. По този начин се премахва линейното картографиране и се дава възможност за намиране на по-сложна хиперравниа, която да раздели преобразуваните данни. Проблемът, който се появява при този вид SVM е, че за решаване на реални проблеми е трудно да се намери такава хиперравнина, която да раздели обектите с висока прецизност.

**7.2. Клъстеризация**

Kлъстеризация (групиране) е задача за разделяне на дадена извадка от данни на непресичащи се еднородни множества, наречни клъстери. Понятието „клъстер“ се превежда като „струпване“, „грозд“. Всеки клъстер се състои от подобни обекти, а обектите от различните клъстери съществено се различават. Aгрегирането на подобните обектите в групи се нарича клъстеринг. Целта на клъстеризацията е търсене на съществуващи структури.

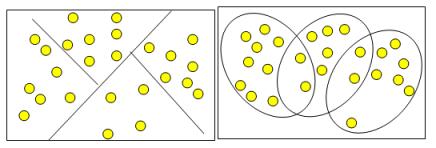
Eдна от разликите с класификацията е в това, че **класовете от данни не са предварително** определени (Таблица 7.1).

**Таблица 7.1.**

**Сравнение между класификация и клъстеризация[[4]](#footnote-4)**

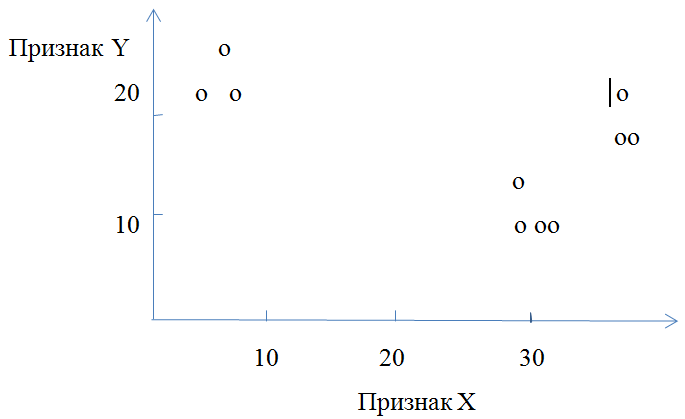
|  |  |  |
| --- | --- | --- |
| **Характеристика** | **Класификация** | **Клъстеризация** |
| Контролируемост на обучението | Контролируемо | Неконтролируемо |
| Стратегия | Обучение с учител | Обучение без учител |
| Наличие на маркировки за класа | Обучаващото множество се съпровожда от маркировки, указващи класа към който се отнася наблюдението | Маркировките на класа на обучаващото множество са неизвестни |
| Причина за процеса | Да се класифицират нови данни на основата на създадения модел | Да се установят съществуващите клъстери от данни |

Класовете може да бъдат пресичащи се или непресичащи се (фиг.7.4). Различните методи могат да създават клъстери с определени размери (малки или големи), да са различно чувствителни към шума и т.н. В резултат на прилагането на различни методи може да се получат нееднакви резултати и това е нормално, тъй като има особености в работата на различните алгоритми.



Фиг.7.4. Пресичащи се и непресичащи се клъстери

Процедурата на клъстерния анализ е следната: нека набор от данни се състои от примери, които имат по два признака Х и Y. Променливите Х и Y са представени на диаграма на разсейването (фиг. 7.5). Примерите, които по стойностите на Х и Y са сходни, принадлежат на един клъстер; обектите от различни клъстери не са сходни.



**Фиг. 7.5. Примерна диаграма на разсейването**

Критерий за определяне сходството и различието между клъстерите е разстоянието между точките на диаграмата на разсейване. Най-разпространеният начин е изчисляване на евклидовото разстояние между 2 точки i и j:

Dij=√ (xi-xj)2 + (yi-yj)2

Aкo е необходимо да се намери разстоянието между 2 точки в тримерното пространство в случай на евклидово разстояние се ползва същата формула с добавяне на още 1 събираемо за третата координата.

Клъстерният анализ се основава на 2 предположения:

-променливите допускат по принцип разделяне на съвкупността от обекти на клъстери;

- изборът на мащаба за измерване на данните е правилен. Данните може да принадлежат на несъпоставими интервали, например Х е в интервала 100-700, а на Y в интервала 0-1. Този проблем се решава с предварителна нормализация на данните (привеждане към единен интервал). Мащабиране на стойностите на числови данни към предварително зададен числов интервал се нарича нормализация на данни.

Основен метод за **индивидуална нормализация на данни** е линейното преобразуване. При този метод дадена стойност *v* се заменя с *v’* чрез формулата:

в която *minv* и *maxv* са границите на проявление на оригиналната величина, а *new\_minv* и *new\_maxv* границите на трансформираната величина.

Това е възможно най-елементарният и надежден подход, който може да се прилага към всякакви категории данни.

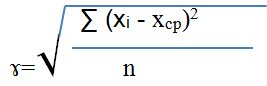
За нормировка може да се ползва и средно аритметичната стойност

1

хср = ----- ∑ xi , заi oт 1 до n

n

и средноквадратично отклонение:

 , заi oт 1 до n

Toгава нормираната i- та стойност xi на променливата х е:

xi -хср

х iср= ------------------

ɤ

За дадена променлива се извършва нормировка на всичките й стойности.

Процесът на клъстеризация зависи от избрания метод.

Разработени са стотици различни алгоритми за клъстеризация. Те могат да се представят така:

* Йерархични методи.
* Нейерархични методи

Нейерархичните методи имат по-голяма стабилност по отношение на шума, неправилния избор на показатели, включването на незначими променливи в набора, участващ в клъстеризация. Цената, която трябва да бъде платена за тези предимства на метода, е че анализаторът трябва предварително да определи броя на клъстерите, броя на повторенията или правилото за спиране, както и някои други параметри на клъстера. Това е особено трудно за начинаещи.

Ако няма предположения за броя на клъстерите, препоръчително е да се използват йерархични алгоритми. Ако обаче размерът на извадката не позволява това, възможно е да се извърши поредица от експерименти с различен брой клъстери, например, да се започне разделянето на групата на две и постепенно да се увеличива броя им, за да се сравнят резултатите. Нови клъстери се формират, докато се изпълни условие за преустановяване на делението. Така се постига доста голяма гъвкавост на клъстеризация.

Йерархичните методи, за разлика от нейерархичните не определят броя на клъстерите, но изграждат цялото дърво от вложени клъстери. Предимството на тази група от методи в сравнение с нейерархичните методи е тяхната яснота и възможност да се получи подробна представа за структурата на данните, както и тяхната нагледност.

При използване на йерархични методи е възможно лесно да се идентифицира шум в даден набор от данни и в резултат да се подобри качеството на данните. Тази процедура е в основата на алгоритъма за клъстеризиране в две стъпки. Подобен набор от данни може по-късно да се използва за извършване на не-йерархично групиране.

Още един аспект за йерархичните методи е, че не могат да работят с големи набори от данни.

Недостатъците на групите методи са следните: при йерархичните е необходимо много време, докато се разделят наблюденията в групи, а при нейерархичните е необходимо преди да се започне да се определи броя клъстери. Предимствата на вторите пред йерархичните са, че се работи бързо, не се влияе от променливи данни и не е необходимо да има точно определено разстояние между групите.

За да се определи ефективността на конкретен метод се прави следното:

* разглеждат се няколко предварително известни различни групи, смесват се по случаен начин. Провежда се клъстеризация за възстановяване на първоначалното разделяне. Съвпадението на обектите в получените и началните групи е показател за ефективността на метода.
* Установяване на контролни точки и проверка на получените клъстери.
* Определяне на стабилността на клъстеризация чрез добавяне в модела на нови променливи.

**Алгоритъм k- means (k-средни)**

Най-често срещан сред нейерархичните методи e алгоритъмът k-means, наречен също бърз клъстер анализ. Пълно описание на алгоритъма може да се намери в работата на Hartigan and Wong (1978). За разлика от йерархичните методи, които не изискват предварителни предположения за броя на клъстерите, е необходимо да има хипотеза за най-вероятния брой клъстери.

Алгоритъмът k-средни стойности конструира k клъстери, разположени на възможно най-големи разстояния един от друг. Основният тип задачи, които алгоритъмът k-средни решава, е наличието на предположения (хипотези) за броя на клъстерите и че те трябва да бъдат колкото е възможно по-различни. Изборът на числото k може да се основава на резултатите от предишни изследвания, теоретични съображения или интуиция.

Предимства на алгоритъма k-средни:

• лесна употреба;

• скорост на използване;

• яснота и прозрачност на алгоритъма.

Недостатъци на алгоритъма k-средни:

• Алгоритъмът е твърде чувствителен към шума, което може да изкриви средната за клъстера стойност. Възможно решение на този проблем е използването на модификация на алгоритъма - к-медиен алгоритъм;

• Алгоритъмът работи бавно с големи бази от данни. Възможно решение на този проблем е да се използва извадка от данни.

**7.3. Прогнозиране**

Прогнозирането е свързано с определени тенденции на динамика на конкретен обект или събития на основата на ретроспективни данни.

Преди началото на прогнозирането трябва да се отговори на следните въпроси:

* Какво ще се прогнозира;
* При какви времеви параметри;
* С каква точност на прогнозата.

При отговор на първия въпрос се определят променливите, които ще се прогнозират. Напр., сумата на продажбите, ниво на производство и т.н. При избора на променливи следва да се отчита достъпността на ретроспективните данни, предпочитанията на лицата, вземащи решения и т.н. Често при тези задачи се предсказва не самата променлива, а измененията на стойностите й.

При втория въпрос се определят период, хоризонт и интервал на прогнозиране. Периодът означава основната единица време, за която се прогнозира. Напр., месец. Хоризонт- брой на периодите в бъдеще, които покрива прогнозата. Напр., ако търсим прогноза за 12 месеца с данни за всеки месец, периодът е месец, а хоризонтът- 12 месеца. Интервалът е честотата, с която се прави нова прогноза.

При избора на параметри трябва да се отчита, че хоризонтът на прогнозиране не трябва да е по-малък от времето, необходимо за реализация на решението, взето на основата на тази прогноза. Само тогава прогнозирането има смисъл.

С увеличаване на хоризонтът на прогнозиране точността като правило се намалява, а с намаляването му- се повишава. Тогава може да се подобри качеството на прогнозиране намалявайки времето, необходимо за реализация на решението, за което се реализира прогнозата и следователно намалявайки хоризонта и грешката на прогнозиране.

Точността се характеризира с грешка на прогнозиране. Тя зависи от използваната система за прогнозиране. Най-разпространените видове грешки са:

* Средна грешка- изчислява се с просто усредняване на грешката на всяка стъпка
* Средна абсолютна грешка. Ако е равна на 0, то имаме съвършена прогноза.
* Сума на квадратите на грешките (средноквадратична грешка). Изчислява се като сума от квадратите на грешките. Това е най-често използваната оценка на точността на прогнозата.
* Относителна грешка.

Разглеждат се също краткосрочни прогнози- 1-3 периода напред; средносрочни- 7-12 периода напред и дългосрочни.

Най-разпространените методи за прогнозиране са НМ и регресионният анализ.

**Регресионният анализ** е направление в математическата статистика, в което се изучават и оценяват възможните функционални зависимости между две или повече случайни величини. Той дава оценка на връзката между една зависима и една или повече независими променливи. Като резултат се получава регресионно уравнение, което е модел на явлението, определено от изследваните величини. При определянето на функционалните зависимости между величините трябва да има предпоставки за причинно-следствена последователност на измерваните величини. Регресионният анализ дава отговор на въпроса кои са причините. Целите на регресионния анализ са да се определи в каква степен зависимата променлива варира като функция от измененията на независимата променлива. Когато има повече от една независима променлива анализът се нарича факторен анализ.

1. Y е зависима от Х променлива, ако Y= f(X) [↑](#footnote-ref-1)
2. Han Jihan. Data mining – concepts and techniques. Elsevier 2011. [↑](#footnote-ref-2)
3. Няма равнина в двумерното пространство, която да разделя единия вид резултати от другия. [↑](#footnote-ref-3)
4. Щерев Й., Анализ на данни, изд. „Фабер“, В.Търново, 2010,стр.136 [↑](#footnote-ref-4)